UNIVERSAL
LIBRARY

OU_162342

UNIVERSAL
LIBRARY

OUP—24—4-4-69—5,000

OSMANIA UNIVERSITY LIBRARY

Call No. 311 / B63E Accession No. P.G. 18083

Author Blair, M. M.

Title Elementary Statistics with general application. 1952.

This book should be returned on or before the date last marked below.

A WORD TO THE TEACHER

This book is the result of fifteen years of practical experience in teaching statistics to more than 3,000 college students ranging from the sophomore to the graduate level. It was developed and taught by the laboratory method. Much of the material is pitched to the sophomore level. Some of it is more difficult. After teaching the subject for thirty semesters the author is convinced that a large body of fairly difficult statistical theory and methods can be taught to sophomores and juniors if it is presented in a practical non-technical way. He is also convinced that statistics should first be taught to most students from the standpoint of preparing them to be *consumers* of statistics instead of *creators* of researchers. Those who learn first to consume, to understand, statistics which others have prepared will if interested and as the need arises begin to do creative work.

This text is set up in four parts to meet the teacher's needs for subject matter on various levels of difficulty adapted to students of successive levels of age and attainment. The following combinations of the four parts are suggested for the various levels.

1. For Junior College students and for Sophomore classes in Senior colleges, use Chapters 1 to 9, inclusive, and the easier portions of Chapters 13 to 19, inclusive, with special emphasis on Chapters 1 to 9 and 13, for a one-semester course.

2. For Normal Schools and Teachers' Colleges, use Chapters 1 to 9, inclusive, and Chapters 13, 20, 23 and 25 for a one-semester course.

3. For Schools of Business Administration and Commerce and Schools of Agriculture, use Chapters 1 to 9, inclusive, and

Chapters 13 to 19, inclusive, for the first semester and Chapters 10 to 13 and 20 to 25, inclusive, with reviews of Chapters 14 to 19 for the second semester.

4. For a one-semester course for advanced or graduate students who have had one or more semesters of statistics, use Chapters 12 to 25, inclusive, with heavier assignments and original problems.

Since the first edition was published eight years ago, the author has experimented with the presentation of this material to 2000 students on the graduate and undergraduate level in two universities. The present rearrangement of the subject matter and suggestions as to methods of presentation are the results of that practical teaching experimentation.

Because of the gradual and progressive difficulty of the material of this text from Part I through Part IV, it will be found to serve well as a coherent and progressive outline for a series of lectures on statistics which might add additional materials from other texts or from original sources. It may well serve as the *core material* for many types of courses which are varied in detail to meet special requirements. It is not thought that any teacher will expect to use all the material in this text in one semester. Rather will he select from the several Parts those particular and limited materials for which his course is designed. It is believed that by this adaptive, selective method much better results will be obtained. The wise teacher will always *fit* his subject matter to his students.

TEACHER'S USE OF EXERCISES AND PROBLEMS

The teacher will notice that no specific and detailed instructions are given in most cases as to the use of the exercises and data at the end of each chapter. The author has two reasons for this omission.

First, most of the data included are suitable for many uses and for the computation of a large number of statistical measures. To specify a few specific problems in each case would be to limit the full possibilities of the use of the data. To specify all profitable

computations would be to lengthen unduly the assignments and to make them seem confusing to the students.

Second, different teachers in different classes of varying degrees of maturity and with varying objectives will wish to emphasize different phases of statistical methods with their assigned problems. To achieve the best results each teacher will wish to determine and assign his own problems for the particular purpose in mind. This practice vitalizes the work. It also makes possible the use of the same data for many semesters without routine and monotonous duplications. Such a method makes the teacher, the pupil, and the text, partners in a vital problem for a specific objective. It gives teaching and study a sense of greater zest and practical purpose.

REFERENCE BOOK

The businessman and practicing statistician will also find this text a valuable reference book. The use of complete worksheets and the detailed solution of problems not only makes the learning process quicker and easier for the student but it enables the businessman in his office and the research worker to have a *model-to-go-by* in organizing his statistical work. The large number of reference tables enables the businessman, the student, and the research worker to find the measure of error, the required formula or the specific information desired with the least time and effort. Business executives will find the materials in the book easy to locate and use. It will be a convenient *Handbook of Statistics*.

M. M. B.

Tulsa, Oklahoma
Feb. 22, 1952

A WORD TO THE STUDENT

Statistics is considered a dry subject by many college students. It often is. The purpose of this book is to supply an adequate quantity of moisture in the form of simple, well-organized material presented with a minimum of mathematics and a strong emphasis on practical, commonplace, everyday application to the problems of life.

More and more, statistics is taught in most colleges from the standpoint of *consumption* rather than with a view to *creation*. Not one student in ten, perhaps not one in fifty, will ever become a productive statistician or a research worker, or even a statistical clerk. But our civilization has reached the point where it is impossible to read intelligently the baseball scores, the stock market reports, the weather report, the business section of the daily paper, trade magazines, popular books on geography, travel, economics, sociology, agriculture, or production without an elementary knowledge of statistics. The reason for this condition is that statistics are a kind of shorthand, a time-saving device in a busy world to say a great deal in a small space and in a short time.

Today much advertising is done on the printed page, poster, and over the radio in the form of statistics. Banks, labor unions, government agencies, insurance companies, railroads, retail stores, military organizations, and factories employ vast displays of statistical material for the information of their employees, their owners, and for the general public. The person who has no knowledge of these simple uses of statistics is definitely handicapped in the modern world. This book is presented to you as a quick and easy way of acquiring this necessary information. Its subject

matter cannot be learned without study, but with an ordinary amount of conscientious effort the average student can acquire this knowledge in a single semester. It is well worth your effort. You will have daily use for many of these methods. You will need them in your business.

ACKNOWLEDGMENT

The author makes no claims for originality in subject matter. He has chosen and limited his materials for the purpose of meeting the needs of that large and ever-increasing body of students and businessmen who must have an elementary knowledge of statistics. The book is designed for beginners. It is prepared for those who must consume statistics rather than create them. It is an *elementary* text. The author has leaned heavily on a large number of his predecessors and associates in this field to whom he expresses his indebtedness and appreciation.

I am indebted to Professor R. A. Fisher, also to Messrs. Oliver and Boyd, Ltd., of Edinburgh, for permission to reprint Tables No. 3, Table of Chi-Square; Table No. 4, Table of t; Table No. 6, 5 percent; 1 percent points of the distribution of z, from their book, *Statistical Methods for Research Workers.*

I am indebted to Professor George W. Snedecor and also to Collegiate Press, Inc., Ames, Iowa, for permission to reprint Table 10.3, Points for Distribution of F, from their book, *Statistical Methods.*

I am indebted to Dr. Mordecai Ezekiel and also to John Wiley and Sons, Inc., New York, for permission to use portions of Appendix 1, Methods of Computation, from their book, *Methods of Correlation Analysis.*

I am indebted to Professor Frederick C. Mills and also to Henry Holt and Company, New York, for the methods of presenting Least Squares and the Analysis of Small Samples in their book, *Statistical Methods, Revised.*

I am indebted to Professors Davies and Yoder and also to John

Wiley and Sons, Inc., New York, for the use of the table on Squares, Square Roots, and Reciprocals from their book, *Business Statistics*, Second Edition.

I am also indebted to my teacher and friend, the late Professor Harry Jerome of the University of Wisconsin, as well as to my teachers, Professors Selig Perlman, John R. Commons, and George S. Wehrwein, all of the University of Wisconsin, for their aid and encouragement, and for their methods of presenting statistical problems in the teaching of theory. I wish to acknowledge my indebtedness to Professor Harry G. Brown, of the University of Missouri, for the inspiration of his effective methods in the organization of subject matter and teaching large groups of students; and also to the late Professor James Harvy Rogers, of the University of Missouri and later of Yale University, for his effective use of statistics in teaching.

I am indebted also to my colleague of the Department of Sociology of Oklahoma Agricultural and Mechanical College, Professor William H. Sewell, for his aid in reading and constructive criticism of the entire manuscript.

My indebtedness is deeper than I can ever repay to the more than 3,000 students who have passed through my classes in statistics during the last fifteen years and on whom I have practiced in the selection, organization, and presentation of the subject matter and methods of this book.

I wish also to express my appreciation for the knowledge of practical statistical methods and their application to management and research problems which I gained while employed in the Statistical Department of the B. F. Goodrich Co., Akron, Ohio, The General Tire Company, Akron, Ohio, The Mercantile Trust Co., Little Rock, Arkansas, and the United States Reduction and Refining Company, Colorado Springs.

I wish further to express my indebtedness for the many kindly and constructive criticisms of the First Edition which I have received from my many friends who are teachers and users of statistics. As far as is consistent with my basic purpose to make *this text easy for teachers to teach* and *for undergraduate students to*

understand, I have incorporated many of those suggestions into this Revised Edition.

I wish finally to express my deep appreciation to my daughter, Ann Blair, for her continual aid in typing and checking the manuscript, and to Professor John D. Gemmill, of Tulsa University, for his checking of the manuscript and proofs of the entire book.

M. M. B.

CONTENTS

Part One: The Analysis of Large Samples

1. THE MEANING AND USE OF STATISTICS . . . 1
2. GRAPHIC PRESENTATION 11
3. TABULATION 47
4. RATIOS AND PERCENTAGES 72
5. FREQUENCY DISTRIBUTIONS 93
6. AVERAGES 112
7. DISPERSION, SKEWNESS, AND VARIATION . . 150
8. LINEAR REGRESSION 176
9. CORRELATION AND DETERMINATION 215
10. NON-LINEAR REGRESSION 240
11. MULTIPLE AND PARTIAL CORRELATION . . . 285
12. TABULAR ANALYSIS 317
13. PROBABILITY, SAMPLING, AND ERROR . . . 346

Part Two: Types of Business Fluctuations

14. THE ANALYSIS OF TIME SERIES 398
15. SECULAR TREND 421
16. SEASONAL VARIATION 443
17. CYCLICAL FLUCTUATIONS 463

18. INDEX NUMBERS 487
19. PRACTICAL APPLICATION OF INDEX NUMBERS . 513

Part Three: The Analysis of Small Samples

20. THE ANALYSIS OF SMALL SAMPLES 522
21. THE ANALYSIS OF VARIANCE 557
22. THE ANALYSIS OF COVARIANCE 586

Part Four: Organization of Statistical Problems and Reports

23. ORGANIZATION OF A STATISTICAL PROBLEM . . 608
24. COLLECTING AND EDITING STATISTICAL DATA . 620
25. PREPARATION OF STATISTICAL REPORTS . . . 644

APPENDIXES
I. GLOSSARY OF SYMBOLS AND INDEX OF FORMULAS 656
II. APPENDIX OF TECHNICAL TERMS 668
III. MATHEMATICAL APPENDIX, SQUARE ROOTS, SIMULTANEOUS EQUATIONS, AND LOGARITHMS . 683

TABLES
I. ORDINATES OF THE NORMAL PROBABILITY CURVE 693
II. AREAS OF THE NORMAL CURVE OF ERROR IN TERMS OF ABSCISSA (ABSICA) 694
III. RELATIONS BETWEEN r AND z FOR VALUES OF z FROM 0 TO 5 695
IV. SQUARES, SQUARE ROOTS, AND RECIPROCALS, 1–1000 696
V. COMMON LOGARITHMS OF NUMBERS FROM 1 TO 10,000 TO FIVE DECIMAL PLACES 707
INDEX 729

Part One

The Analysis of Large Samples

CHAPTER 1

THE MEANING AND USE OF STATISTICS

Statistics is the theory and method of analyzing groups of related numbers in order to discover their relationships and meanings. A single number does not lend itself to statistical analysis. The wage of one workman for a single month is too limited a quantity for comparative study. There must be at least two numbers before comparisons can be made or relationships discovered. The monthly wage of a workman for two or more months, or the wages of two or more workmen, is a fit subject for study. Not a datum, but data are the subject matter of statistics.

In common usage the term "statistics" is employed to signify any group of figures or data. A common statement is that the lecturer quoted many statistics or that the book is dry and uninteresting because it is full of statistics. Such usage makes "statistics" equivalent to "numbers." This is a limited, inadequate, and, in a sense, a wrong use of the term. The word "data" is the proper term with which to designate raw, or unanalyzed, numbers. In the true sense, statistics appear only when the data have been organized and analyzed and the relationships existing among them have been stated in summary or relative terms.

Statistical methods are ways of thinking that are almost universal. Millions of persons who have never heard of statistical theory follow its basic principles every day in making the common necessary decisions of life. The housewife in shopping for groceries reads the various food advertisements and inquires as to prices of various articles at several stores until she arrives at a workable idea of the level, average, and range of prices. In her daily efforts to get the most of quality and quantity for her money, she is, perhaps, without knowing it, using sound and adequate statistical methods. The new boy at the party as he dances, talks, and laughs with the various girls quickly forms an opinion as to their grace, ease, culture, charm, and sophistication, and unconsciously compares them with the girls back home or in other communities as higher or lower, better or worse. If the young man had obtained his information concerning these girls from a long formal questionnaire which gave in quantitative measurements the income and occupation of their parents, the kind of car they drove, their years in school, their sports, hobbies, opinions, and social and religious associates, and a hundred other points, all of which had been summed, averaged, correlated, and stated in terms of ratios, indexes, coefficients, and other formal statistics, he would have had a fuller and more accurate measure of this group of girls, but the main difference would have been that the questionnaire formal method was only more mathematical and detailed, though not essentially different in principle from his method at the party. In all our social relationships we are compelled to make decisions about associates and friends on the basis of samples. When boys and girls complete high school or college, stores, mills, and factories employ them on the basis of the data in their transcripts and recommendations. A young man and a young woman on the sample of a brief conversation and chance acquaintance may make a date, and later, on the inadequate sample of a few months' courtship, assume the heavy responsibilities of married life. Because life is so short that we cannot know all about anything, we have to make our decisions on the basis of sample information. This is the statistical method. Since no sample can be perfect, the difference between our methods of making our daily decisions

and the most refined statistical methods is one of degree and not of kind. We literally live by statistics.

Our common expressions in the home and shop and on the street reveal the statistical basis of our daily thought. "You can't make a silk purse out of a sow's ear" is a homely way of saying that there is a high correlation between ability and achievement or between character and results, and that this particular individual has such a low mark on the X axis that we must expect him to have a low mark on the Y scale. When you say, "Birds of a feather flock together," you are indicating that the trend, or regression line of character is positive, that good people seek good people and bad people seek bad people, and that the evil characters with which this particular person associates indicate that he is of low character himself. The expression, "Penny wise and pound foolish," means that there is a negative regression line, or inverse trend between saving on small items and wasting on large ones. "If you go in at the big end of the horn, you will come out of the little end" expresses the same statistical concept. In fact the relations of trend or regression as developed in Chapter 8 and the ideas of correlation analyzed in Chapter 9 are only highly exact and mathematical methods of expressing relationships which we use every day in the common affairs of life.

The close relationship between statistical theory and methods and the universal behavior of mankind may be illustrated at great length, but a few more will suffice at this point. The expression, "From Dan to Beersheba," or "From Maine to California," or "From here to Landsend," means a wide range of data. "The exception that proves the rule" is a way of saying that the deviation of the items from the mean, average or trend merely proves that there is a stable or dependable average or base. "He is lost in the woods," or "He can't see the town for the houses," means that one is confused by the vast amount of data and has failed to analyze them into their true mean, trend, and relations.

Suffice it to say that we all, often without realizing it, select friends, sweethearts, wives, husbands, business partners, professions, automobiles, clothing, real estate, vacations, and investments, and avoid enemies, dangers, losses, and death as long as

we can by means of the basic principles of statistics as computed in averages, dispersions, trends, correlations, indexes, errors, and sampling. In studying the science and methods of statistics, therefore, we are only acquiring a clearer understanding and more accurate measurement of principles and rules we have long been using and cannot get along without.

Statistics is the theory and method of analyzing groups of measurements. A series of measurements having been made of stock prices, or distances to the stars, or gas pressures, or yields of wheat, or rainfall, or plumbers' wages, or heights of children, or interest rates, or births, or exports, or of anything, statistics supplies the theory and methods for classifying, organizing, and computing representative values of these measurements. When experimenters and investigators have completed their projects they often have a body of data so complex, varied, and voluminous that their meaning is not clear. The real trends and relationships are so obscured by a mass of details that careful analysis is required to bring out the significant and principal points. The processes and rules of analyzing these numbers so that their true inner meaning is revealed is the science of statistics.

One might ask the simple question, How tall are grade school children? The attempt to get an answer would raise a number of other questions, such as: Which grades, first to six, or first to eight? White or colored children? Native white or foreign born? Scandinavian or Italian children? etc.

Having determined the particular population of children to be studied, we measure the height of one child. This single datum is not sufficient for a statistical study. We then measure another child and find that he is taller or shorter than the first. We can now make comparisons, but our sample is too small to obtain significant results. We, therefore, continue to measure children until we have recorded the heights for some thirty, forty, or even one hundred, or more. We are perplexed by the wide variation of heights. Some children are only forty inches tall; others are sixty or even seventy inches tall. The original question remains, What is the height of grade school children? The answer requires statistics. We cannot use all the numbers as an answer. That

would be completely confusing. The shortest height does not adequately represent the group; it is too short. The greatest height has no better claim; it is too tall. We must compute an average, or perhaps, various kinds of averages. Even this is not sufficient. Measures of scatter of the heights around their mean as well as relationship with age, sex, weight, health, race, and other qualifying conditions and related variables must be computed before we have the full answer. The simple question of the height of grade school children requires the science of statistics for an adequate answer.

Another common question is, What is the price of wheat? This query suggests others, such as, What kind of wheat? Spring or winter wheat? White or red wheat? Hard or soft wheat? Wheat on the farm or in the city elevators? For a complete analysis many other questions must be answered to delimit accurately and scientifically the measurements desired. Having decided what kind of wheat to measure, and where and when to measure its price, we still are faced with the problem of differentiating between several types of prices and price movements, such as, daily changes, seasonal variation, secular trends, and cyclical fluctuations, besides accidental or occasional factors. An adequate measurement of the price of wheat requires the science of statistics.

Similar requirements would arise in the measurement of the tensile strength of steel, the cost of going to college, the cost of textbooks, the price of theater tickets, the rent of dress suits, students' grades, the distance to the moon or Venus, the beginning salaries of teachers or engineers, or any other set of measurements. Any factor for which varying measurements may be obtained requires statistics for the analysis of the data. Such factors are called *variables*, because they change from one measurement to the next. The quantity of rainfall is a variable, changing from day to day, and month to month, as well as from year to year. Similar variability is manifest in length of day, heat, wind velocity, road accidents, fuel consumption, food intake, health, the death rate, the interest rate, the tax rate, and practically all other human activities and natural phenomena.

DIVISIONS OF STATISTICS

Statistics as a science may be divided into statistical theory, statistical methods, and statistical results or measures.

Statistical theory is largely mathematical, deductive, and logical. It rests on the inherent relationships of numbers as expressed in algebra, geometry, analytics, and calculus. Statistical theory explains the reasons for and the relationships between statistical methods, computations, and results. It supplies the logical outline and scheme of relationships on which any sound and dependable analysis of numbers must rest. Statistical theory explains why the various statistical methods must be set up in specific forms for particular results and the degree to which each statistic is dependable. It is the body of quantitative and numerical logic which guides the statistician to desirable and dependable results. Clear and accurate thinking is as necessary in the analysis of numbers as in any other field of science. The student should be certain always that he understands the full reason for his computations and why and to what extent they are significant. While a knowledge of calculus is desirable in the study of statistics, it is not absolutely necessary. Students with a fair command of algebra can operate the formulas and understand their results. This text is designed to meet the needs of students who have as little as one semester of algebra. With that limited background they can manipulate the formulas and have a working idea of their meaning. To become an able statistician, however, the student should have at least two years of mathematics.

Statistical methods are the devices for achieving the desired ends explained in the theory. They are the outward manifestations of the inward logical relations of the theory. Since a method is a means to an end, its relative excellence and appropriateness depend on the specific end sought. There are often several ways of doing approximately the same thing, or reaching identical results. Some methods are shorter, or easier, or give more complete analysis, or are more logically related to particular types of data. In the following chapters often several similar or slightly varying methods will be given. This is done not to confuse the student,

but to equip him more completely for effective work. Just as a dentist or surgeon has many kinds of instruments for different operations, and just as an automobile mechanic has many tools for the various parts of an automobile, so a statistician must have many methods of manipulating different types of data, or the same data for various purposes. Geographic, or place data, require a type of treatment different in some respects from that required by time data. Percentages, ratios, and relative numbers generally must be manipulated by another method than that used on absolute numbers. Rates of speed, or time changes require treatment differing from that required by percentages or place data.

There are various degrees to which data may be analyzed. In some cases a simple summation, or a simple average, is all that is required. In other cases it is necessary to know the scatter of the items around the average of the data. Sometimes one must measure the relationship between two or more variables. These statistics may be computed in terms of the original data, as pounds, bushels, miles, dollars, or any other concrete unit. At other times, however, a ratio or percentage comparison is preferred. A farmer may say that the average yield of his wheat is 20 bushels, and that his poorest acre produced only 7 bushels, while the best acre yielded 31 bushels, and that the total production of his farm was 1500 bushels this year and 1200 last year. These statistics are stated as absolute numbers. He might, however, state his results as follows: (1) the poorest acre produced 35% of the average, (2) the best acre yielded 155% of the average and the total production this year was 25% larger than last year. When two or more variables are included in one study, a still greater number of comparisons are possible and methods of computation available.

The total of statistical methods available at present is the result of the accumulation of half a century or more. Some of the more simple formulas such as summation, arithmetic mean, and ratios are as old as mathematics. Some of the more recent methods of analysis perfected in the past quarter of a century or even during the last decade are much superior to the older methods in both completeness and accuracy. The purpose of the author in pre-

paring this text is to choose from all available methods those which are most simple, fundamental, and useful for beginners in statistics.

A *statistic* is a summary, or representative measure computed from a sample. The average height of grade school children computed from a sample of individual measurements of heights, is a statistic. The measure of the range of grade school children's heights from the shortest to the tallest is a statistic. The average change of the price of wheat from last year to this year computed from a sample of price quotations, is a statistic. Any single computation or measure based on a sample of data and conforming to statistical theory and methods is a statistic. A group of these computations is statistics. The presentation or statement of a statistic or of several statistics for the information and guidance of individuals, corporations, schools, groups, or governments is very important, and in our present-day world, is an extensive business.

Many farmers and small merchants read carefully the price quotations, deliveries of grain and livestock, yields, weather reports, export figures, consumption trends, tables and charts in their trade journals as well as in the daily newspapers. These persons are *consuming* statistics. They are attempting to guide themselves and to rationalize their business activities by statistics. All large corporations, railroads, public utilities, trading and manufacturing concerns depend largely on statistics of one kind or another for management guidance. The United States Government and the several states and most cities are continually preparing and publishing large volumes of statistics. If all price statistics were removed from all papers, magazines, radio and telegraphic reports for a single day, the business world would be paralyzed. If all statistics now available were removed from the world for one year, utter economic chaos and ruin would result. We live in the midst of statistics. We live and direct our lives by means of statistics, from prices to weather reports. All human activities from our simplest daily purchases to the making of national laws, treaties with foreign nations, court decisions, and the prosecution of wars turn largely on statistics.

Statistics as a science in its broader meanings is not only a means of guiding our judgment in making the daily decisions of

life, but also *is a device or tool for discovering new truth.* It is one of the most powerful engines for research. Continually by means of statistics new and important relationships are brought to light. The value of new fertilizers, of new methods of advertising, of new machines, of new chemical processes is measured by statistics. The meaning of hitherto unknown biological, psychological, physical, economic, social, and technical relationships and factors is revealed or clarified by statistics. Those who expect to do effective research work must be well trained in its theory and methods.

Elementary statistical theory and methods present the basic elements which are universal in their application and without which one can go no further in the science. They are as useful and necessary in agriculture, biology, and genetics as they are in economics, sociology, and education.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapters I, II, III, IV. Irwin, Inc., Chicago, 1941.

Chaddock, R. E., *Principles and Methods of Statistics*, Chapters I, II, III. Houghton Mifflin Company, Boston, 1925.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapters I, II. Prentice-Hall, New York, 1948.

Davies, G. R., and Crowder, W. F., *Methods of Statistical Analysis*, Chapters I, II. John Wiley & Sons, New York, 1933.

Holzinger, K. J., *Statistical Methods for Students in Education.* Ginn & Company, Boston, 1928.

Mills, F. C., *Statistical Methods, Revised.* Henry Holt & Company, New York, 1938.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chapter I. Harcourt, Brace & Company, New York, 1943.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics.* McGraw-Hill Book Company, New York, 1951.

Waugh, A. E., *Elements of Statistical Methods*, Chapters I, II, III. McGraw-Hill Book Company, New York, 1938.

REVIEW QUESTIONS

1. What is the difference between a statistic and statistics?
2. Explain in detail the difference between statistical theory and statistical methods.
3. What is the statistical meaning of the proverb, "He can't see the forest for the trees"?
4. What is meant by "consuming statistics"? Who consumes statistics and why?
5. What is the meaning of the statement, "Statistics is the method of analyzing groups of numbers"?
6. How is statistics a tool of research?
7. Does anyone besides statisticians use statistical methods? Explain.
8. What is the difference between the ordinary "day-by-day common sense approach" to the problems of life and the "statistical approach"?

CHAPTER 2

GRAPHIC PRESENTATION

In teaching statistics from the consumer's viewpoint it is logical to place graphic presentation first. The graph, chart or picture is the device which brings the meaning of the statistical analysis and conclusions to the mind of the consumer. It is the language by which he learns to read the ideas of the author of statistical studies and researches. He must learn to read before he can comprehend.

This method also has the advantage that graphic presentation is the easiest part of statistics to teach as well as to learn. The average student can here be challenged by evident simple ideas of which he already has some knowledge. The practical application of statistics to everyday life can be demonstrated. This is especially helpful to those who are less proficient in mathematics, as most consumers of statistics are.

Psychologists long ago discovered that the eye takes in more knowledge than the ear. Display advertising in both newspapers and magazines occupies most of the space in these publications. Our highways are littered with billboards, and television has proved to be more efficient than radio in attracting and holding consumer interest. In presenting graphics early in a statistics course the same superior psychology is employed.

Of all the many available methods of making graphs the six most widely used are presented here. They are:

1. Line curves or graphs
2. Bar charts
3. Band charts
4. Pie diagrams
5. Statistical maps
6. Pictographs

The simplest to understand, the easiest to make, the most adaptable to many uses, and the most widely employed type of chart is the *line graph*. It enables one to present more information of a more complex nature in a perfectly understandable form than any other kind of chart. It requires the least technical skill to make.

BASIC REQUIREMENTS FOR LINE GRAPHS

1. In such graphs, the independent variable is always placed on the X-axis and the dependent variable on the Y-axis. If time is one of the variables, it is almost always considered to be the independent factor and is usually placed on the horizontal axis.

2. Quantity or value is measured on the Y-axis. The scale of the Y-axis should always begin at zero (0), even if the lowest Y-figure associated with any X period or value is far above zero. It is deceptive to begin the Y-scale at a point above zero without clearly indicating this fact as is done in Figs. 1 and 2, in which

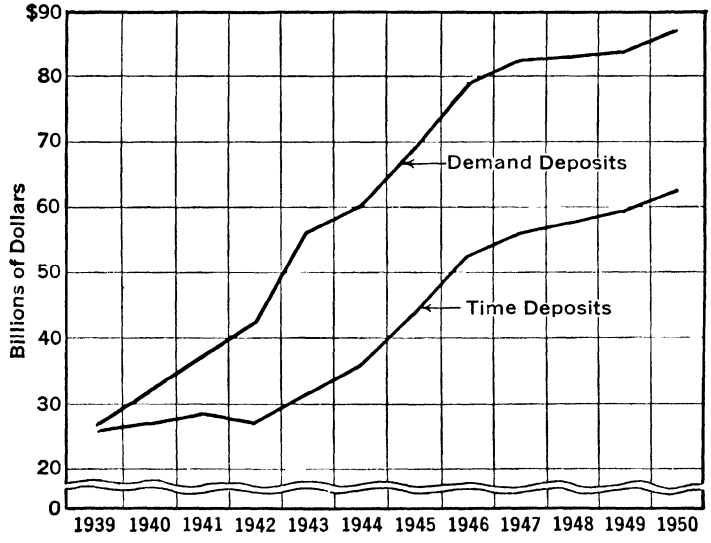


FIG. 1. Total of all demand deposits and all time deposits in all United States Commercial, Savings, Postal, and Reserve Banks. (*Statistical Abstract of U.S.*, 1951)

the broken space indicates that a part of the scale has been omitted. If economy of space does not make it necessary to cut the *Y*-axis short, it is always better to show it in full as is done in Figs. 3 and 4.

3. In a line graph in which time is shown on the *X*-axis, it is necessary to plot the point of value, or datum, for each time period at the *middle of the space* allotted to that time period as is done in Figs. 1, 2, and 3, when the *Y*-values are averages or totals for that period.

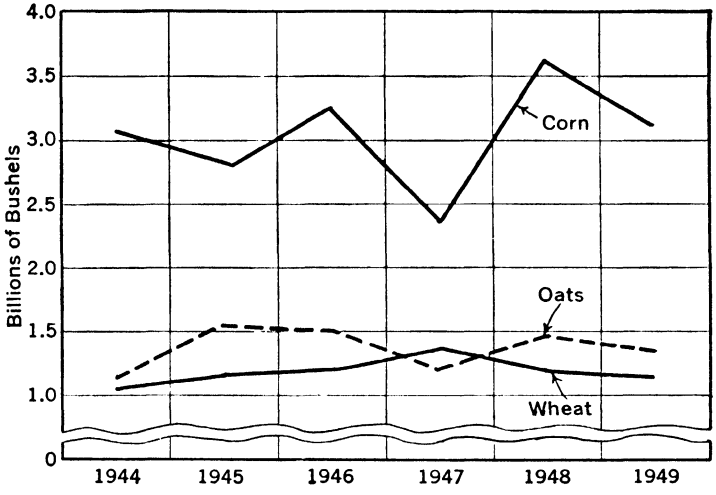


FIG. 2. Production of wheat, oats, and corn in the United States, 1944–1949. (*U.S. Statistical Abstract*, 1951)

4. If the *Y*-quantities are to show cumulative values in which the quantity for each succeeding period is added to the preceding one, the point is plotted at the right-hand edge of the space allotted to the period, as is shown in the cumulative line in Figs. 10 and 42. This procedure is necessary because the entire quantity is not obtained until one reaches the end of the period in which it is being accumulated.

5. All graphs should have a title and number and should give the source of the data. Ordinarily the graphs of a chapter and usually of an entire book or report should be numbered consecutively as is done in this text.

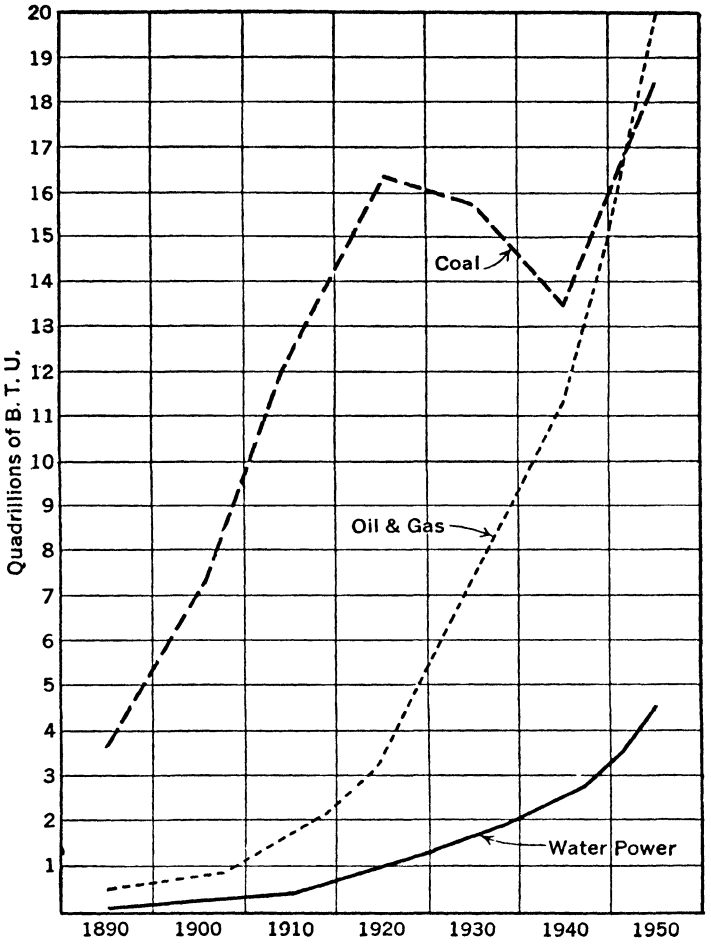


FIG. 3. Energy produced from mineral fuels and water power in the United States from 1890 to 1950. (*U.S. Statistical Abstract*, 1950)

6. If more than one line is shown on a graph, the meaning of each line should be clearly designated by a legend or by writing in the graph close to the line the nature of the data as is done in Figs. 2, 3 and 4.

Figures 4 and 5 should be studied together. Figure 4 shows data on life insurance in the United States for total, ordinary,

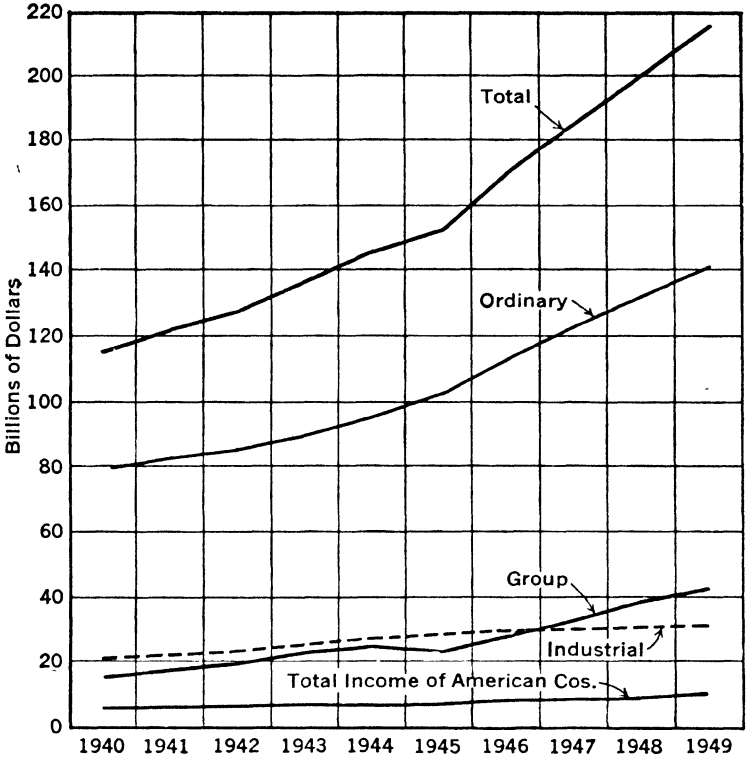


FIG. 4. Total life insurance — ordinary, group, and industrial — in the United States and total income of U.S. insurance companies, 1940–1949, inclusive. Arithmetic scale. (*Life Insurance Fact Book*, 1950, pp. 6, 41)

group, industrial, and the earnings of U.S. companies, all on the natural or absolute scale. On such a scale the larger values such as total and ordinary show wide fluctuations, while the smaller values of group, industrial, and income show little change from year to year. Figure 5 is made on semi-logarithmetic paper. It has the absolute or natural scale on the X-axis for time periods, but has a ratio, percentage, or logarithmic scale on the Y-axis. This type of chart reduces the variations of all the series to the same percentage base. In fact, it shows that group insurance actually has a larger percentage growth and variation than either the total or ordinary. On the log chart the income of the U.S.

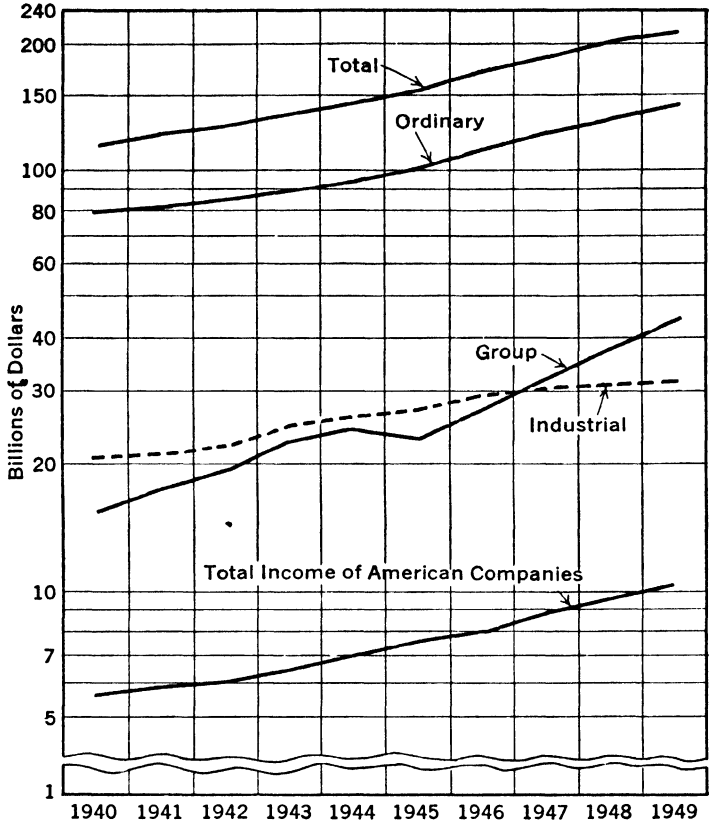


FIG. 5. Total life insurance — ordinary, group, and industrial — in the United States and total income of U.S life insurance companies, 1940-1949, inclusive. Logarithmetic scale. (*Life Insurance Fact Book*, 1950, pp. 6, 41)

companies shows as great or greater rate of growth than that of total insurance while in Fig. 4, it appears as nearly a flat line.

This percentage change cannot be shown on the natural scale. The ratio chart or semi-logarithmetic graph is of great value in the measurement of cyclical and seasonal fluctuations in sales, production, employment, and other business activities which management needs to compare percentage wise. It is especially useful in showing the percentage changes of small and large de-

partments or branches in the same store, factory or corporation, and of revealing the ratio fluctuations of small and large plants in the same industry. It is also useful in political, sociological, psychological and educational data.

Principles of Construction. 1. *Normal or absolute scale charts give* EQUAL SPACES TO EQUAL ABSOLUTE *values.* This means that an increase or decrease of 10 would receive the same space on the graph whether the 10 were added or subtracted to 100 or 1000 or even 10,000. If the chart were drawn to give 10, a space of .1 of an inch, it would receive this space regardless of whether it fell at the top or bottom of the chart.

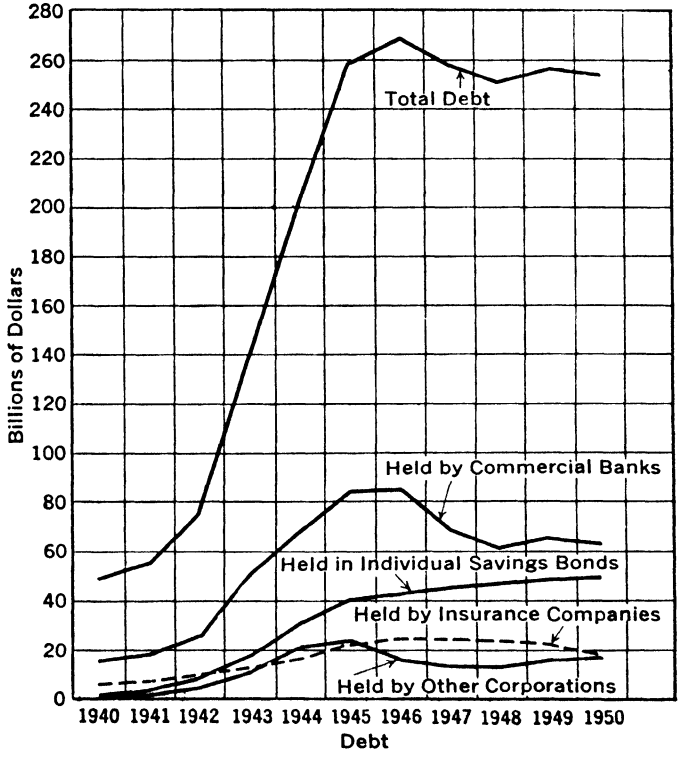


FIG. 6. Total United States and principal holders, 1940–1950. (*Federal Reserve Bulletin*, July 1951. Arithmetic scale)

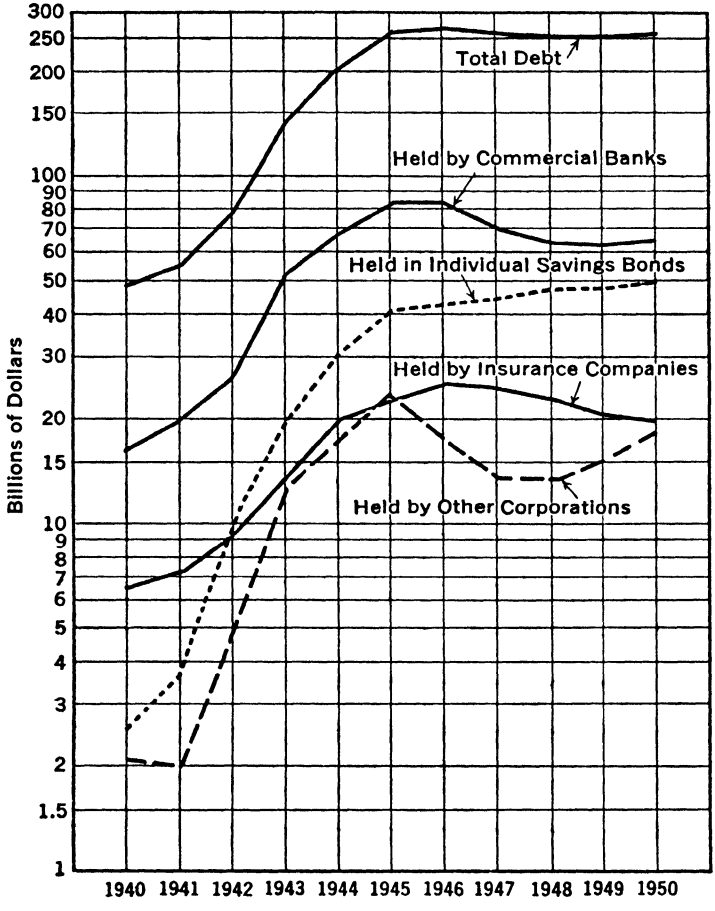


FIG. 7. Total U.S. debt and principal holders, 1940–1950. Logarithmic scale. (*Federal Reserve Bulletin*)

2. *Ratio or semi-logarithmic charts* GIVE EQUAL SPACES TO EQUAL PERCENTAGE CHANGES, regardless of where they fall on the graph. This means that a 10% increase in quantity is given the same space on the chart whether it occurs at the bottom or at the top. If a 100% increase is allotted one inch of space on a chart, a 50% increase receives one-half inch, and a 10% increase, one-tenth of an inch. Ten added to 10 is a 100% increase, but 10 added to 20 is only a 50% increase and would receive only one-

half as much space, while 10 added to 100 is only a 10% change and would receive only one-tenth as much space. This means that the higher one goes on the chart, the larger must be the absolute increase to make the same percentage increase. At the bottom of the chart 5 added to 10 is a 50% increase, but higher on the graph, 500 must be added to 1000 to make a 50% change, and still higher, 50,000 must be added to 100,000 to make the same percentage change. *Equal percentage changes receive equal space on a ratio chart.*

The scale runs from the bottom to the top of the chart on page 18.

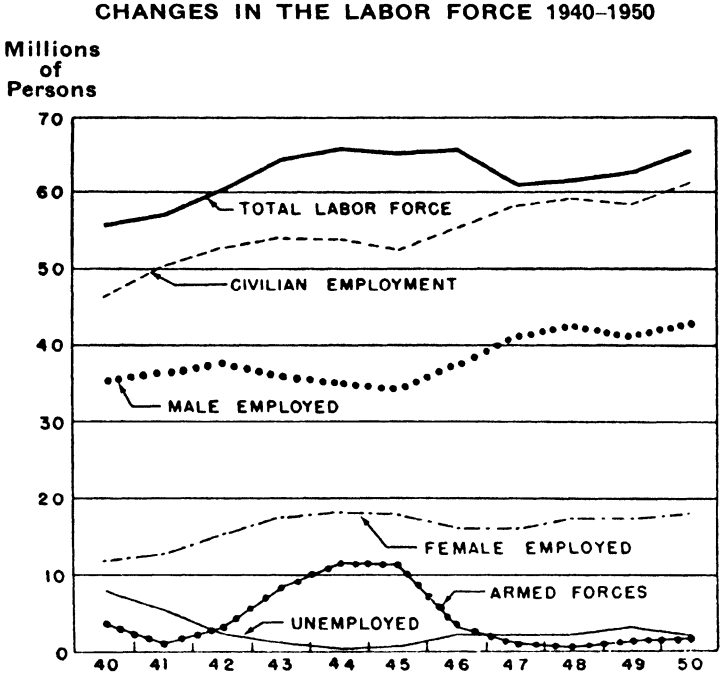
Scale	Percentage Changes	
100	11⅑	One added to 1 is a 100% increase.
90	12.5	Ten added to 10 is a 100% increase.
80	14⅐	Both have the same space on the chart.
70	16⅔	
60	20	One added to 2 is a 50% increase.
50	25	Ten added to 20 is a 50% increase.
40	33⅓	Both changes have the same space on
30	50	the chart which is one-half of the space
20	100	given to 100% changes.
10	11⅑	
9	12.5	All the other spaces are in propor-
8	14⅐	tion to the size of the percentage
7	16⅔	changes.
6	20	
5	25	
4	33⅓	
3	50	
2	100	
1		

It should be noted that on a line chart the several lines showing the several variables may cross each other without giving the slightest confusion to the reader if the adjacent lines are drawn in a different script or color as is shown in Fig. 7. It is preferable to draw solid lines unless they touch or cross at an angle which would cause the reader to lose their identity. In such cases the following variations may be used to insure clarity:

1. ————————————————
2. — — — — — — —
3. - - - - - - - - - - - - - - - - - - -
4. — - — - — - — - — - —
5. . . . . .

The student can readily devise other types of lines if more variation is necessary.

Figures 6 and 7, showing the Total Debt of the United States up to February 1951 and the principal owners of U.S. securities indicate the relative advantages of the two types of line graphs.



Source: Department of Commerce

FIG. 8. Changes in the labor force in the United States, 1940–1950. (From *Business and Real Estate Trends*, a quarterly letter by Home Federal Savings and Loan Association, Tulsa, Oklahoma, and used with their permission)

Percentage, ratio, and semi-logarithmic graphs at the present time are widely used by businessmen and research workers in the fields of economics, political science, sociology, psychology, and education to make ratio comparisons among variables of wide absolute ranges. Physicists, chemists, and engineers find them of equal value.

The student should become thoroughly familiar with their construction and use. It is often more important to know whether sales have increased or decreased 10% than it is to know that the change is $5000. What percentage increase in sales resulted from an advertising expenditure of $10,000, or of increasing the advertising budget 10%? What has been the percentage decrease in the death rate? What per cent of freshmen finally graduate from college? We live by means of comparison in a world of variables, and the rate of change is an essential part of the information necessary to wise decisions. The semi-logarithmic chart is a ready device to this end.

In plotting data on semi-logarithmic charts such as Figs. 5 and 7, it is necessary to begin the Y-scale with the numeral 1 at the bottom line, or with 10, or 100, or 1000, or one million or ten million or any number which consists of 1 by itself or 1 followed by one or more zeros. The beginning number of the Y-scale on the second bank must be ten times as large as the first number on the first bank. If the first bank begins with 10, the second bank must start with 100 and the third bank with 1000. This ratio must always prevail on the Y-scale of a semi-logarithmic chart.

The slope of a line as well as its location is significant on a ratio chart. In Fig. 9, line A indicates an increase at a constant rate. Line B indicates a decrease at a constant rate. Line C indicates an increase at an increasing rate. Line D shows a decrease at an increasing rate. Line E shows an increase at a decreasing rate, but line F indicates a decrease at a decreasing rate. The slope and direction of the line indicates the nature of the change and its steepness, the amount or size of the change, that is the rate of change. *The steeper the line is, the higher or larger the rate of change is.*

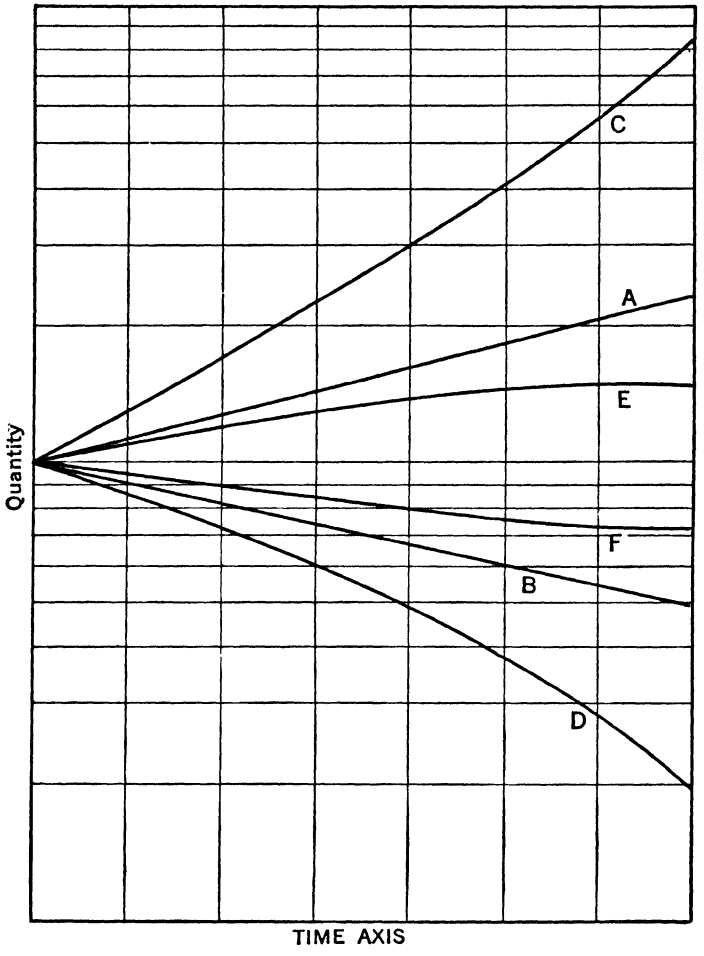


FIG. 9. Showing various rates of increasing and decreasing changes on semi-logarithmic scale

Figure 10 is an illustration of the *Z*-chart which is designed to show three phases of a business simultaneously with different scales on the two ends of the chart. The top line shows a 12-months' moving total of production and the diagonal line a cumulative monthly production, both measured on the left-hand *Y*-scale. The lower line measures the actual monthly production on the

right-hand *Y*-scale. These three quantities essential in marketing and production management may be plotted simultaneously from month to month through the year.

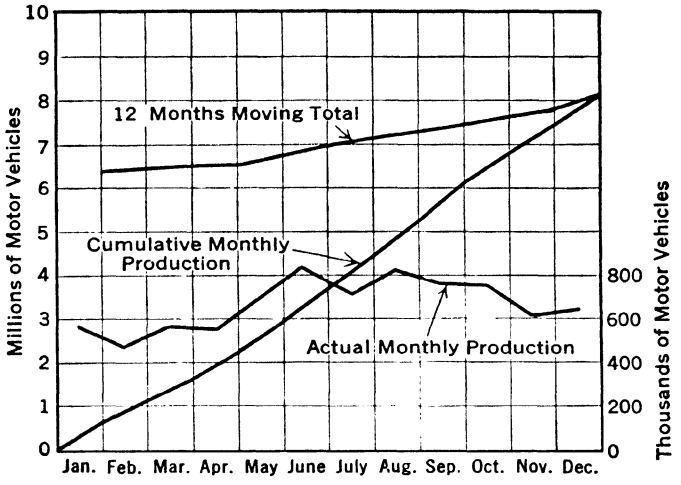


FIG. 10 Actual monthly production, cumulative monthly production, and twelve months' moving total of motor vehicle production in U.S., 1950. (*Survey of Current Business*)

Another highly specialized type of line chart is the Lorenz Curve which is used to measure the degree of the uniformity of scatter of data over an area. Both variables are reduced to percentages and their associated percentage values plotted on the graph. The curved line joining the plotted points indicates the distribution. The line of uniform distribution is a straight diagonal line extending from zero to 100%. The greater the curvature of the actual distribution is the more uneven it is. Fifty percent of the population of New Jersey is crowded into twelve percent of its area while, at the other end of the scale another ten percent of its population occupies sixty percent of its area.

This type of chart is quite useful in marketing, delivery, production, wealth, population, and related economic, educational, and social studies. Many other specialized types of line charts

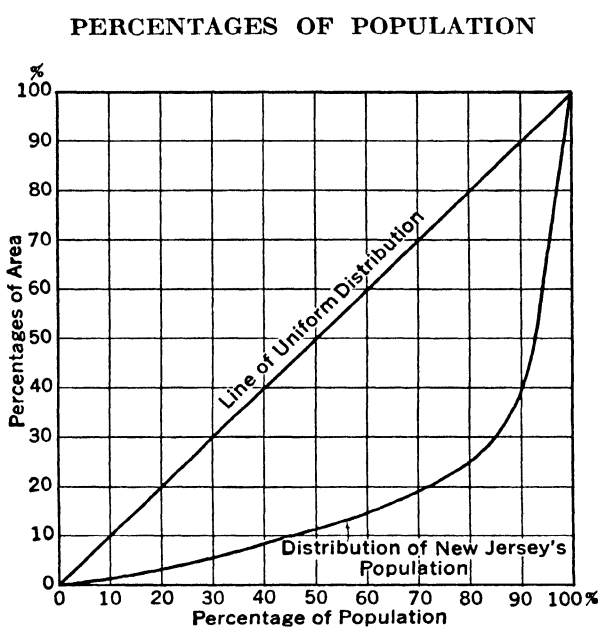


FIG. 11. Lorenz Curve. Uniformity of density of population in New Jersey in 1950. Source: U.S. Decennial Census, 1950

have been devised and may be found in the sources indicated at the end of this chapter, but these are the kinds most widely used.

BAR CHARTS

The second most widely used type of graph is the bar chart. It is more striking to the eye and arrests the attention of the reader with more force than the line graph, but it is much more limited in its possible variations and in the load of data it can carry without confusing the reader. Bar charts should preferably be limited to single bars for purposes of clarity as is shown in Fig. 12, giving the population of leading Texas cities for 1950. In this chart there is one bar for each city. The data given are quite limited, but their relative sizes are so strikingly clear that the complete idea is obtained at a glance. Such graphs are especially effective for the popular presentation of simple relationship.

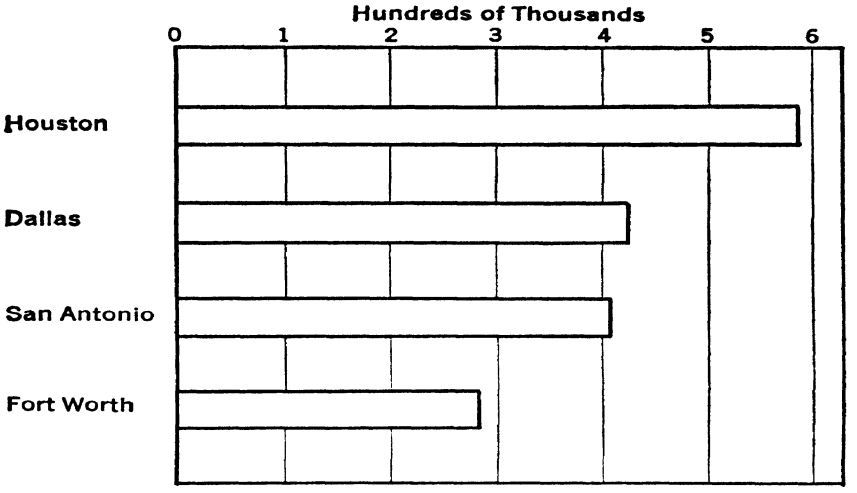


FIG. 12. Population of four principal Texas cities, 1950. Single bar graph. (*U.S. Decennial Census*, 1950)

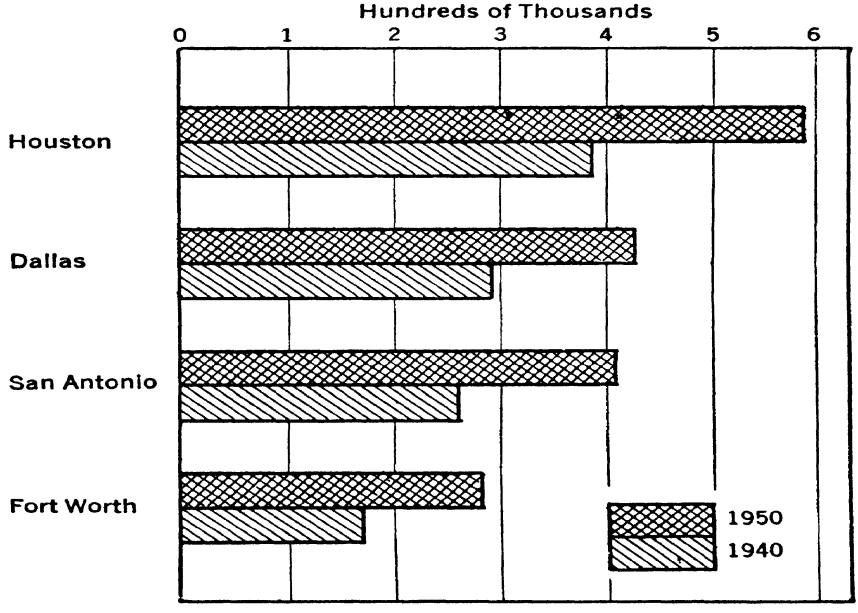


FIG. 13. Population of four principal Texas cities, 1950 and 1940. Double bar graph. (*U.S. Decennial Census*, 1950)

In Fig. 13 the double bar chart is employed. It makes a comparison for each variable between two dates or areas. This chart shows the population of leading Texas cities for the two census years of 1950 and 1940. The double bar chart is nearly as easy to read as is the single bar and it has the advantage of presenting twice as much data. The two sections of the double bars should be so clearly distinguished that no confusion can result. Multiple bar charts require a legend or other designation to indicate clearly the meaning of the several bar segments.

In Fig. 14, the triple bar chart is used. It is much more difficult to read than the single or double bar, but it gives three comparisons. If it is used, the segments of the bars should be so clearly

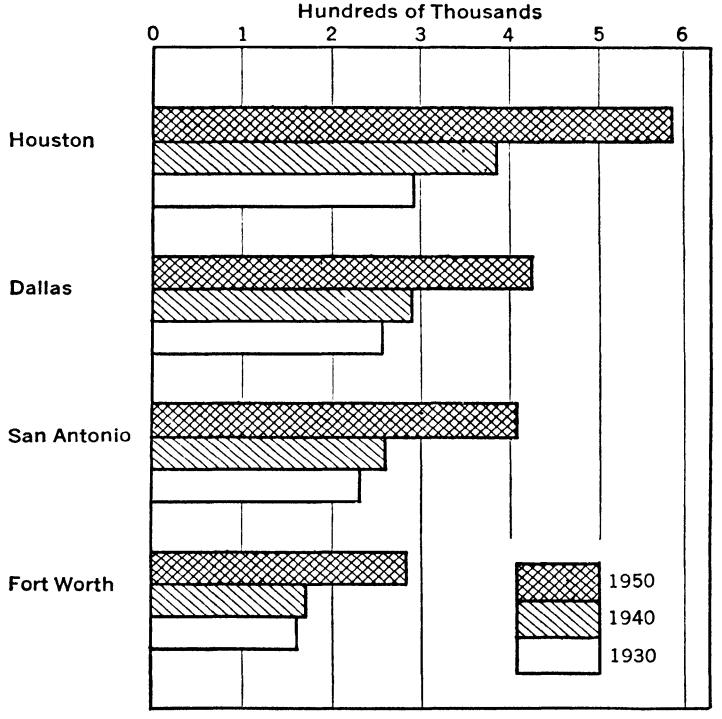


FIG. 14. Population of four principal Texas cities, 1950, 1940, and 1930. Triple bar graph. (*U.S. Decennial Census*, 1950)

distinguishable in cross-hatch or color that they can be read quickly and that no confusion will result.

In any ordinary circumstance a triple bar is as complex a bar chart as should ever be used. Bars with four or five segments result in more confusion than they do in clarity. If there are more than two or three variables to present, they should be charted in a line graph or in separate single bar charts.

Time is sometimes placed on the *Y*-axis in bar charts as is done in Figs. 12, 13, and 14, but usually if time is to be emphasized as an independent variable, the chart becomes a *vertical* bar chart with time on the *X*-axis. Figure 15, measuring the Finished Stocks of Gasoline on Hand in the U.S. 1939–1951, is such a chart, designed to show the fluctuations of Gasoline Stocks through succeeding periods of prosperity, depression and war activity. This information could more easily be shown on a line graph, but the bar chart presents the facts in a more glaring and forceful form. Such a graph is an excellent way to emphasize a single important time series.

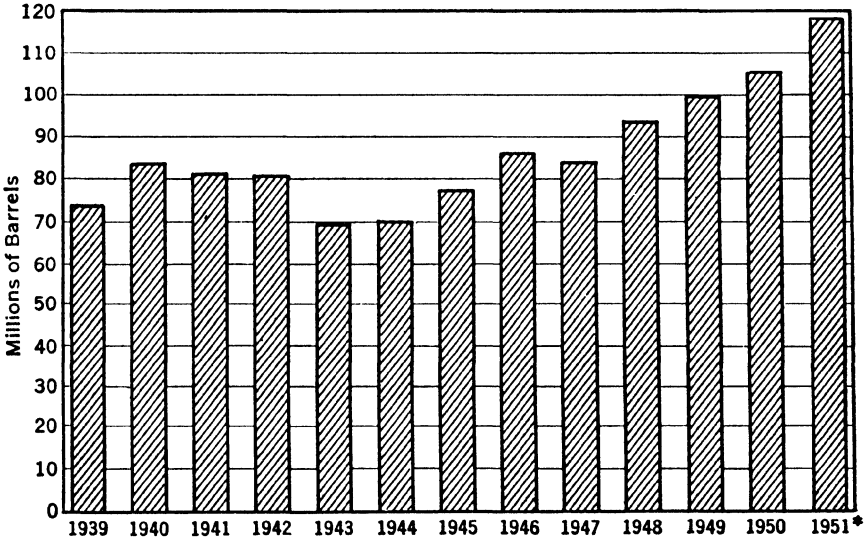


FIG. 15. Average stocks of finished gasoline on hand in the United States, 1939–1951, inclusive. (*Survey of Current Business*)

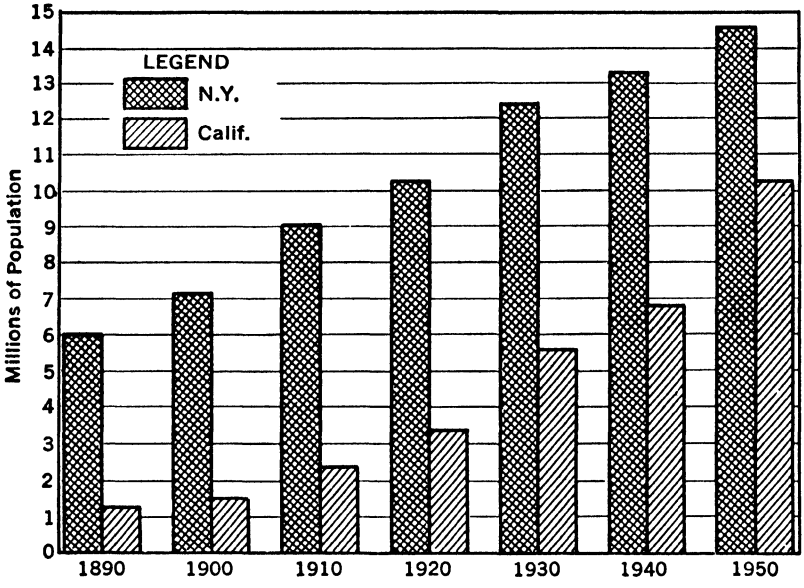


FIG. 16. Population growth of New York and California, 1890–1950, inclusive. (*U.S. Decennial Census*, 1950)

The double vertical bar chart is frequently used to compare two closely related variables through the same time periods. Figure 16, showing the population of New York State and California from 1890 to 1950 inclusive, is a good example of this type of chart. These data could be shown on a line graph, but the double bars emphasize the relationships more clearly. The recent rapid growth of California is emphasized. It is not advisable to make vertical bar charts more complex than double bar graphs. Triple and quadruple bars are too confusing to the reader to be practical. For more than two variables on the same time graph, line charts should be used.

For pure scientific analysis and in all other cases in which no psychological appeal is desired, the simple line graph is preferable. But in all cases in which the mind of the reader must be struck or challenged with an idea, the bar chart is more effective.

In making both vertical and horizontal bar charts the number of guide lines from the quantity scale across the graph should be

only sufficiently numerous to give the approximate length of the bars. It spoils a bar chart to have it covered with guide lines so close together that they obscure the bars. *The bars must stand out,* or the bar chart loses its force and value. In the horizontal bar chart, if the lower bars are shorter than those above, the guide lines do not need to be extended to the bottom of the graph at the lower right-hand corner.

In making bar graphs the numbers represented by the bars should never be written at the top of vertical bars or at the right-hand end of horizontal bars, as is shown in Fig. 19. Such a practice obscures the length of the bars, especially if they are slender, and deceives the reader as to their actual and relative length. The numbers or data represented by the bars should be written either,

1. inside the hollow bar as in Fig. 17,
2. at the base end of the bar on the opposite side of the base line, as in Fig. 18.
3. or in adjoining table, but never as shown in Fig. 19.

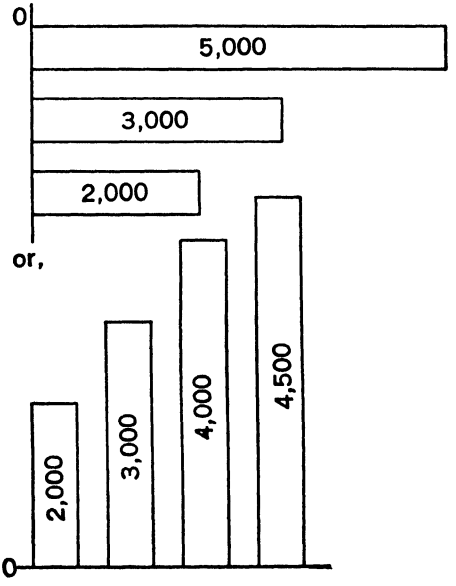


FIG. 17. Correct forms for bar charts

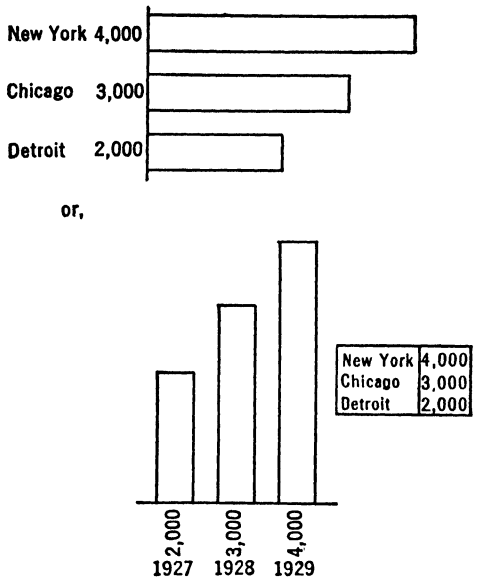


FIG. 18. Correct forms for bar charts

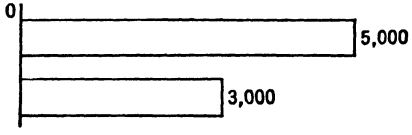


FIG. 19. Incorrect forms for bar charts

FREQUENCY POLYGONS AND HISTOGRAMS

Frequency distributions are best pictured in either frequency polygons or in histograms. These graphs are explained in Chapter 5, pages 105–6. Figures 38, 39 are frequency polygons showing the same information as the histograms in Figs. 36, 37. The student should note carefully the method of construction used in these graphs and their special meaning in statistical analysis. They measure varying quantities at an instant, point or period of time. They are useful in indicating how far samples deviate from the normal curve of error and in making comparisons of these deviations.

BAND CHARTS

A band chart is a type of line graph which shows the total for successive times periods broken up into subtotals for each of the component parts of the total. The spaces between the successive lines are filled in with cross-hatch or color to emphasize the quantity of each factor for each time period. This type of graph is especially useful in dividing total costs into component costs, total sales into department or district or individual salesman's

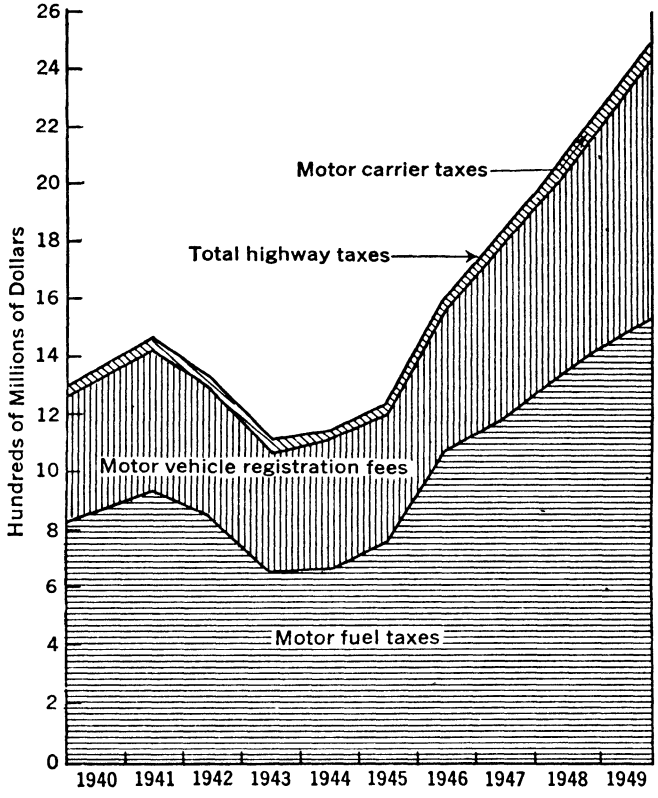


FIG. 20. Total highway user revenue and its subdivisions, 1940–1949 inclusive. (*Tax Economics Bulletin,* American Petroleum Institute, Vol. XV, No. 6, Nov.–Dec., 1950. Used with permission of the American Petroleum Institute)

sales, total production by states, plants or industries and other such relations.

The lower band measures the amount of taxes paid on all motor fuels, gasoline, etc. The second band shows the taxes paid for registering motor vehicles. The third narrow band next to the total line represents the taxes paid by motor carriers.

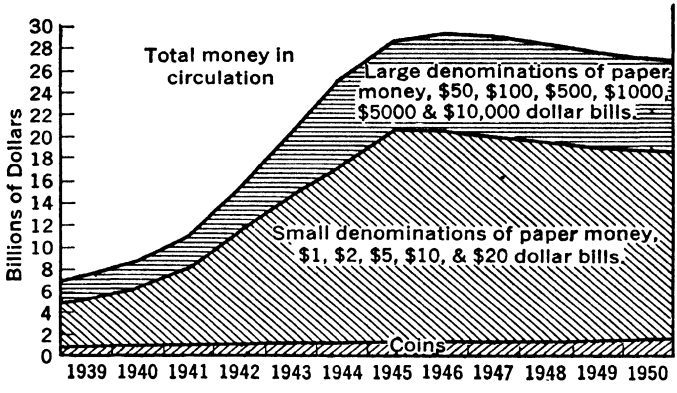


FIG. 21. Total money in circulation and quantities of principal types of money in the United States, 1939–1950. (*Federal Reserve Bulletin*)

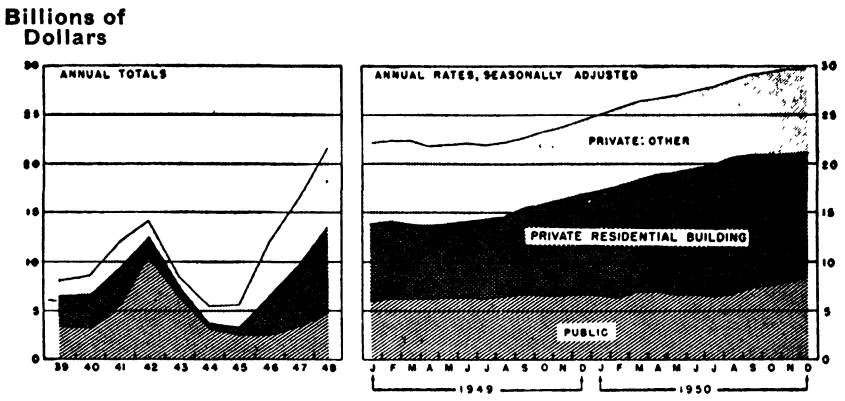
In Fig. 21, is indicated the amounts of the various types of money in circulation in the U.S. and the total from 1939 to 1950 inclusive. The lower band shows the total of all coins. The middle band measures the total of all paper money of small denominations, $1 to $20 bills inclusive. The top band shows the total of large denominations, $50 to $10,000 bills inclusive. The top line is the grand total.

These two graphs are sufficient to indicate how a series of subtotals can be emphasized in a band chart along with their grand total. It is much more impressive than a mere line graph.

Figure 22 has the advantage of breaking the details of the last two months down into monthly comparisons.

PIE DIAGRAMS

A pie chart is a device for reducing a percentage total of 100 to its component parts. The diagram consists of a circle, the cir-



Sources: Departments of Commerce and Labor

FIG. 22. Value of construction put in place, by type, 1939–1950. (From *Business and Real Estate Trends*, a quarterly letter by Home Federal Savings and Loan Association, Tulsa, Oklahoma, and used with their permission)

cumference of which is divided into 100 parts, each representing 1.%. Each component part of the total is given its percentage part of the total. This type of table is useful whenever there is need for picturing the relative size of the several parts of a whole. Its adaptations are almost limitless. It may show the total sales by departments, states, districts, salesmen or branches; production by departments, plants, states or nations; total costs by items, etc.

In making percentage pie charts the following points should be carefully observed:

1. Divide the circumference of the circle into 100 equal spaces, each representing 1.%.
2. Reduce the data to be plotted to percentages for each class based on the total of all classes as 100%.
3. Locate zero (0) percentage at the top of the circle chart and measure to the right for increasing percentages, coming around clockwise, to the top again for 100%.
4. Write in each segment of the chart its class name or indicate the meaning of each segment in an accompanying legend.
5. For comparisons between two or more time periods, or

between two or more areas or classes, use separate charts for each one.

6. The largest item in each series should be placed first and the others should follow in declining order of size, with the smallest or "All others" last.

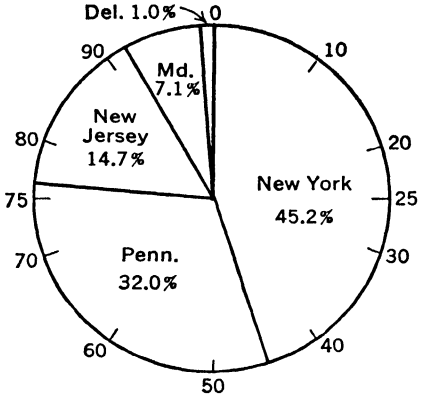


FIG. 23. Population of Middle Atlantic States as percentages of total population of all five states, 1950. (*U.S. Decennial Census*, 1950)

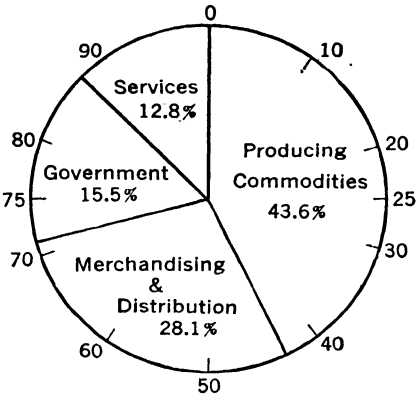


FIG. 24. Percentage distribution of total non-farm wages paid in U.S., 1950. (*Federal Reserve Bulletin*, February, 1951, p. 217)

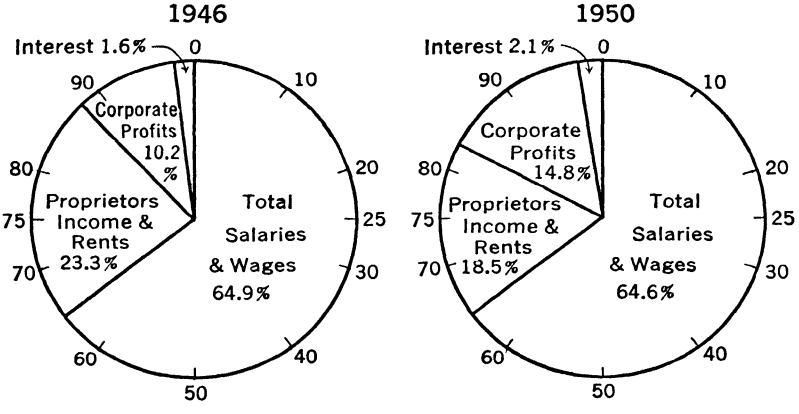
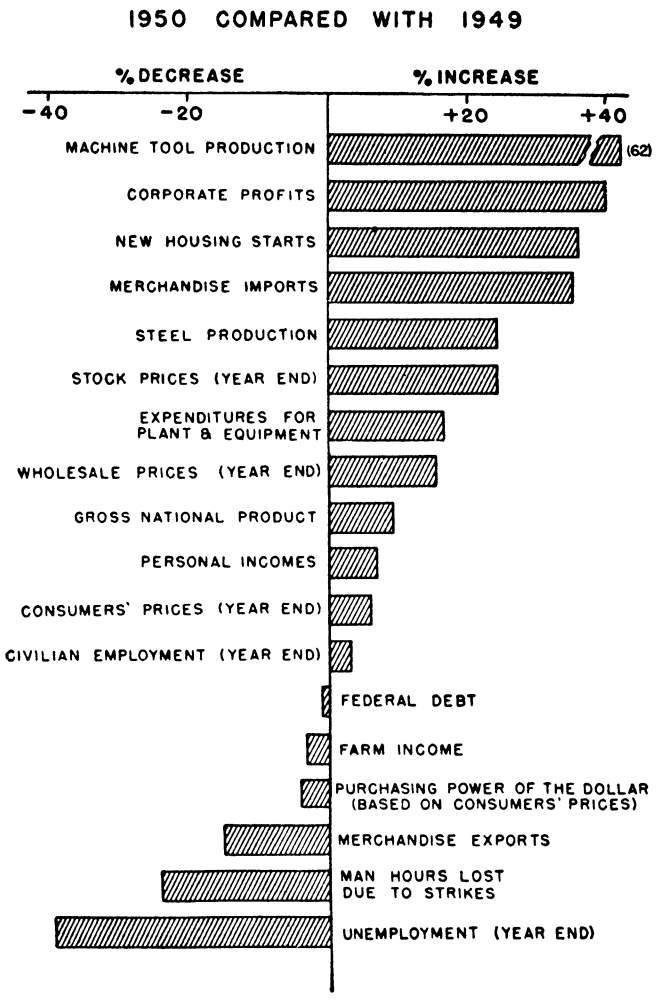


FIG. 25. Percentage distribution of national income among the major income classes: (1) total salaries and wages, (2) proprietors and rental income, (3) corporate profits, and (4) interest, for 1946 and 1950. (*Federal Reserve Bulletin*, February, 1951, p. 216)

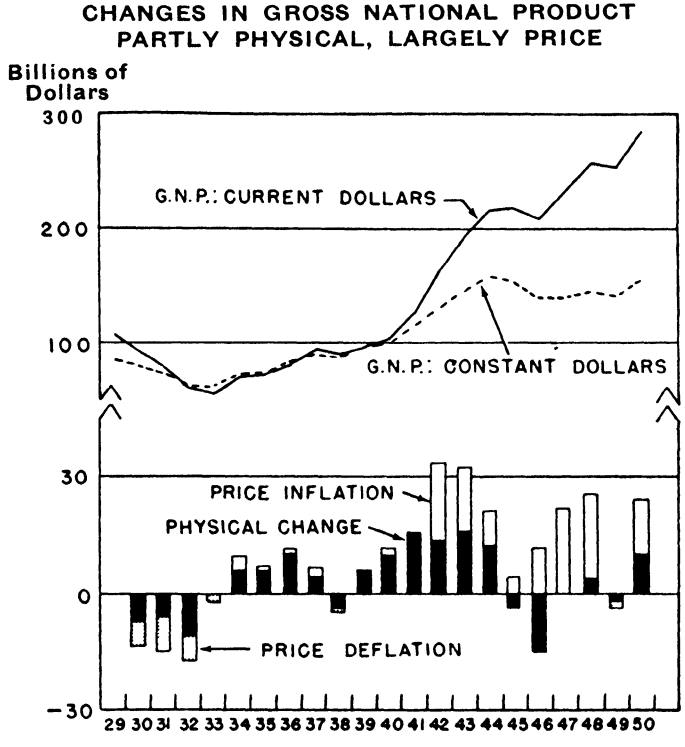


Sources: *Survey of Current Business* and *Federal Reserve Bulletin*

FIG. 26. Percentage increase or decrease in volume of principal economic activities between 1949 and 1950. (Copyright, 1951, *Business and Real Estate Trends*, a quarterly letter by Home Federal Savings and Loan Association, Tulsa, Oklahoma, and used with their permission)

PLUS AND MINUS CHARTS

Business and other human activities may be presented in charts which measure up and down, or plus and minus from a zero line or point. The comparisons may be made in either absolute or percentage forms. Figures 26 and 27 indicate the possibilities of this widely used type of graphic presentation. This type of chart is especially useful in measuring profit or loss, inventory and similar changes from one period or place to another.



Sources: *Survey of Current Business* and *Federal Reserve Bulletin*

FIG. 27. Changes in national product, partly physical, largely price. (From *Business and Real Estate Trends*, a quarterly letter by Home Federal Savings and Loan Association, Tulsa, Oklahoma, and used with their permission)

STATISTICAL MAPS

Most data bear some relationship to space. They may be identified with some area or location. The statistical map is a widely used method for the presentation of such data. The maps may be made by the (1) cross-hatch method or (2) the dot method. Cross-hatch maps are usually employed to make comparisons of averages or totals between relatively large areas.

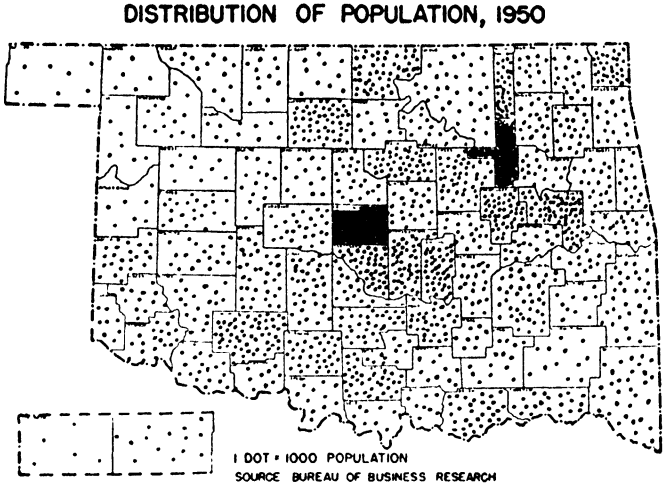


FIG. 28. Small dot map showing distribution of population of Oklahoma in 1950. (*Oklahoma Business Bulletin*, Bureau of Business Research, University of Oklahoma, April, 1951)

Cross-Hatch Maps. The error in such maps is that a large area, a state or county for instance, is all given the same density or rating while there may actually be a great deal of variation within its boundaries. Kansas and Nebraska both have a much more dense population in the east and a very sparse settlement in the west. The same disparity occurs in California, Arizona, and even in New York. Cross-hatch maps should always be understood as comparing average or total figures between areas without regard to the internal variations of the areas.

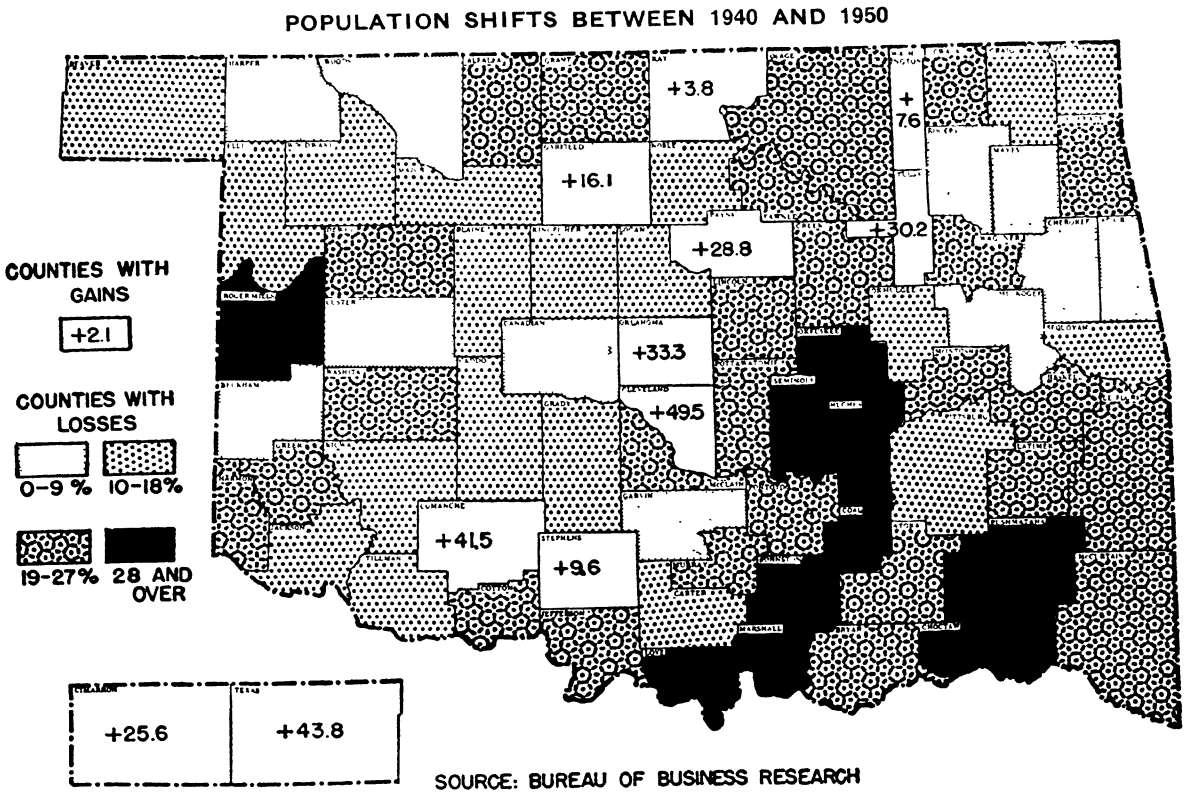


FIG. 29. Cross-hatch map of population shifts in Oklahoma from 1940–1950. (*Oklahoma Business Bulletin*, Bureau of Business Research, University of Oklahoma, April, 1951)

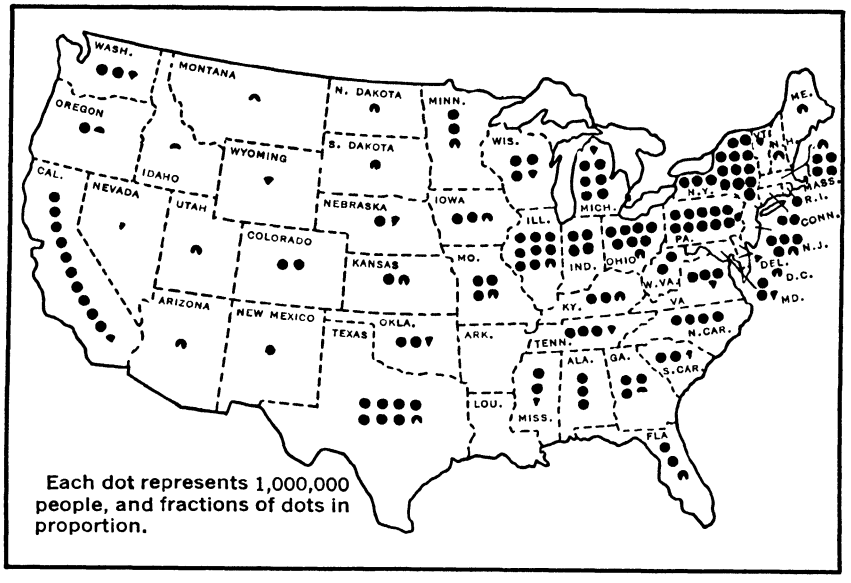


FIG. 30. Density of population of the U.S. showing the amount of population in each state without exactly locating the data. (*U.S. Decennial Census*, 1950)

Small Dot Maps. The type of dot map shown in Fig. 28 is useful in measuring the variations or gradations of any kind in any area, large or small. Population, sales, wealth, production, deaths, automobiles, or any other continuous variable may be so measured. The small dots are placed as near as possible in the exact location on the map where the population or activity exists. Such maps are excellent for detailed analysis.

Rules for Making Small Dot Maps. 1. Decide how many units of data should be represented by one dot.

2. Have the number of units of data equal to one dot small enough that the dots will just cover the most dense portion of the map solidly.

3. Scatter the dots promiscuously over the smallest unit for which the data are given. The smaller the geographical unit for which exact data can be obtained is, the smoother the distribution of the dots over the map will be. Minor civil divisions are better

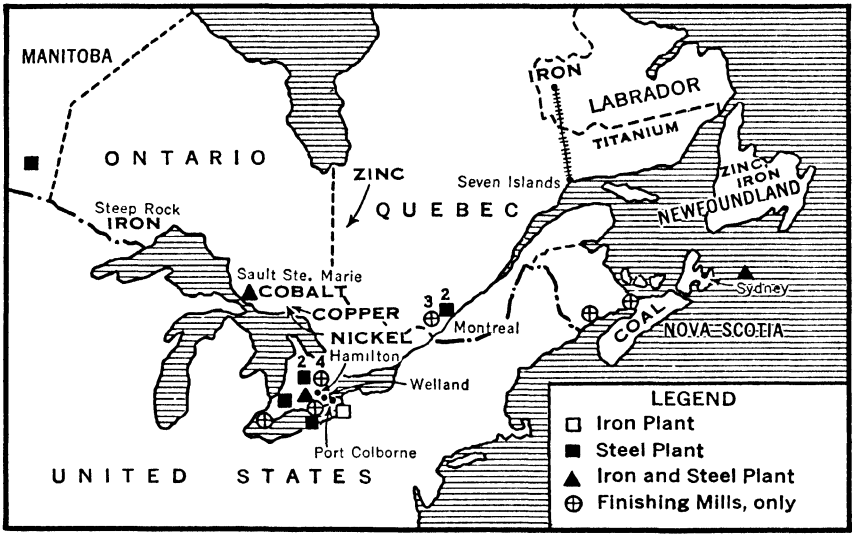


FIG. 31. Pin-head map used to locate specific plants, offices, marketing and manufacturing areas or cities, mines, or transportation lines. (*Steel Facts*, No. 107, April, 1951, p. 3)

than counties, but frequently data can be obtained only for counties.

Large Dot Maps. Large dot maps merely indicate the amount of population or other data in an area without exactly locating them in the area. Figure 30 shows that Texas has approximately 7.7 millions of people but does not show that more than three-fourths of these are located in the eastern half of the state. The map of Arizona does not show that one-third of the states people live in the Salt River Valley. Such maps are useful only for making large area or total comparisons. Frequently this is the only type of comparison needed. Unless some detailed use is to be made of county, city, or minor political subdivision data, it is a waste of time to make the more difficult and expensive small dot maps.

Pin Head Maps. The pin head map constructed by sticking thumb tacks or pins with heads of different colors in specific

locations on a map to indicate some specific activity or data, is a very useful device for business management purposes or any other organizational controls.

The methods of graphic presentation listed above include all the variations in charts likely to be used by the average statistician or ordinary business, educational or political organization. If the student should need other variations of these basic methods or more complex patterns, he should consult the excellent and complete work, *Graphic Presentation* by Willard C. Brinton, published by Brinton Associates, New York City, 1939, or one of the references at the end of this chapter.

Graphic presentation has become a fine art and deserves the careful study of all students of statistics whose research has produced something worthy of presentation to the public. It is one of the principal devices by which the public and most college students consume statistics. It has also a wide use in business analysis and advertising. Even the mathematical research statistician finds it useful in clarifying his own ideas for further investigations.

PICTOGRAPHS

As the name indicates a pictograph is a combination of a graph and a picture, or literally picture writing. It is a graph made up of a number of pictures.

There has been a rapid expansion in their use in recent years. They are widely employed in advertising, in the daily press, in popular magazines, and in books designed for popular appeal. They add little or nothing to rigorous statistical analysis but have educational value and popular appeal in the presentation of statistical analysis. In increasing competition with television, video, the press will, perhaps, be compelled to use more pictographs, to sustain their competitive position. In the more serious and scientific statistical studies, they should be used conservatively, if at all; but where popular appeal is desired, they are effective. They are especially useful in popular educational programs in grade and high schools and in economic education for large groups of employees.

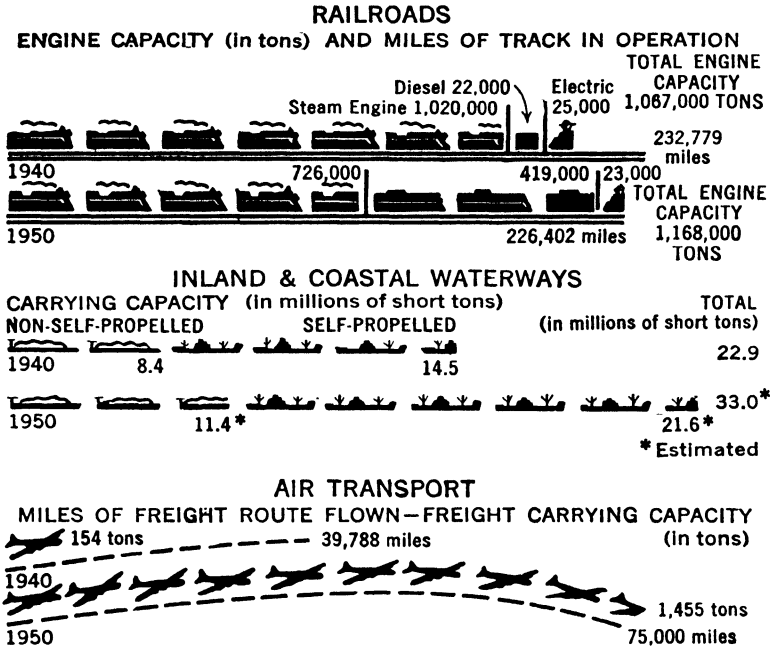


FIG. 32. Changes in carrier's equipment and capacity, railway engines, inland and coastal waterways, and air transport, 1940–1950. (Dun & Bradstreet, Inc. Used with permission)

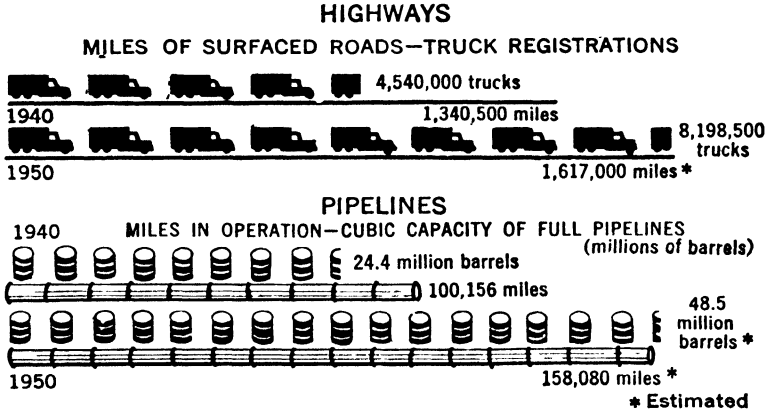


FIG. 33. Changes in carrier's equipment and capacity, highways and pipelines, 1940–1950. (Dun & Bradstreet, Inc. Used with permission)

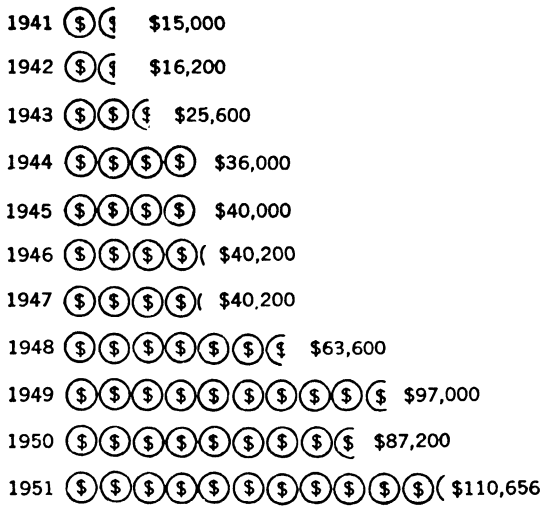


FIG. 34. New membership income of American Management Association, 1941–1951. (*Annual Report of American Management Association*, 1950–1951, p. 14)

SUMMARY

1. Graphic presentation is the use of lines, bars, diagrams, maps, or pictures to illustrate the meaning and relationships revealed in the analysis of data.

2. On the arithmetic or absolute scale, equal absolute numerical values are given equal spaces.

3. On ratio or logarithmic scales equal spaces are given to equal percentage values.

4. Line graphs can carry more complex relationships or data without confusion than other types of charts.

5. Bar charts should never contain more than triple compound bars and preferably never more than double bars. More than this leads to confusion.

6. Pie or circle charts are usually employed to divide a total value into its percentage subdivisions.

7. Pictographs or statistical graphs composed of pictures are employed more for popular appeal than for scientific accuracy and are of special value in advertising, popular educational appeals or propaganda. The line graph is most widely employed in strictly scientific analyses.

8. If numerical values are employed in bar charts they should never be written at the top or right-hand end of the bars. To do so tends to over magnify the length of the shorter bars.

9. The scale on the *Y*-axis should always begin at a base of zero (0) in order to prevent distortion of the quantities presented in the graph.

10. In making small dot maps the area of greatest density should appear as a solid patch of dots, the dots touching each other, in order to serve as an accurate base for comparison with less dense areas.

11. Large dot maps do not give detailed locations of the data, but only the comparative quantities among larger areas.

12. The type of chart or figure used should be determined by the nature of the data, the type of analysis, and the persons to whom appeal is made.

SELECTED REFERENCES

Brinton, W. C., *Graphic Presentation.* Brinton Associates, New York, 1939.

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics,* Chapter XIII. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics,* Chapters IV, V, VI. Prentice-Hall, Inc., New York, 1948.

Ezekiel, M., *Methods of Correlation Analysis.* John Wiley & Sons, Inc., New York, 1941.

Mills, F. C., *Statistical Methods, Revised.* Henry Holt and Company, New York, 1938.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics,* Chapters V, VI. McGraw-Hill Book Company, New York, 1951.

REVIEW QUESTIONS

1. Why is graphic presentation necessary?

2. Why should a graph always begin its quantity scale at zero? Explain.

3. What is meant by the "natural or absolute scale"? For what should it be used?

4. What is the "percentage or ratio scale"?

5. What is "semi-logarithmic" paper? For what is it used? Explain.

6. Compare the equality as spaces on a "log-chart" with a "natural-scale" chart.

7. What are the principal advantages of the line graph over other charts? Explain in detail.

8. What are the principal advantages and limitations of the bar chart?

9. Where should "time" usually be placed on a graph? Why?

10. What are the principal uses of the "Z-chart"? Explain in detail.

11. Explain the uses of the "band chart." Give illustrations of its uses.

12. Explain the proper construction of a "pie chart" and illustrate its common uses.

13. Name four kinds of statistical maps and explain the uses and limitations of each one.

14. Of what use is graphic presentation to the research statistician?

15. What are "histograms" and "frequency polygons"? For what are they used? How are they constructed?

16. What is the purpose of graphic presentation in business statistics? Explain in detail.

EXERCISES

I. 1. Make one line graph on the arithmetic scale and another on the logarithmic scale for the following data on the population growth of the United States and its three principal sections. Source: 1950 census (data in 1000's).

Census	United States	North	South	West
1890	62,947	39,817	20,028	3,102
1900	75,994	47,379	24,523	4,091
1910	91,972	55,757	29,389	6,826
1920	105,710	63,681	33,126	8,903
1930	122,775	73,021	37,858	11,896
1940	131,669	76,120	41,666	13,883
1950	150,697	83,939	47,197	19,561

2. Make a horizontal single-bar chart of the population of the three sections, North, South, and West for 1950.

3. Make a vertical double-bar chart of the population for North and South for 1920, 1930, 1940, and 1950. (South and West may be used instead.)

4. Make a pie-percentage chart of the three areas, North, South, and West for 1950 (or any other year). (The teacher may vary the combinations used from semester to semester.)

II. Population of Principal American Cities (1000).

City	1950	1940	1930	1920
New York	7,835	7,455	6,930	5,620
Chicago	3,606	3,367	3,376	2,702
Philadelphia	2,064	1,931	1,950	1,824
Los Angeles	1,958	1,504	1,238	577
Detroit	1,838	1,623	1,568	994

Make the same charts for the cities that were made for the United States and its sections.

CHAPTER 3

TABULATION

Tabulation in its broadest sense *is any orderly arrangement of data in columns and rows.* This definition includes worksheets. A worksheet, however, should be distinguished from the ordinary table, because in a worksheet the data are always arranged according to some narrow and exact mathematical requirement and include the mathematical computations of the specific data for a specific purpose, such as a mean, a regression line, or correlation.

Any meaningful tabulation must be based on some logical classification of the data. It is hardly conceivable that any intelligent person would jumble data in a table of meaningless confusion. Some basis of logical arrangement may be taken for granted in any table. Sometimes the principle of which classification is made is simple and is revealed to the reader by a glance at the table. At other times the principle of organization is complex and obscured by a mass of details. In any case, some systematic principle of analysis should be worked out for each table before tabulation is begun.

There are four basic principles for tabulation:

1. Time sequence;
2. Spatial sequence, or location;
3. Quality, or order of characteristics;
4. Quantity, or size.

1. Time sequence is a universal human experience. We are continually conscious of the lapse of time and the sequence of events in time. The operations of a business, a social event, or a journey may be classified according to the time order in which

they occur. All time series may be so organized. This is a fruitful basis for classification if there is any logical relation between the order of past events and those of the present or expected future. The business cycle, seasonal, or month-to-month variation in sales or production, changing inventory in a store, or the operations of agriculture are illustrations of this principle with which the student is familiar. It has a wide application.

2. Spatial sequence, or geographic location, is as universal as time. Data may be collected without reference to time, or duration may be eliminated by taking data from many locations at one instant or point of time. The population data of any single census are for one point of time. They measure the number of persons by areas at the same point of time. In fact, all data of all kinds in any single census are for a point in time. They cannot be classified on a time sequence basis, but only on the bases of space, quality, or quantity. A succession of censuses is necessary to form a time series. Geographic subdivisions are a common and very useful basis for classification. Sales, production, wages, interest rates, deaths, births, rainfall, and any other data that may be associated with space may be classified by location.

3. Quality and quality variations are universal facts in nature and in our thinking processes. Adverbs and adjectives and their comparisons make up a large part of all languages. We are continually comparing good, better, and best; slow, slower, and slowest; sweet, sweeter, and sweetest. We could not buy, sell, or conduct business, or carry on social intercourse, or establish or operate political organizations without quality distinctions. This is one of the widest and most useful bases for tabulation. Careful and accurate discrimination among qualities, physical, mental, and social, is essential to its proper use.

4. Distinctions between quantities are among the simplest and most necessary of our experiences. Much, more, and most; little, less, and least are quantitative comparisons necessary in sales, purchases, production, transportation, distribution, and consumption. Data which can be stated in numerical form can be tabulated according to quantity variation. Man has developed a large number of units for measuring quantity such as the

gram, ounce, pound, ton, inch, foot, yard, rod, mile, B.T.U., pint, quart, barrel, meter, dollar, franc, second, hour, and light-year. On the basis of such units, most data can be classified and tabulated. Bills of lading, inventory sheets, pay rolls, population figures, number of births, imports, school enrollment, etc , are quantity tables.

TABLES

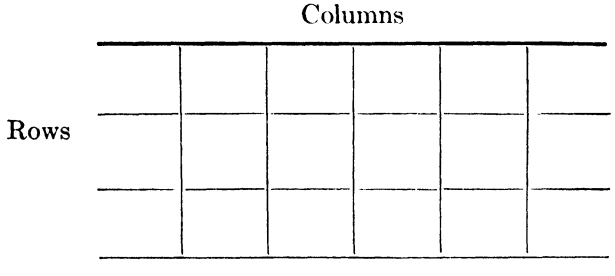
Various authors have classified and defined tables as general and specific, primary and secondary, summary and detailed. These definitions are somewhat useful but in the main are too vague and inaccurate to be of much value in guiding students in making tables. The best classifications rest on the *form* of the tables. The following types of tables may be clearly differentiated.

1. Complete cross-classification tables
2. Incomplete cross-classification tables
3. Percentage comparison tables
4. Frequency tables
5. Cumulative tables

Essential Parts of a Table

Before the various types of tables are analyzed in detail, it is necessary to have a clear idea of the essential parts of any table.

Body of Table. The most important part of any table is the body of the table. This part is composed of the columns and rows or crossed vertical and horizontal lines and spaces which include the data.



The organization of this part of a table is of prime importance. It *is* the table. Its form is determined by the (1) bases of classi-

fication, (2) the number of classes and subclasses desired, (3) the relationships which it is desired to show among classes. Before the body of a table can be formulated, it is necessary to study the data in detail and the purposes for which they are to be analyzed. If *all* characteristics of the data are to be revealed, there must be a class or subclass for each characteristic. If only *some* of the relationships are to be emphasized, the number of classes may be reduced. In any case, the statistician must decide what he wishes to reveal about the data before he can set up the form of the table. If all inter-relationships are to be shown, a complete cross-classification table is necessary. If only some of the relationships are to be emphasized, an incomplete cross-classification table is sufficient. The beginning student may have to experiment with several forms before he secures the most appropriate one. A clear and adequate table is the result of much careful thought. No table is better than the planning that goes into its formation.

Captions and Stubs. A *caption* is a heading of a column or subcolumn. A *stub* is the heading or description of a row, or

(Captions)

			A	B	C						
		A def			B def			C def			
		d	e	f	d	e	f	d	e	f
(Stubs) M	q r s									
	q									
	r									
	s									
N	q r s									
	q									
	r									
	s									
O	q r s									
	q									
	r									
	s									

cross-section of a table. If the table is to be complete, it will have one or more general class captions and several subclass captions. It will likewise have several stubs. The preceding table pattern is illustrative of these relationships.

Table Totals. There are two sets of locations for table totals. If it is desirable to emphasize the totals, they should be placed at the top and at the left. This is their location in all tables in the various United States censuses and in many business reports. Since the totals are the largest and in many respects the most important items in the table, they should be placed in the positions of emphasis. If, however, there is no purpose for emphasizing the totals, or one wishes to follow the traditional routine summation of data, the totals may be placed at the bottom of the table and at the right. Accounting forms usually place totals at the bottom as in the balance sheet and profit and loss statement. Statistical analyses usually place the totals at the top of the table and follow the total with its breakdown into subclasses.

Locations of Emphasis in Tables. Since the purpose of a statistical table is to make the meanings of data stand out in bold relief, it is necessary to exploit the *locations of strongest contrast* for comparisons and contrasts.

1. The strongest position for comparison or contrast is the *vertical consecutive* location, or the placing of one item immediately above another in the same columns, as

98765
01234

In this position the entire comparison is taken in at once without moving the eye.

2. The second strongest position for comparison or contrast is the *horizontal consecutive* location, or the placing of one item in adjoining column in the same row, as

98765	01234

In this position the entire comparison or contrast is taken in with one slight movement of the eye.

3. The third strongest position of comparison or contrast is the *vertical alternate* location, or the placing of the items one above the other in the same column with one intervening item, as

98765
57975
01234

In this position the eye must pass over one intervening item to make the comparison, which delay slightly weakens the contrast.

4. The fourth strongest position of comparison or contrast is the *horizontal alternate* location, or the placing of both items in the same row with one intervening item, as

98765	57975	01234

In this position the eye must move farther to grasp the significance of the items to be compared, which movement requires a longer lapse of time and more effort.

5. Other positions of still weaker emphasis may be employed by placing the items still farther apart in the columns or in the rows. From these statements the basic principle of location for emphasis is derived, that *that location in a table is strongest for comparison or contrast which forces the idea into the consciousness of the reader with the least expenditure of time and energy on his part.* It makes the idea stick out like a sore thumb.

IDENTIFICATION OF TABLES

Every table is a distinct creation for some specific purpose. As an individual structure it must be named, numbered, and its source of data given that it may be easily and quickly identified and referred to in discussion.

1. Name or Title. Every table must have a title which should answer the three questions of What? When? and Where? with the least expenditure of time and effort on the part of the reader. It is usually effective to place the title of a table in the form of a three-line inverted triangle, as

---

---

which answers in order the three questions of What? Where? and When? Examples of this type of title are

Height of Grade School Children
Stillwater, Oklahoma
1951

and

Production of Open Hearth Steel
United States
1951

or

Merchant Ships Sunk by Submarines
in Atlantic Ocean
May, 1942

2. Number. Although a good title reveals at once the content and meaning of a table, it is a poor means of reference and identification because of its length. In referring to a table in discussion, it is a waste of time to have to repeat its title at every reference. As a matter of economy for reference, all tables should be numbered. The table numbers for a chapter, and preferably for the entire book, should be in consecutive order to facilitate location in the body of the text. If they are numbered by chapters, the numerals should give the chapter number first followed by a period or colon followed by the table number, as 4.3, or 4:3, or 17.2, 17:2, etc. It is usually better to number all the tables in a book in consecutive order, as Table 1, Table 2, . . ., Table 175, etc., or Exhibit 1, Exhibit 2 . . . Exhibit 207, etc. These numbers may be placed above the title at the top of the table or at the

bottom of the table with the source reference. They should be written in a type that is easy to locate and identify.

3. Source Reference. The source of the data used in each table should be given in a reference at the bottom of the table, as

* *Survey of Current Business*, United States Department of Commerce, May, 1951.

or * *Monthly Labor Review*, United States Department of Labor, October, 1951, p. 763.

or * Dun and Bradstreet, Inc., *Dun's Review* (June, 1951) p. 7.

or (1) *The Babsonchart of Business Conditions*, Babson's Statistical Organization, June, 1950.

COMPLETE CROSS-CLASSIFICATION TABLES

In a complete cross-classification table every characteristic of the data included in the table can be read from a single point of data. The term "complete" does not mean that every fact and point in the original data is included in the table, but that *the cross-classification is complete.* For instance, if four classifications are included in the table, the value of every item in the table is revealed in all four classes and can be read from one point in the table.

I. One Classification

TABLE 1

HEIGHTS OF 108 GRADE SCHOOL CHILDREN OF TULSA, OKLAHOMA, APRIL, 1951 *

All Heights	Number of Children
	108
44.5–51.4	30
51.5–58.4	50
58.5–65.4	28

* Source: Worksheet No. 1, Chapter 5.

The one characteristic revealed in this table, that of height, is completely classified, so that the reader can tell at one glance the variations in the number of children in all included classes of heights.

II. Two Classifications

TABLE 2

HEIGHTS AND WEIGHTS OF 108 GRADE SCHOOL CHILDREN OF TULSA, OKLAHOMA, APRIL, 1951 *

All Weights	Heights			
	All Heights	44.5–51.4	51.5–58.4	58.5–65.4
	108	30	50	28
45.5– 70.4	49	26	23	
70.5– 95.4	41	4	25	12
95.5–120.4	18		2	16

* Source: Worksheet No. 1, Chapter 5.

Table 2 was made by tallying the height and weight of each child given in Worksheet No. 1 in its appropriate cell in the table, and then accumulating the totals under the appropriate captions and opposite the appropriate stubs. The figures for each cell are summed for each column and each row in totals at the top and at the left. The student should note how complete this cross-classified information is. This test may be made by placing a pencil point on any figure in the table and asking: "What are the heights and what are the weights of these children?" We may illustrate the point by taking the figure 26 near the center of the table. What are the heights of these 26 children? The answer comes at once. "They are from 44.5 to 51.4 inches tall." How much do they weigh? The answer is complete, "From 45.5 to 70.4 pounds." Try the 2 in the bottom row. "They are from 51.5 to 58.4 inches tall and weigh 95.5 to 120.4 pounds." This is a complete cross-classification table, revealing at once all information the table is designed to give about every item in it.

The classes may be made as numerous and narrow as desired. It is a powerful device for analysis.

III. Three Classifications

TABLE 3

HEIGHTS, WEIGHTS, AND AGES OF 108 GRADE SCHOOL CHILDREN, TULSA, OKLAHOMA, APRIL, 1951 *

All Weights	Heights			
	All Heights	44.5–51.4	51.5–58.4	58.5–65.4
	108	30	50	28
70.5– 90.4 months old	24	24		
90.5–110.4 months old	25	6	19	
110.5–130.4 months old	30		27	3
130.5–150.4 months old	29		4	25
45.5– 70.4 pounds	49	26	23	
70.5– 90.4 months old	20	20		
90.5–110.4 months old	18	6	12	
110.5–130.4 months old	10		10	
130.5–150.4 months old	1		1	
70.5– 95.4 pounds	41	4	25	12
70.5– 90.4 months old	4	4		
90.5–110.4 months old	7		7	
110.5–130.4 months old	18		16	2
130.5–150.4 months old	12		2	10
95.5–120.4 pounds	18		2	16
70.5– 90.4 months old				
90.5–110.4 months old				
110.5–130.4 months old	2		1	1
130.5–150.4 months old	16		1	15

* Source: Worksheet No. 1, Chapter 5.

THE CONSTRUCTION OF A THREE-CLASS TABLE

Making a three-way table is not so easy a task as it appears at first glance. Seventy-five percent of all students fail the first time they attempt it. Since there are only two sides or dimensions to a table, it is necessary to place two classes on one side when three classes are used. Two of the classes may be placed on the top and one on the side, or one on top and two on the side. The principle to follow in deciding which of these two methods to use is, *Place the largest number of subclasses on the side.* The reason for this is that there can be more rows than columns on a normal page. In Table 3 there are 20 rows but only 5 columns.

The second point to consider in making a three-way table is, *When two classes are placed on one side, one of these two classes must be put on the* OUTSIDE *and the other class put* UNDER *it and repeated over and over again as many times as there are subclasses and totals in the outside class.* This is where students stumble. They would do well to "stop, look, and listen" at this point. In Table 4, two classes, *weight* and *age* are placed on the side, and *height* is put at the top. On the side where the two classes are, weight is put on the *outside* and the four subclasses of age in complete form are repeated over and over *under* each subclass of weight. By no other process can a complete cross-classification table be made.

Let us give Table 3 the *pencil point check* to test whether it is a complete cross-classification table. Place the point of your pencil on 15, the lowest right-hand figure in the table. Can all three characteristics of these fifteen items be read from this one point? The answer is, Yes. If we read up the column we find their heights are between 58.5 to 65.4 inches. If we read across on the row, we find they are from 130.5 to 150.4 months old and weigh from 95.5 to 120.4 pounds. All the information about them in the table can be read from this one point. Try any other group in the table and you will get the same results. This principle is basic. It is the final test of the completeness of the cross-classification in any table and applies to all tables regardless of the number of classes and subclasses in them.

A four-way cross-classification table is a series of sections, or

subcells, each one of which is a table in itself. All the sub-totals of all these sections are accumulated in the grand totals sections at the top and at the left. A good practice for the student who is learning to make four-way tables is to give several of the items in this table the *pencil point check*. All four characteristics of each item in this table can be read from one point.

A five-way table could be constructed from these data by adding to the four classes included in Table 4 the fifth class of schools, as Elliot, Jefferson, and Lincoln. Such a five-way table would be too large to include on the single page of a text. If it were set up, it would be preferable to put three classes on the side and leave the two classes of height and sex as they are in Table 4. If a third class of schools were put on the side, it would be put on the outside of the two classes of weight and age and the 12 rows of the present table with their present stubs would be repeated over and over under (1) All Schools, (2) Elliot, (3) Jefferson, and (4) Lincoln, four times in all. The five-way table would have the same width as Table 4, but would be four times as long. This rapid expansion in size of complete cross-classifications with the addition of each new variable requires a large sheet of paper or the breaking up of the total table and the presentation of each separate section or cell on a separate page, if more than four main classifications are used. In the United States Decennial Census, some of the larger tables showing data by states, counties, age, sex, marital condition, etc., cover many pages.

The complete cross-classification table is a powerful tool for the analysis of data. A much more complete discussion of the analytical possibilities of such tables will be given in Chapter 12. Table 2 clearly reveals the relation between the heights and weights of grade school children. Table 3 shows this relationship modified by the age of children.

INCOMPLETE CROSS-CLASSIFICATION TABLES

Frequently an incomplete cross-classification table is sufficient for the analysis desired, and since it is easier to construct, it can be used to advantage in such cases.

TABLE 4

Heights, Weights, Ages and Sex of 108 Grade School Children
Tulsa, Oklahoma, April, 1951 *

All Weights	Total				Male				Female			
	All Heights	44.5–51.4	51.5–58.4	58.5–65.4	All Heights	44.5–51.4	51.5–58.4	58.5–65.4	All Heights	44.5–51.4	51.5–58.4	58.5–65.4
	108	30	50	28	52	15	24	13	56	15	26	15
70.5– 90.4 months old	24	24			10	10			14	14		
90.5–110.4 months old	25	6	19		11	5	6		14	1	13	
110.5–130.4 months old	30		27	3	17		15	2	13		12	1
130.5–150.4 months old	29		4	25	14		3	11	15		1	14
45.5– 70.4 pounds	49	26	23		21	12	9		28	14	14	
70.5– 90.4 months old	20	20			7	7			13	13		
90.5–110.4 months old	18	6	12		8	5	3		10	1	9	
110.5–130.4 months old	10		10		5		5		5		5	
130.5–150.4 months old	1		1		1		1					
70.5– 95.4 pounds	41	4	25	12	25	3	15	7	16	1	10	5
70.5– 90.4 months old	4	4			3	3			1	1		
90.5–110.4 months old	7		7		3		3		4		4	
110.5–130.4 months old	18		16	2	12		10	2	6		6	
130.5–150.4 months old	12		2	10	7		2	5	5			5
95.5–120.4 pounds	18		2	16	6			6	12		2	10
70.5– 90.4 months old												
90.5–110.4 months old												
110.5–130.4 months old	2		1	1					2		1	1
130.5–150.4 months old	16		1	15	6			6	10		1	9

* Source: Worksheet No. 1, Chapter 5.

I. Two Classifications, Incomplete

TABLE 5

HEIGHTS AND WEIGHTS OF 108 GRADE SCHOOL CHILDREN
TULSA, OKLAHOMA, APRIL, 1951 *

Heights				Weights			
Total	44.5–51.4	51.5–58.4	58.5–65.4	Total	45.5–70.4	70.5–95.4	95.5–120.4
108	30	50	28	108	49	41	18

* Source: Worksheet No. 1 Chapter 5.

If we apply the *pencil point check* to Table 5, we at once discover that its cross-classification is incomplete. Let us take the 18 in the last column under Weights. How much do these 18 children weigh? The answer is, "95.5 to 120.4 pounds." How tall are they? There is no answer. The table does not give the information. It is true that there are 28 children from 58.5 to 65.4 inches tall, but there is no way of knowing whether any of the 18 children are included in the 28. The pencil point check proves that the cross-classification of this table is *incomplete.* If there is no present need for such a cross-classification, Table 5 is sufficient. If all we wish to know is the range and frequency of the heights and the range and frequency of the weights without revealing any cross-relations, the above table is the kind to construct, because it is easier, quicker, and cheaper to make than a complete cross-classification table for the same data.

II. Three Classifications, Incomplete

TABLE 6

AGES, HEIGHTS AND WEIGHTS OF 108 GRADE SCHOOL CHILDREN,
TULSA, OKLAHOMA, APRIL, 1951 *

Age					Height				Weight			
Total	70.5–90.4	90.5–110.4	110.5–130.4	130.5–150.4	Total	44.5–51.4	51.5–58.4	58.5–65.4	Total	45.5–70.4	70.5–95.4	95.5–120.4
108	24	25	30	29	108	30	50	28	108	49	41	18

* Source: Worksheet No. 1, Chapter 5.

The data in Table 6 may also be presented in vertical form, as in Table 7.

III. Three Classifications, Incomplete

TABLE 7

AGES, HEIGHTS AND WEIGHTS OF 108 GRADE SCHOOL CHILDREN, TULSA, OKLAHOMA, APRIL, 1951 *

Ages	
Total	108
70.5– 90.4	24
90.5–110.4	25
110.5–130.4	30
130.5–150.5	29
Heights	
Total	108
44.5– 51.4	30
51.5– 58.4	50
58.5– 65.4	28
Weights	
Total	108
45.5– 70.4	49
70.5– 95.4	41
95.5–120.4	18

* Source: Worksheet No. 1, Chapter 5.

If cross-classification is not desired, there is no advantage in including more than one class in one table, except for conveniences of space and reference. The three classes in Table 7 might just as well have been placed in three separate tables.

In Table 8 there is cross-classification between sex and height, between sex and weight, and between sex and age, but there is no cross-classification between height, weight, and age. This is a part cross-classification table. This type of table is sufficient for many purposes and is widely used. Many of the tables in

United States censuses are of this type. They are also useful in business reports and in research work.

IV. Four Classifications, Partly Complete

TABLE 8

AGES, HEIGHTS, WEIGHTS, AND SEX OF 108 GRADE SCHOOL CHILDREN, TULSA, OKLAHOMA, APRIL, 1951 *

	Sex		
	Total	Male	Female
Ages, Total	108	52	56
70.5– 90.4	24	10	14
90.5–110.4	25	11	14
110.5–130.4	30	17	13
130.5–150.4	29	14	15
Heights, Total	108	52	56
44.5– 51.4	30	15	15
51.5– 58.4	50	24	26
58.5– 65.4	28	13	15
Weights, Total	108	52	56
45.5– 70.4	49	21	28
70.5– 95.4	41	25	16
95.5–120.4	18	6	12

* Source: Worksheet No. 1, Chapter 5.

PERCENTAGE COMPARISON TABLES

Percentages are the most generally used method of comparison in the business world, if not in all life's activities. The simple decimal system of tens and hundreds long ago became so fundamental a part of our daily methods of thinking that it may truly be said that all literate persons usually make comparisons on the basis of 100. This fact makes the *percentage comparison table* one of the most useful in statistical presentation. It is especially useful in comparing values which differ greatly in size, such as a large store and a small store, a whole industry and a single plant

in that industry, one department and an entire plant, exports and imports, and other such divergent values.

I. Pure Percentage Comparisons

TABLE 9

EMPLOYMENT IN MANUFACTURING, TRADE, SERVICE, AND GOVERNMENT IN THE UNITED STATES, 1945–1950, IN PERCENTAGES OF TOTALS *

Year	Total %	Manufacturing %	Trade %	Service %	Government %
1945	100.0	46.6	22.9	12.3	18.2
1946	100.0	43.4	25.8	13.9	16.9
1947	100.0	43.9	26.4	14.0	15.7
1948	100.0	43.4	27.0	13.6	16.0
1949	100.0	41.4	27.6	14.0	17.0
1950	100.0	42.4	27.2	13.6	16.8

* Source: Federal Reserve Bulletin, April, 1951, p. 438.

II. Mixed Percentage Comparisons

TABLE 10

GROWTH AND CHANGES OF WORK ENERGY OUTPUT IN THE UNITED STATES, 1850 TO 1944 *

Year	Billions of Horsepower Hours	Total Percent	Percent Supplied by		
			Animals	Humans	Tools
1850	12	100.0	51%	22%	27%
1860	17	100.0	52	21	27
1870	20	100.0	49	22	29
1880	28	100.0	46	21	33
1890	42	100.0	40	18	42
1900	57	100.0	33	15	52
1910	94	100.0	21	12	67
1920	145	100.0	11	8	81
1930	189	100.0	5	6	89
1940	247	100.0	3	4	93
1944	346	100.0	2	4	94

* Source: "America's Needs and Resources." 20th Century Fund. Revised and brought up-to-date by Dewhurst.

TABLE 11

U.S. ADVERTISING IN 1949 *

Medium	Amount in Millions	Percent of Total
Newspapers	$1,995.0	36.6
Radio	633.8	12.2
Magazines	492.5	9.5
Farm Papers	20.5	.4
Direct Mail	755.6	14.5
Business Papers	247.1	4.8
Outdoor	131.0	2.5
Television	63.0	1.2
Misc.	952.7	18.3
	$5,201.2	100.0

* Source: Advertising Federation of America Report to Joint Congressional Committee, March, 1951. *Broadcasting Telecasting*, April, 1951, p. 19.

TABLE 12

SALES AND NET INCOME OF 100 LARGE MANUFACTURING CORPORATIONS IN THE FIRST QUARTERS OF 1950 AND 1951 *

(In Millions of Dollars)

	1950	*1951*	*Change*	*% Change*
Sales, etc., revenues	$6,912	$9,436	+ $2,524	+ 37
Total costs, except taxes	6,026	7,865	+ 1,839	+ 31
Income before taxes	886	1,571	+ 685	+ 77
Fed. income & e.p. taxes	345	930	+ 585	+ 170
Net income after taxes	$ 541	$ 641	+ 100	+ 18
Taxes to income before taxes	39%	59%		
Net income per sales dollar	7.8c	6.8c		

* Source: National City Bank of New York.

TABLE 13

NET INCOME OF LEADING CORPORATIONS FOR THE FIRST QUARTER *

(In Thousands of Dollars)

No. of Cos.	Industry Groups	Reported Net Income After Taxes First Qr. 1950	Fourth Qr. 1950	First Qr. 1951	Percent Change From First Qr. 1950	Fourth Qr. 1950
29	Food products	$ 29,278	$ 41,782	$ 31,790	+ 9	−24
15	Beverages	16,236	30,286	29,992	+85	− 1
9	Tobacco products	16,769	20,953	17,320	+ 3	−17
29	Textiles and apparel	21,102	27,520	26,557	+26	− 3
24	Pulp and paper products	18,948	35,319	34,002	+79	− 4
40	Chemicals, paints, etc.	144,475	194,289	168,469	+17	−13
15	Drugs, soap, cosmetics	28,460	28,150	34,279	+20	+22
23	Petroleum products	251,305	429,151	377,353	+50	−12
24	Cement, glass, and stone	39,413	45,976	45,486	+15	− 1
38	Iron and steel	160,808	168,402	176,218	+10	+ 5
18	Building, heating, plumbing equipment	10,736	25,838	16,318	+52	−37
18	Electrical equipment, radio and television	60,882	106,738	65,871	+ 8	−38
40	Machinery	25,239	32,957	30,314	+20	− 8
11	Office equipment	12,765	19,387	13,991	+10	−28
40	Automobiles and parts	45,577	50,379	48,388	+ 6	− 4
12	Railway equipment	6,523	17,998	12,742	+95	−29
9	Aircraft and parts	10,371	18,787	11,898	+15	−37
58	Other metal products	40,981	94,554	75,777	+85	−20
27	Miscellaneous manufacturing	14,533	27,320	23,830	+64	−13
479	Total manufacturing	954,401	1,415,786	1,240,595	+30	−12
32	Mining and quarrying †	26,441	56,025	38,312	+45	−32
22	Trade (retail and wholesale)	19,415	49,238	26,934	+39	−45
17	Service and amusement industries	14,113	17,583	11,763	−17	−33
550	Total	$1,014,370	$1,538,632	$1,317,604	+30	−14

† Net income is reported before depletion charges in some cases.

* Source: National City Bank of New York, *Monthly Letter*, May, 1951, p. 51.

FREQUENCY TABLES

In Chapter 5, Worksheets Nos. 3, 4, and 5 are frequency tables. The student is asked to check the technique and methods explained in Chapter 5 for the construction of such tables. They have a wide use in statistics as the basis of further analysis. It is unnecessary to repeat at this point the long and detailed explanations of the principles that should be followed in the construction of such tables. The reason that they are called worksheets instead of tables in the body of the text is that their form

is dictated and severely limited by the nature of the mathematical computations which are based upon them. A frequency table should always be constructed in conformity with the four basic principles explained in Chapter 5 and the limitations which the mathematical manipulation of the data require. They are a highly specialized type of table, simple in form but based on narrow and exact principles of statistical analysis which must be followed strictly if dependable results are to be obtained. At this point Chapter 5 should be examined.

I. Cumulative Frequency Table

TABLE 14

FREQUENCIES AND INCREASING AND DECREASING CUMULATION OF FREQUENCIES OF THE EARNINGS OF 188 OIL REFINERY EMPLOYEES IN CHAPTER 6 *

Class Intervals of Oil Refinery Earnings in Dollars	Frequencies No. of Employees in Each Class	Increasing Cumulation of Frequencies	Decreasing Cumulation of Frequencies
$110 to $120	2	2	188
120 to 130	5	7	186
130 to 140	11	18	181
140 to 150	20	38	170
150 to 160	26	64	150
160 to 170	34	98	124
170 to 180	36	134	90
180 to 190	28	162	54
190 to 200	16	178	26
200 to 210	7	185	10
210 to 220	3	188	3

* Source: Worksheet No. 7 in Chapter 6.

CUMULATIVE TABLES

A cumulative table is one which carries a value or series of values through a set of computations and analysis to some specific summation which either increases or decreases the initial figures as the processes of the analysis may require. In its simplest form it is a successive cumulation or decrease of quantities. The

increasing cumulation measures the total number which has come into existence at the end of each succeeding class; the decreasing cumulation measures the number remaining in existence at the end of each class.

An analysis such as may be made by Table 14 is very useful in revealing the number of railway ties, telephone poles, or other fixed capital goods that will be standing or in use at the end of successive periods of time. It is also used in making mortality tables for insurance companies. It can be made to apply to any situation in which there is a changing number of items in successive classes or at succeeding periods of time.

II. Special Cumulative Tables

The best examples of special purpose cumulative tables are the accounting Profit and Loss Statement and the periodic Balance Sheet. In the typical profit and loss statement one begins with (1) total sales, from which is deducted (2) cost of sales, which gives (3) gross returns from sales. From this is subtracted (4) sales expenses which gives (5) net profits from sales.

The ordinary bank statement and the commercial statement are samples of cumulative tables (see pages 68 and 69).

TABLE 15

PORTION OF CONSOLIDATED INCOME ACCOUNT OF UNITED STATES STEEL CORPORATION FOR 1949 *

Sales and Revenue		$2,293,257,287
Cost of Sales		1,721,527,637
Gross Profit from Sales		$ 571,729,650
Selling, general and administration expenses	$ 90,152,806	
Pay under pension plan	7,129,924	
Provision for doubtful accounts	1,186,665	
Depreciation and depletion	119,661,652	
Taxes (other than income)	71,014,171	
		289,145,218
Operating Profit		$ 282,584,432

* Source: *Moody's Manual of Investments*, Industrials, 1950, p. 2668.

Among other tables of this type are the perpetual inventory card, bills of lading, individual accounts, installment purchases accounts, income tax forms, and many other tables, forms, statements, and accounts universally employed in the business world. An intimate acquaintance with them is an essential part of a business education and a mastery of business statistics.

The five types of tables analyzed in this chapter cover all the essential points in tabulation. The few types, however, may be varied and combined into an almost endless variety of tables which are suited to every kind and degree of statistical presentation.

TABLE 16

CONSOLIDATED BALANCE SHEET OF SEARS, ROEBUCK & COMPANY
DECEMBER 31, 1950 *

Assets	
Cash	$190,008,871
Marketable Securities	4,945,582
Notes and Accounts Receivable	75,263,320
Inventory	321,396,365
Prepaid Expenses	13,672,596
Total Current Assets	$605,286,734
Investment in Allstate Insurance Company	4,848,347
Foreign Subsidiaries	19,783,521
Other secured Investments	8,608,605
Property, Plant and Equipment	290,518,544
Less Reserve for depreciation	120,588,867
	$169,929,677
Total Assets	$808,456,884
Liabilities	
Accounts payable	43,592,543
Accrued Taxes	100,463,562
Accrued Salaries and Wages	32,872,595
Unfilled Orders	16,812,174
Total Current Liabilities	$193,740,874
Capital Stock	191,298,972
Earned Surplus	423,417,884
Total Liabilities	$808,456,884

* Source: *Moody's Manual of Investments*, Industrials, 1950, p. 2410.

The purpose of the present chapter is to describe the various types of tables and the technique of their construction. Chapter 12 will indicate how these various types of tables may be employed in statistical analysis. This kind of analysis is widely used in all types of research work and in business reports and should be mastered thoroughly by the student. Without such mastery he will find himself handicapped at many points in his later work.

TABLE 17

MR. JOHN DOE
STATEMENT OF ACCOUNT WITH THE
FIRST NATIONAL BANK, CENTERVILLE, NEW YORK
JANUARY, 1951 *

Checks	Deposits	Date	Balance
		December 31, 1950	$325.54
$ 25.50		January 2, 1951	300.04
13.43		January 2, 1951	286.61
4.80		January 2, 1951	281.81
.56		January 3, 1951	281.25
15.00		January 3, 1951	266.25
3.51		January 5, 1951	262.74
12.40		January 5, 1951	250.34
40.00	$250.00		
	201.00	January 9, 1951	661.34
7.50		January 12, 1951	653.84
587.50		January 12, 1951	66.34
9.31		January 15, 1951	57.03
10.00		January 19, 1951	47.03
5.82		January 19, 1951	41.21
15.00		January 21, 1951	26.21
10.00		January 22, 1951	16.21
4.95		January 27, 1951	11.26
10.00		January 30, 1951	1.26

* Source: Accounts, First National Bank, Centerville, New York, February 1, 1951.

Cumulative tables either forward or backward may be used to measure either time sequence or space relationships for any type of quantitative data.

SUMMARY

1. Tabulation is an orderly arrangement of data.
2. Tabulations are usually based on one or more of the following four bases: (1) Time sequence, (2) Spatial sequence, or location, (3) Quality, or (4) Quantity.
3. Captions are the designations or names of the columns at the top of a table.
4. Stubs are the designations or names of the lines at the left side of a table.
5. A complete cross-classification table is one in which the descriptive captions and stubs for any data may be read from one point in the table. All information on these particular data is available from the point of its location in the table.
6. The point of greatest emphasis in a table is the point at which the values to be compared are closest together. This is a vertical contiguous position, or one directly above the other.
7. All tables should have brief descriptive titles and numbers for identification or reference.
8. All tables should have a statement, usually at the bottom, indicating the source of the data tabulated.
9. Tables are the most extensively used method of organizing and presenting data either in the press or in statistical studies.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter VIII. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter III. Prentice-Hall, Inc., New York, 1948.

Ezekiel, M., *Methods of Correlation Analysis*, Chapters X, XI. John Wiley & Sons, Inc., New York, 1941.

Mudget, B. D., *Statistical Tables and Graphs*, Houghton Mifflin Company, Boston, 1930.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics*, Chapter IV. McGraw-Hill Book Company, New York, 1951.

REVIEW QUESTIONS

1. What is tabulation?
2. What are the logical bases of tabulation? What are the characteristics of each?

3. What considerations should determine the basis of tabulation in any particular table?

4. Why is it better to classify tables according to form instead of use and purpose? Explain.

5. What are the five principal forms into which tables may be classified?

6. What is required to make a complete cross-classification table?

7. How may one check a table to determine whether it is a complete cross-classification table? Explain.

8. What are the essential parts of a table?

9. What are captions, stubs, columns, and rows? Explain fully.

10. What is meant by positions of emphasis? What are the four strongest ones in the order of decreasing strength? Why? Explain fully.

11. What information should the title of a table give? Why?

12. When more than two classes are used in one complete cross-classification table, where must the third or fourth classes be placed? Explain.

13. What are the principal uses of percentage tables? Explain fully.

14. Name and explain the uses of four cumulative tables.

15. What are the principles on which frequency tables should be constructed? Explain.

16. What are the prime essentials in producing an essential table?

EXERCISES

1. The data in Worksheet No. 56, Chapter 2 is suitable for practice in making tables. Any 10 to 20 lines of data across the Worksheet is sufficient for a short exercise in table making. In fact, 3 variables at one time are sufficient.

2. The exercise at the end of Chapter 2 is suitable for table making in this Chapter 3. The data for any 10 to 20 States are sufficient for one laboratory exercise.

CHAPTER 4

RATIOS AND PERCENTAGES

In its simplest form a ratio is a quotient, or the numerical quantity obtained by dividing one number by another. The quotient obtained by dividing 75 by 25 is 3. The divisor, 25, is the base with which 75 is compared. In its complete form the statement of the ratio is 75 is to 25 as 3 is to 1, or 75 : 25 :: 3 : 1. By this device the larger numbers are reduced to the simple comparison of 3 to 1. The ratio of the area of Kansas to the area of Ohio is 81,774 ÷ 40,740, or 2.007 to 1, or 2 : 1. The ratio of the area of Texas to the area of Rhode Island is 262,398 ÷ 1,067, or 246 to 1.

Just as the total of a series of numbers is the simplest representative quantity for the single group, so the ratio is the simplest form of comparative figure between two or more numbers. We use such comparisons frequently in the daily affairs of life. "She is *one* girl in a million," "He can whip twice his weight," "I would rather have it ten to one," "Six of one and half a dozen of the other," "It isn't half so good," "Two to one," and many other common expressions illustrate the universal use of ratios.

Ratios are frequently stated as percentages, as 75 is 300% of 25, or 75 : 25 :: 300 : 100. The area of Ohio is 50% of the area of Kansas, or 40,740 : 81,774 :: 50 : 100. Such expressions as, Production is down 20%, Prices are up 10%, The cost of living is up 30%, Exports have fallen off 50%, reveal how important ratios expressed as percentages are in our modes of thinking and living. We are so accustomed to thinking in terms of percentages that the ratio 200 : 100 is as clear to most persons as the ratio 2 : 1. Other bases such as 10, 1,000, 10,000, or even 1,000,000 are used

for special types of ratios. The rule to be followed in selecting the base is, *The base chosen should be sufficiently large to permit the numerator to be a whole number or a whole number plus a short fraction but should be sufficiently small to prevent more than three digits appearing in the numerator to the left of the decimal point.* The death rate of Negroes in New York in 1938 is stated as 14.1 per 1,000. If the base were reduced from 1,000 to 10, the numerator would appear as .141 which is less easy to understand than 14.1. On the other hand, if the base is changed to 100,000 the numerator would be 1410 which violates the basic principle of the ratio which is *to reduce larger numbers to smaller ones for the sake of simplicity.*

BASES OF RATIOS

A great variety of bases may be used for computing ratios. Ten of the most important bases are (1) Total to total, (2) Part to total, (3) Part to part, (4) Part to average, (5) Present time to former, (6) Given area or distance to standard area or distance, (7) Standard unit (school, family, class, etc.), (8) Arbitrary units, 1, 10, 100, 1,000, etc., (9) Effect to cause, (10) Dependent variable to independent variable. Most of the ratios in general use rest upon one of these bases.

1. Total to Total. This type of ratio is used to compare one entire group with other entire groups. Per capita income is such a ratio. The total national income is divided by the total population and the quotient written as a certain number of dollars, as

$$\frac{\text{National Income}}{\text{Total Population}} = \frac{\$269,400,000,000}{154,492,000} = \$1,743.78$$

according to the Statistical Abstract, 1951. All per capita ratios fall in this class, such as per capita production of wheat, cheese, shoes, steel, etc. and per capita consumption of gasoline, oranges, tobacco, etc. For each ratio the total production or consumption figure is divided by the total population of the same geographic area. It is possible to have per capita ratios for city, county, state, nation, world or any other unit area.

Per Capita National Debt

For 1951 the per capita national debt of the population of the United States from data in the Federal Reserve Bulletin and in the U.S. 1951 Decennial Census population estimates is $1,652.20.

$$\frac{\text{National Debt}}{\text{Total Population}} = \frac{\$255,251,000,000}{154,492,000} = \$1,652.20$$

2. Part to Total. This base of ratios is used to compare an item or sub-total of a group to the total of the group. It is used in business to compare department or individual results with those of the entire plant or industry. The part is divided by the total times 100, which expresses the result as a percentage. It is illustrated by the following examples:

$$\frac{\text{Demand Deposits, Member Banks}}{\text{Demand Deposits, All Banks}} \times 100$$

$$= \frac{\$86,772,000,000}{\$100,740,000,000} \times 100 = 86.13\%$$

$$\frac{\text{Gold Certificates, N.Y. F.R. Bk.}}{\text{Gold Certificates, All F.R. Bks.}} \times 100$$

$$= \frac{\$6,814,944,000}{\$19,868,402,000} \times 100 = 34.30\%$$

(Source: Federal Reserve Bulletin, July, 1951.)

3. Part to Part. Sometimes it is desirable to state the relationships among the several parts of a whole in terms of one or more of the parts. Overhead costs may be measured as a ratio to direct costs. Radio advertising cost may be expressed as a percentage of magazine or newspaper advertising expense. A family's expenditures for clothing may be stated in a ratio to food outlay. Sales in the several sections of a department store may be computed as a percentage of shoe sales, or piece goods sales, or fur-

niture sales. The value of the by-products of a factory may be measured as ratios of the value of the principal product.

4. Part to Average. In a department store the average sales for all the departments may be the best base for ranking the several divisions. The average sales of all the salesmen of a store may be the best standard for measuring the efficiency of each member of the staff. The average teaching load of the members of a faculty is perhaps the best standard for measuring the specific load of any one member of the group.

5. Present Time to Former Time. In computing temporal, or time series, index numbers a former specific time period is always used as a base. If No. 2 hard winter wheat sold at .682 a bushel in 1940, the base period, and in 1948 sold at $2.05 a bushel, the ratio is $2.05 \div .682$, or 3.0 which is stated as a percentage 3.0×100 or 300%. If the wages of press-drill operators were .75 an hour in 1939, the base period, and were $1.44 an hour in 1949, the ratio is $1.44 \div .75 = 1.92$ which expressed as a percentage is 192%. If the production of bituminous coal was 460,772,000 tons in 1940, the base period, as is 630,624,000 tons in 1947 the ratio is

$$\frac{\text{1947 production}}{\text{1940 production}} = \frac{630{,}624{,}000}{460{,}772{,}000} = 1.369 \text{ or } 136.9\%$$

6. Standard Area or Distance. It is frequently advantageous to use a standard geographic area or political or economic area as the base of a ratio, percentage or index. Population per square mile, average production per acre, improved farm land per farm, or per county, or per state, are samples of such common ratios. Often a standard distance is used as the base, as cost of highways per mile, telephone poles per mile, or airplane accidents or deaths per 1,000,000 miles flown, gallons of gasoline consumed per hundred miles traveled, etc. In geographic, economic, business, population and social analysis, such indexes or ratios are necessary.

7. Standard Conventional Units. Society has developed or accepted many institutional or custom-controlled units as bases for measuring relationships. Among these units are the indi-

vidual, the family, the farm, the school, the class, precinct, ward, county, state, and nation. Income per family is such a ratio. It is obtained by dividing the total income for a given area by the number of families in that area. The ratio is expressed in dollars per family. The housing ratio of rooms per family is of the same type. Cattle per farms, library books per school, failing grades per class, policemen per ward, miles of road per county are all ratios based on conventional units as bases. One principal objection to such units is that they are not uniform among themselves for any group. The size of the family is a varying unit from two persons to several times that number. Farms vary from three acres to many thousands of acres. Counties vary greatly in population and area. In spite of this lack of uniformity in any type of conventional unit such ratios persist and are widely used.

8. Arbitrary Units. For any type of scientific study to give dependable results, the units of measurement must be definite and employed accurately. In many fields it is possible to set up exact units arbitrarily which result in more uniform measures than is possible with any conventional unit. This is especially true in the physical sciences. The volt, ampere, watt, kilowatt, B.T.U., foot pound, and horse power, are such units. The ton-mile, passenger-mile, and train-mile are arbitrary units used in railway transportation. Flight-hours, light-years, pupil-hours, class-hours, "standard tests," degrees of temperature, latitude and longitude, and even clock hours are a few other examples of the many arbitrary units in general use.

One of the most common units for ratios is one, 1. All per capita ratios are based on this unit. If the numerator of such a ratio is a fraction, it is preferable to increase the size of the denominator until the numerator becomes a whole number. If the ratio is $\frac{.5}{1}$ it is better to state it as $\frac{5}{10}$ or $\frac{50}{100}$. If the ratio is $\frac{.005}{1}$, it is preferable to state it as $\frac{5}{1000}$ or $\frac{50}{10{,}000}$. If the ratio were .000273, it should be written as $\frac{27.3}{100{,}000}$, or 27.3 per 100,000.

The denominator should be kept as small as possible but should be sufficiently large to make the numerator primarily a whole number. It is entirely meaningless to say that the death rate is .0113 per person. But to state the ratio as a death rate of 11.3 per 1,000 is easily understood. Arbitrary bases of 1, 10, 100, 1,000, 10,000 or even 10,000,000 are widely used and readily understood. The most general arbitrary base is 100, or a percent.

9. Cause and Effect. In any comparison in which there is a cause and effect relation between the quantities, the one judged to be the cause should be used as the base or denominator in the ratio. Factory output per man hours worked, new sales to dollar or lines of advertising, light or power produced per kilowatt consumed, plant growth per unit of fertilizer, taxes collected for thousand dollars of property, typing speed gained per hour of practice, and deaths from syphilis per 1,000 infections are illustrations of this type of ratio.

10. Independent to Dependent. In scientific statistical analysis the variables compared are listed as dependent and independent. The dependent variable is the one for which values are to be computed from related values of the other variable. In computing ratios the independent factor should be used as the base. A common ratio of this type is pounds of weight per height in inches for an individual. Dollar income per dollar of sales, the number of theater-goers per 1,000 of population, the ratio of imports to exports, and the age of wives as a ratio of the age of husbands are examples of this relationship. It does not require a degree of causation, although such a relation may exist. All that is necessary is that the factor considered to be the independent variable shall be the base of the ratio.

RATIOS BETWEEN LIKE ITEMS

In order for a comparison to yield a meaningful, logical and significant result, the two quantities compared must be of the same quality or kind and must be stated in identical or comparable units. In other words, the variables compared must differ in *quantity only* as measured in *identical units*. The simplest examples

of this requirement of ratios are those units stated in arbitrary common denominate numbers such as inches, pounds, feet, tons, dollars, gallons, barrels, etc. How many inches make a pound? is a meaningless question. How many gallons make a dollar? is equally foolish. But the number of *pounds* of cured pork obtained per *100 pounds* of hogs slaughtered is an important and significant ratio. The number of *gallons* of gasoline obtained per *100 gallons* of crude oil run to stills is a legitimate ratio and very valuable as a measure of oil-refining efficiency. The number of *tons* of coke obtained per *100 tons* of raw coal placed in the coking ovens is a useful and correct ratio. The number of *dollars* earned as income per *dollar* or *100 dollars* invested is an accounting ratio that is universally employed and understood.

In death rates and birth rates the comparison is between *persons* and *persons*. The number of persons or individuals dying in a given time period is compared with the number of persons or individuals living during that period. The vitality of cabbage seeds or beet seeds is the ratio of *seeds* germinating per *100 seeds* planted. The accuracy of shooting or bombing is measured by the ratio of the *shots* or *bombs* that hit the target per *100 shots* or *bombs* fired or dropped. In all the cases listed above and in many others which the reader will readily call to mind the ratio is a simple statement of the relation between two quantities both measured in identical units, pounds to pounds, tons to tons, dollars to dollars, person to person, seeds to seeds. Such simple ratios are easy to understand and to use.

RATIOS BETWEEN APPARENTLY UNLIKE UNITS

Business, education, science and the activities of life generally are so complex that we are frequently compelled to make comparisons between quantities which are measured in unlike units. How many *inches* make a *pound?* may be a foolish question, but the question, how many inches of sausage make a pound or how many inches of quarter-inch rope make a pound is sensible. All per capita ratios involve this double unit type of comparison. *Dollars* of income per *person*, *bushels* of wheat consumed per *per-*

son, *gallons* of gasoline used per *person* or per *car*, or per *mile* or *100 miles*, *pounds* of beef consumed per *person* per *year*, *pairs* of shoes purchased per *person* per *year*, *kilowatts* of electricity used per *person* per *year*, or per *light*, or per *refrigerator*, the *hours* of labor per *automobile* produced, or *house* painted, or *acre* of land plowed, or *children* taught in school, or the *tons* of coal burned per *train mile* are only a few of the most common ratios of this type.

The device by which such ratios are made meaningful and valid is a *common denominator*. Inches and pounds, hours and miles, dollars and tons and other conglomerates of mixed units cannot be compared as such. The common denominator used to join such diverse units in a significant ratio is the idea, NUMBER or QUANTITY. The full statement of such a diverse ratio is the *number* of bushels of wheat per the *number* of persons consuming wheat. The *number* of gallons of gasoline consumed per the *number* of 100 miles traveled is the full and correct form of the ratio. The *number* of tons of coal burned per the *number* of train miles is the real ratio. The actual comparison is not between tons and miles, but between *number* and *number*, or *quantity* and *quantity*. Such comparisons are as logical and valid as ratios between identical units, because the numerator and the denominator of the fraction are identical. *Both are numbers.* This simple device of a common denominator makes possible the use of hundreds of the most important ratios which otherwise would be meaningless.

$$(1)\quad \frac{\text{Number of gallons of gasoline consumed}}{\text{Number of miles traveled}}$$

is as valid a ratio, as

$$(2)\quad \frac{\text{Gallons of gasoline obtained}}{\text{Gallons of crude oil run to stills}}$$

for in the second case the ratio is between

$$\frac{\text{Number of gallons of gasoline obtained}}{\text{Number of gallons of crude oil run to stills}}$$

In both cases the ratio is between

$$\frac{\text{number}}{\text{number}}$$

In fact most of the ratios in common use of the diverse type depend on the common denominators of *number* or *quantity*. The

$$\frac{\text{Number of children born}}{\text{Number of women of child bearing age}}$$

is a correct and necessary measure of the birth rate.

The three points for all makers and users of mixed ratios to keep clearly in mind are:

(1) The units and limitations of the numerator,
(2) The units and limitations of the denominator must be clearly expressed in the ratio,
(3) The relationship between the numerator and denominator must be clearly stated.

The standardized birth rate is a ratio which will illustrate these points.

$$\frac{\text{Number of live births per year}}{\text{Number of women 15 to 45 years old at the middle of the year in 1000's}}$$

Point No. 1. The numerator is defined to exclude still births. The period measured is limited to a year. No distinction is made between races, or national origins. These and other distinctions could be made. If it is intended to make them they should be clearly stated.

Point No. 2. The denominator is defined to exclude all women under 15 years of age and all over 45. The further restriction that these women are of these ages at the middle of the year is made. The age limit might have been placed as of the end of the year or of the beginning of the year. Whatever social or national restriction is placed on the numerator must also be placed on the denominator.

Point No. 3. The ratio between the numerator and denominator is stated in terms of 1,000's of women of child bearing age. A ratio of 141.7 would mean that *on the average* 141.7 children were born during the year for every 1,000 women between the ages of 15 to 45 inclusive.

AVERAGING RATIOS

Ratios cannot be correctly averaged by the simple average used for absolute numbers. The average of the heights of ten men would be found by adding the ten heights and dividing by ten. The average of the ratios of labor turnover for factories could not be accurately determined by this method, because of the difference in the number of employees and separations for the various plants. A weighted average must be used in which the original data are used as weights.

TABLE 18

Factory	Total No. of Separations per Year	Annual Average No. Employed	Ratio of Labor Turnover
1	150	460	32.6
2	792	3,765	21.0
3	2,902	3,658	79.3
Totals	3,844	7,883	Mean 48.7

The individual ratios are

1. $\frac{150}{460} = 32.6$
2. $\frac{792}{3{,}765} = 21.0$
3. $\frac{2{,}902}{3{,}658} = 79.3$

The ratio of the totals is

$$\frac{3{,}844}{7{,}883} = 48.7$$

If the separate ratios were already available, the same total result could be obtained as follows:

$$\begin{aligned} 32.6 \times \ 460 &= \ 150 \\ 21.0 \times 3{,}765 &= \ 792 \\ 79.3 \times 3{,}658 &= 2{,}902 \\ 7{,}883 &= 3{,}884 \end{aligned}$$

$$\frac{3{,}884}{7{,}883} = 48.7\% \text{ Weighted Average}$$

A simple arithmetic average of the three ratios is 44.3%.

$$\begin{array}{r} 32.6 \\ 21.0 \\ 79.3 \\ \hline 3\overline{)132.9} \end{array} = 44.3 \quad \text{Incorrect Average}$$

If one does not wish to use the large numbers of the original denominators, he may reduce them to percentages of their total and use these percentages as weights.

Numbers	Percentages
460	6
3,765	48
3,658	46
7,883	100

Ratios	Percentage Weights
32.6	$\times$ 6 = 195.6
21.0	$\times$ 48 = 1008.0
79.5	$\times$ 46 = 3657.0
	100 = 4860.6

$$\frac{4860}{100} = 48.6$$

This figure, 48.6, is almost identical with the weighted average of 48.7 obtained above.

POPULAR RATIOS

There are many ratios in general use with which the student may be familiar which will illustrate the practical value of ratios in the analysis of quantitative data. Since a ratio is the simplest numerical comparison which can be made between two quantities, there is a wide use for them in all statistical studies. The student should become familiar with those in common use.

Accounting Ratios

1. Current Ratio =

$$\frac{\text{Dollar Value of Current Assets}}{\text{Dollar Value of Current Liabilities}}$$

The current ratio is the total current assets divided by the total current liabilities. For the ordinary commercial firm a ratio of 2 : 1 is usually considered a safe minimum. The following figures from the Sears, Roebuck & Co., January 31, 1949, will illustrate this ratio:

$$\frac{\text{Current Assets}}{\text{Current Liabilities}} = \frac{\$605,276,659}{\$229,569,233} = 2.6$$

2. The Net Worth Ratio =

$$\frac{\text{Dollar Value of Net Worth}}{\text{Dollar Value of Total Debts}}$$

A safe ratio would vary among industries. For railroads the ratio might be 1 : 1, but for manufacturing firms it should be ordinarily more than 2 : 1. The ratio may be illustrated by the figures of the Marshall Field & Co., Chicago, December 31, 1948.

$$\frac{\text{Net Worth}}{\text{Total Debts}} = \frac{\$41,333,590}{\$28,274,024} = 1.46$$

3. Fixed Assets Ratio =

$$\frac{\text{Dollar Value of Net Worth}}{\text{Dollar Value of Fixed Assets}}$$

This ratio measures the security of creditors. If the net worth is large in proportion to fixed assets, creditors are in a more secure position than when the ratio is low. The balance sheet of Montgomery Ward & Co., January 31, 1949, will illustrate this ratio.

$$\frac{\text{Net Worth}}{\text{Fixed Assets}} = \frac{\$442,029,387}{39,861,935} = 11.1$$

4. The Turnover Ratio =

$$\frac{\text{Dollar Value of Sales}}{\text{Dollar Value of Merchandise (or Inventory)}}$$

This ratio measures the speed or efficiency with which inventory is turned into sales. In a chain grocery store the entire stock may be turned over 25 to 30 times a year. In a furniture store the ratio may be only 2 : 1. The figures for the May Department Stores Co., for the year ending January 31, 1949, will illustrate this ratio.

$$\frac{\text{Sales}}{\text{Inventory}} = \frac{\$407,266,466}{41,026,330} = 9.9$$

5. Collections Ratio =

$$\frac{\text{Dollar Value of Credit Sales}}{\text{Dollar Value of Receivables}}$$

If this ratio is high, collections on credit sales are poor.

Dun & Bradstreet, Inc. publish fourteen ratios on 72 lines of business including averages for hundreds of firms. These ratios based on large numbers of firms serve as a kind of basic standard against which the ratio of any particular firm may be compared to determine whether it is above or below average.

There are many other accounting, sales, and management ratios of great aid to the business executive which cannot be included here. Those presented above indicate how large a part such simple statistical ratios play in the operation of the modern business world, and how useful they are even for the smallest firm.

Agricultural Ratios

1. The Corn-Hog Ratio =

$$\frac{\text{Dollar Value of 100 lbs. of Live Hogs}}{\text{Dollar Value of 1 bushel of Corn}}$$

Since it requires on the average about 11 bushels of corn to produce 100 lbs. of live hog weight, the corn-hog ratio should be about 11. During the past thirty years it has ranged from 7 to 17 but has fallen between 9 and 12 about two-thirds of the time. When the ratio is high it pays to raise hogs. When the ratio is low it pays to sell corn.

2. Farm Tenancy Ratio =

$$\frac{\text{Total No. of Tenant Operated Farms in Area}}{\text{Total No. of Farms in Area}} \times 100$$

This ratio is usually reduced to a percentage. The ratio is as low as 3% to 4% in counties of New England and as high as 90% in Mississippi. It measures the percentage of farms that are operated by tenants. The figures for 1940 are as follows for

Maine	*N.H.*	*Bolivar County Mississippi*
$\frac{\text{Tenants}}{\text{Total Farms}} = \frac{2.519}{38{,}980} = 6.5$	$\frac{1.054}{16{,}554} = 6.3$	$\frac{10.843}{12{,}005} = 90.3$

Economic Ratios

1. Man-Land Ratio =

$$\frac{\text{Total Population in Persons}}{\text{Total Land Area in Square Mile}}$$

or for the United States

$$\frac{\text{Population}}{\text{Area}} = \frac{154{,}492{,}000}{3{,}022{,}387} = 51.1 \text{ population per sq. mi.}$$

This is an important ratio in measuring the ratio of resources to population, or population density. The Man-Land Ratio may also be stated in reverse as the number of acres per person as:

Land-Man Ratio =

$$\frac{\text{Total acres of Tillable Land}}{\text{Total Population of Area}}, \text{ or for}$$

Iowa, $\frac{34{,}281{,}948}{2{,}609{,}748} = 13.1$ acres per capita

Indiana, $\frac{19{,}444{,}456}{3{,}917{,}904} = 5.0$ acres per capita

Rhode Island, $\frac{233{,}549}{786{,}324} = 0.3$ acres per capita

2. Efficiency Ratios =

(1) $\frac{\text{Output in Product}}{\text{Input in Power}}$

This is an engineering ratio with an economic implication.

(2) $$\frac{\text{Dollar Value of Products Sold}}{\text{Dollar Value of Production Costs}}$$

This is the economic ratio measuring the profits of business, or the efficiency of the business as measured by profits. The figures of the Chrysler Corporation for 1940 illustrate this ratio.

$$\frac{\text{Sales}}{\text{Costs}} = \frac{\$744,561,239}{\$633,606,187} = 1.2$$

3. Net Labor Turnover =

$$\frac{\text{Total number of replacements for period}}{\text{Average working force for period}} \times 100$$

This ratio may be illustrated as follows, for the A & B Rubber Company.

$$\text{Net Labor Turnover} = \frac{\text{Replacements for Month}}{\text{Average Working Force for Month}} \times 100$$

$$= \frac{782}{8265} \times 100 = 9.4\%$$

Educational Ratios

The best known educational ratio is, perhaps, the Intelligence Quotient, IQ. It is stated as follows:

1. $$\text{IQ} = \frac{\text{Mental age of pupil in years}}{\text{Chronological age of pupil in years}} \times 100$$

If the mental age is 12 years, while the chronological age is 10, the IQ = $\frac{12}{10} \times 100 = 120$. If the mental age was 9 years, while the chronological age was 11, the IQ would be $\frac{9}{11} \times 100 = 82$.

2. $$\text{A.Q.} = \frac{\text{Test Score}}{\text{Mental Age}}$$

The Achievement Quotient is the measure of accomplishment as compared with ability. It measures how well the student applies himself to the task.

Population Ratios

1. Crude Death Rate = $\dfrac{\text{Number of deaths occurring in the area during the year}}{\text{Number of people living in that area at mid-point of that year}} \times 100$

This ratio is called the Crude Death Rate because it makes no allowances for errors, the varying age and sex composition of different populations. While it may be quite accurate for some areas, in other areas the error might be large. The Refined Death Rate would as far as possible eliminate these errors. The ratio is multiplied by 1,000 to express the rate as a whole number instead of a fraction. Other death rates are the "corrected," "adjusted," and "standardized."

For Wisconsin in 1947, this ratio is

$$\text{Crude Death Rate for Wisconsin in 1947} = \frac{33{,}219}{3{,}219{,}000} \times 1000 = 10.3$$

2. Crude Birth Rate = $\dfrac{\text{Number of births occurring in area during the year}}{\text{Number of people living in area at the mid-point of that year}} \times 1000$

This ratio is called the *crude* birth rate because the number of births does not depend on the total population but on the number of women of child-bearing age. A given population might have a very low birth rate because most of that population consisted of old persons. For Ohio in 1947 this ratio is

$$\text{C.B.R.} = \frac{187.311}{7{,}754{,}000} \times 1000 = 24.2$$

3. General Fertility Rate = $\dfrac{\text{Number of births to women of all ages}}{\text{Number of women 20–44} \times 1000 \text{ (or 15–45)}}$

This ratio may be varied to cover only specific age groups or, specific race, residence, nativity, married, or other groups. It is a much more accurate measure of the fertility of a given popula-

tion than is the Crude Birth Rate. For Minnesota in 1947 this ratio is

$$\text{General Fertility Rate for Minnesota} = \frac{\text{Number of Births}}{\text{Number of Women 15 to 44}} \times 1000$$

$$= \frac{75.577}{645.329} \times 1000 = 117.1$$

4. $$\text{Sex Ratio} = \frac{\text{Number of males in population of area}}{\text{Number of females in population of same area}} \times 1000$$

For the United States in 1950 this ratio is

$$\frac{75{,}154{,}000}{76{,}464{,}000} \times 100 = 98.3$$

For the youngest and oldest portions of the population the ratios are:

$$\frac{\text{No. of males in U.S. under 5 years}}{\text{No. of females in U.S. under 5 years}} \times 100$$

$$= \frac{7{,}712{,}000}{7{,}394{,}000} \times 100 = 104.3\%$$

$$\frac{\text{No. of males in U.S. 75 or over}}{\text{No. of females in U.S. 75 or over}} \times 100$$

$$= \frac{1{,}578{,}000}{1{,}878{,}000} \times 100 = 84.0\%$$

(Source: U.S. Decennial Census, 1951.)

Railroad Operating Ratios

1. Revenue for freight train mile =

$$\frac{\text{Operating freight revenue in dollars per period}}{\text{Total number of freight train miles for period}}$$

A train mile is the movement of a train for one mile. Since trains vary in length and load, the ratio represents the relationship only as an average of all trains.

$$\frac{\$6{,}533{,}868{,}763}{698{,}564{,}116} = \$9.35$$

(Source: *Statistics of Railways in U.S.*, 1945, pp. 194, 196.)

2. Freight revenue per ton-mile =

$$\frac{\text{Operating freight revenues in dollars per period}}{\text{Total number of ton-miles hauled}}$$

A ton-mile is a ton of freight hauled one mile. It is a uniform standard unit. Since some freight pays a much higher rate than other kinds of freight, the ratio represents the relationship only as an average.

$$\frac{\$6,533,868,763}{726,044,111,000} = \$0.009$$

(Source: *Statistics of Railways in U.S.*, 1945, pp. 196–198.)

3. Traffic Density Ratio =

$$\frac{\text{Total net ton-miles per year}}{\text{Total miles of track operated}} =$$

$$\frac{726,044,111,000}{228,473} = 3,177,810$$

(Source: *Statistics of Railways in U.S.*, 1945, pp. 194–198.)

There are many other operating ratios and financial ratios used by railways given in the source listed above.

The Survey of Current Business carries monthly the following airplane operation data.

1. Express and freight ton-miles flown
2. Mail ton-miles flown
3. Passenger-miles flown
4. Total miles flown
5. Express and freight carried
6. Passengers carried

SUMMARY

1. A ratio is a quotient, or a fraction.
2. A percentage is a quotient times 100.
3. The *base* of a ratio is the denominator of the ratio fraction.
4. Ten of the most common bases of ratios are (1) total to total, (2) total to part, (3) part to part, (4) average to part, (5) former time to present time, (6) standard area or distance to given area or distance, (7) standard

unit (school, family, class, etc.), (8) arbitrary units as 1, 10, 100, 1000, etc., (9) cause to effect, (10) independent variable to dependent.

5. Ratios may be correctly combined or added only when they are weighted by the individual ratio denominators.

6. Both the numerator and the denominator of a ratio must be clearly defined as to:

(1) The units and limitations of the numerator,
(2) The units and limitations of the denominator, and
(3) The relationship between the numerator and denominator must be clearly stated. There must be a common unit of comparison between them, such as quantity or number.

SELECTED REFERENCES

1. Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapters XI, XII, Irwin, Inc., Chicago, 1941.

2. Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter VII, Prentice-Hall, Inc., New York, 1948.

3. Walker, Helen M., *Mathematical Essentials for Elementary Statistics*, Henry Holt & Co., New York, 1936.

REVIEW QUESTIONS

1. What is a ratio? What is the difference between a percentage and a ratio?

2. Explain what is meant by the *base* of a ratio.

3. Explain the relationship expressed in each of the following types of ratios:

(a) Total to total
(b) Part to total
(c) Part to part
(d) Part to average
(e) Former time to present time
(f) Standard area ratios
(g) Standard distance ratios
(h) Arbitrary units ratios
(i) Cause and effect ratios
(j) Dependent to independent ratios

4. Explain fully how ratios may be set up between unlike units.

5. Explain how ratios may be averaged.

6. Give examples of the following ratios:

(a) Accounting ratios
(b) Agricultural ratios
(c) Economic ratios
(d) Educational ratios
(e) Population ratios
(f) Railroad operating ratios

7. What is the crude birth-rate ratio?

8. What is the man-land ratio?

EXERCISES

1. Compute the man-land ratios for:

State	Area Square Miles	Population 1950	Answer Persons per Sq. Mi.
New England	63,206	9,305,767	147.2
South Atlantic	268,431	20,972,788	78.1
Mountain States	857,836	5,041,829	5.8
New York	47,929	14,743,210	307.6
California	155,652	10,472,348	67.3
Nevada	109,802	158,378	1.4

2. Compute the ratio of population (population as base) to acres of farm land.

State	Population 1950	Farm Land Acres	Answer Acres per Person
Ohio	7,901,791	20,888,004	2.64
Connecticut	1,994,818	1,289,134	0.64
Wyoming	288,707	522,936	1.81
Kansas	1,898,519	34,902,226	18.38

3. Compute per-capita sales for:

City	Population 1950	Retail Sales $	Answer $ per Person
Atlanta	327,090	485,210,000	1,483.
Chicago	3,606,436	4,249,603,000	1,178.
Denver	412,856	477,520,000	1,157.
Houston	594,321	684,281,000	1,151.
Los Angeles	1,957,692	2,176,561,000	1,111.
Wilmington, Del.	109,907	183,438,000	1,669.

4. Compute the death rate for:

State	Population 1947 (Estimated)	Deaths 1947	Answer per 1000
Alabama	2,833,010	26,347	9.3
Arkansas	1,910,700	15,095	7.9
Idaho	525,270	4,780	9.1
Illinois	8,364,820	93,686	11.2
Maine	881,330	9,959	11.3
New Jersey	4,632,780	48,181	10.4
Oregon	1,508,300	13,501	8.9
Virginia	3,009,580	29,193	9.7

5. Compute the crude birth rate (rate per 100 population) for:

State	Population 1947 (Estimated)	Births 1947	Answers per 1000
Alabama	2,833,010	88,116	31.1
California	9,796,370	245,889	25.1
Iowa	2,595,850	63,858	24.6
Missouri	3,898,700	90,060	23.1
New York	14,177,630	323,250	22.8
North Carolina	3,700,880	112,877	30.5
Texas	7,095,070	198,662	28.0
Utah	640,820	21,724	33.9

CHAPTER 5

FREQUENCY DISTRIBUTIONS

RAW DATA

Data is the term used to designate the raw measurements or numbers which are to be analyzed. The data are a sample drawn from a larger body of measurements, facts, or numbers called a *population* or *universe*. A population or universe is an *entire field of data*, such as the heights of all school children, the lengths of all cotton lint, the wages of all carpenters, the prices of all common stocks, the yields of all acres of wheat, or the lengths of all oak leaves. A population is usually *a large field of data*. A sample is usually a relatively *small portion* of that large area chosen for statistical study. Populations, however, vary greatly in size. Some are without approachable limits, such as the number of bacteria, oak leaves, insects, stars, amoeba, price quotations on potatoes, exchanges of goods and services of many kinds, snowflakes, and grains of sand. Even the number of human beings on earth now exceeds 2,250,000,000 and is increasing. In fact, most populations with which we are concerned are so large that we can never view or measure all their items. On the other hand, there are some populations which are quite small. Egyptian vases of certain dynasties are few. Coins of ancient Ephesus and boats of the Vikings and objects of many other interesting topics of study are scarce. In some cases, although a sample was expanded to the utmost workable limit, it could include only a small portion of the universe. In other cases the sample might include most or even all of the available data.

DISCRETE AND CONTINUOUS SERIES

In some types of data the unit is not divisible. It must be taken and counted in whole units or none. Such series of data are said to be *discrete*. One may have one-half pint of water, nine-tenths of a bushel of wheat, or three-fourths of a ton of coal, but half a boy, one-fourth of a soldier, or the tenth of a wife is an impossibility. Any series in which the units are indivisible entities is *discrete*. *Each unit of data is separate and complete.* A *continuous series* is one in which the units may be measured in fractions of any size no matter how small, so that there is a continuous flow of items of data with gradations infinitely minute. A ton of coal can be divided into hundredweights and these into pounds and the pounds into ounces and these into ten-millionths of an ounce and so on without theoretical limits. Such small units of data touch each other in a continuous flow of measurements.

A fuller and more exact treatment of samples and sampling will be developed in Chapter 13. It is sufficient at this point to recall that in statistics we are concerned primarily with the analysis of samples selected from larger groups of data. Ordinarily these samples when first selected consist of record sheets, or schedules, or questionnaires, which are not arranged in any systematic or regular order. They are simply collections of figures, such as the pile of sales slips in a store at the end of the day, or a pile of questionnaires as they were returned by mail over a period of days. Worksheet No. 1 is such a sample.

THE STATISTICAL ANALYSIS OF SIMPLE DATA

The beginning student is likely to think that statistical analyses are abstruse, difficult, highly mathematical procedures which are applicable only to mysterious data obtained in the recesses of a laboratory or experiment station. Quite the opposite is true. Statistics can dissolve experimental data of the most difficult types, but by far the widest and perhaps more important use of statistics is in finding correct answers to the large number of simple, commonplace, everyday problems of life. To illustrate this

simple, practical, universal use of statistics, this text employs four of the simplest series of data in the world — the wages of laborers and the age, heights and weights of school children. The heights of persons is a matter of universal knowledge. The wages of laborers are familiar to everyone. These humble under-foot data are as appropriate for statistical analysis as the most rare and remote series. Often the analysis of simple things is more interesting, useful, and valuable from a practical standpoint than those that are more remote. The elementary student may well begin his study in statistics with series of data within his own narrow experiences.

The data on the height, weight, and age of 108 school children were obtained by taking from the home room records of the elementary schools of Tulsa, Oklahoma,[1] the measurements for each tenth child on the school roll. Since the children were enrolled alphabetically in each of the twenty-five rooms of the elementary city schools used in this sample and there was no special selection or bias of any kind as to age, sex, economic status, or mental age, it is thought that the sample fairly represents the normal school grade distribution of these three measurements. Since Tulsa has a population drawn from every state in the nation, the sample would seem to be fairly representative of the general white elementary school population of the United States.

WORKSHEET NO. 1

In Worksheet No. 1, the data are mixed indiscriminately in the order in which they were taken from the school records.

The Array

The first step in the analysis of data is to arrange them in some systematic order according to some logical plan. The simplest and easiest method is to sort the data according to size, ranging from the smallest item to the largest. Such an arrangement is called an *array*. Worksheet No. 2 is such a group of arrays, first for age, then for height, and last for weight. In each case the

[1] Source: Student Records of Dr. Frank R. Pauly, Research Director of Tulsa Public Schools, 1951. Used with his cooperation and consent.

WORKSHEET NO. 1

AGE, HEIGHT, AND WEIGHT OF 108 TULSA, OKLAHOMA, SCHOOL CHILDREN OF GRADES 1 TO 6 INCLUSIVE, APRIL, 1951

Age Months	Height Inches	Weight Lbs.	Age Months	Height Inches	Weight Lbs.	Age Months	Height Inches	Weight Lbs.
140	61	108	101	54	60	89	48	55
78	49	57	111	57	63	112	53	68
100	52	60	79	47	56	120	56	81
148	63	114	88	50	61	139	58	101
79	46	47	150	63	119	147	63	111
77	45	46	85	48	48	148	60	100
88	51	72	80	48	51	118	56	85
99	53	66	112	54	66	90	49	59
105	55	85	111	55	95	101	53	66
123	58	105	130	59	112	114	54	67
150	60	81	137	60	84	150	64	115
143	61	90	128	59	90	148	62	110
120	58	85	78	48	53	115	56	73
104	54	65	105	55	75	124	56	77
108	56	81	109	52	61	100	50	62
82	49	53	79	50	62	110	56	68
82	50	55	149	63	115	95	50	63
76	46	47	137	60	84	115	57	82
107	52	66	130	59	90	147	62	93
140	57	77	128	55	65	145	64	117
118	56	67	112	52	62	117	54	81
141	61	91	104	50	63	103	55	83
102	52	64	89	50	64	96	52	70
81	47	53	91	49	55	121	55	64
83	51	68	87	47	56	144	59	84
77	48	75	116	53	66	138	61	87
143	59	111	113	54	68	146	59	107
122	57	74	135	57	80	117	55	76
132	59	76	129	55	74	93	52	71
109	53	69	133	55	68	98	49	56
107	52	62	140	60	105	138	59	92
77	48	75	129	56	80	125	57	90
80	49	63	89	51	55	124	56	75
101	55	91	87	51	71	99	49	50
134	60	106	128	56	80	126	56	82
99	53	76	131	60	102	147	64	113

WORKSHEET NO. 2

THREE ARRAYS OF DATA ON 108 GRADE SCHOOL CHILDREN OF TULSA, OKLAHOMA, APRIL, 1951

Array No. 1 Age in Months			Array No. 2 Height in Inches			Array No. 3 Weight in Lbs.		
76	101	128	45	52	57	46	66	82
77	102	128	46	52	57	47	66	83
77	103	128	46	53	57	47	66	84
77	104	129	47	53	57	48	66	84
78	104	129	47	53	57	50	67	84
78	105	130	47	53	57	51	67	85
79	105	130	48	53	58	53	68	85
79	107	131	48	53	58	53	68	85
79	107	132	48	54	58	53	68	87
80	108	133	48	54	59	55	68	90
80	109	134	48	54	59	55	68	90
81	109	135	48	54	59	55	69	90
82	110	137	49	54	59	55	70	90
82	111	137	49	54	59	56	71	91
83	111	138	49	55	59	56	71	91
85	112	138	49	55	59	56	72	92
87	112	139	49	55	59	57	73	93
87	112	140	49	55	60	59	74	95
88	113	140	49	55	60	60	74	100
88	114	140	50	55	60	60	75	101
89	115	141	50	55	60	61	75	102
89	115	143	50	55	60	61	75	105
89	116	143	50	55	60	62	75	105
90	117	144	50	55	61	62	76	106
91	117	145	50	56	61	62	76	107
93	118	146	50	56	61	62	76	108
95	118	147	51	56	61	63	77	110
96	120	147	51	56	61	63	77	111
98	120	147	51	56	62	63	80	111
99	121	148	51	56	62	63	80	112
99	122	148	52	56	63	64	80	113
99	123	148	52	56	63	64	81	114
100	124	149	52	56	63	64	81	115
100	124	150	52	56	64	65	81	115
101	125	150	52	56	64	65	81	117
101	126	150	52	56	64	66	82	119

smallest figure is placed first and the others following it next in size and so on with the largest number last. An array at once reveals three facts concerning the data:

1. The range of the items is the inclusive distance from the smallest to the largest. The range in age is 76–150 months, inclusive, in heights 45–64 inches, inclusive, and in weights, 46–119 pounds, inclusive.

2. The item that occurs most frequently. This number is called the mode. There are 12 children 56 inches tall which is a greater number than of any other height. Five children weigh 66 pounds and five weigh 68 pounds. In the array of weights there are two modes, or it is a bi-modal group. In the array of ages there are ten different groups of three items each. They are 77, 79, 88, 99, 101, 112, 140, 147, 148, 150. There is no distinct or true mode among the ages, as might well be expected among a large number of children who had started to school at ages of from 72 to 76 months during twelve successive semesters. The mode, therefore, is frequently a poor measure of the concentration of data.

3. The middle item in each array, easily located by counting halfway through it, gives a rough idea of the average of the sample. Since there is an even number of items in these arrays, 108 children, the middle or median is the average of the middle two items. For ages it is 112.5, or the average of 112 and 113. For heights the median is 55, and for weights 74. From these three arrays of data in a few moments we have arrived at a considerable amount of information concerning these children, which may be summarized as follows:

	Range	*Mode*	*Median*
Age in Months	75	none	112.5
Heights in Inches	20	56	55
Weight in Pounds	74	66 or 65	74

We may reasonably conclude that these grade school children average about 112.5 months in age and vary from about 38 months younger to about 38 months older than this average; that they average about 55 inches tall and vary about 10 inches on each side of this average; that their median weight is 74 pounds and they

range about 37 pounds on each side of that weight. For a very simple problem these results may be sufficient, but in many cases a more complete analysis will be desired.

Tally Sheets

When we have discovered the above information we have about exhausted the knowledge easily obtainable from the arrays. More complex and detailed methods are necessary to reveal other still obscure relationships. The next logical step is the creation of a tally sheet. This may be done first by listing the number of items falling in each unit of data. In fact such a tally sheet with one unit class is an excellent device for setting up an array.

WORKSHEET NO. 3

FREQUENCY DISTRIBUTION OF THE HEIGHTS OF 108 TULSA, OKLAHOMA GRADE SCHOOL CHILDREN

Height	No.	Height	No.	Height	No.
45	1	52	𝍸 111	59	𝍸 111
46	11	53	𝍸 1	60	𝍸 1
47	111	54	𝍸 1	61	𝍸
48	𝍸 1	55	𝍸 𝍸	62	11
49	𝍸 11	56	𝍸 𝍸 11	63	111
50	𝍸 11	57	𝍸 1	64	111
51	1111	58	111		
		Total 108			

If similar tally sheets were made for age in months or weight in pounds they would contain 75 and 74 classes respectively, some of which would be vacant and none of which would contain more than five items. Such frequency distributions contain too many classes to be either convenient or economical for further analysis.

Class Intervals

The purpose of arranging data in class intervals for statistical analysis is to save time and expense. Usually some inaccuracies result from the substitution of class intervals for the original individual items but if the classes are properly chosen the errors

are quite small and the time saved fully justifies the shorter method.

For data with a given range the width of the classes and the number of classes are interrelated. The wider the classes the fewer there may be. The larger the number of classes the narrower they must be. How many should there be? How wide should they be?

Rules for Selecting Class Intervals [1]

1. *Number of Classes.* One may compute the number of class intervals to use in setting up a frequency distribution as follows:

N = the number of items in the sample
n = the exponent to which the numeral 2 is raised

When $N = 2^n$, the number of classes should fall between $n - 1$ to $n + 3$, inclusive

FORMULA NO. 1

No. of C should fall within $n - 1$ to $n + 3$ when $N = 2^n$.

TABLE 19

APPROPRIATE NUMBER OF CLASS INTERVALS FOR SAMPLES OF VARIOUS STYLES

$n - 1$ to $n + 3$	2^n
2 to 6	$2^3 = 8$
3 to 7	$2^4 = 16$
4 to 8	$2^5 = 32$
5 to 9	$2^6 = 64$
6 to 10	$2^7 = 128$
7 to 11	$2^8 = 256$
8 to 12	$2^9 = 512$
9 to 13	$2^{10} = 1024$
10 to 14	$2^{11} = 2048$

[1] From B. C. Glover, Westerville, Ohio. 1944.

2. *Width of Classes.* Given a series of items, N, of known range, a suitable [1] class interval, i, may be chosen from Formula No. 2.

$$i = \frac{\text{Range}}{1 + 3.322 \text{ Log } N}$$

For the data on the Tulsa Grade School Children the width of classes would be as follows:

AGES

$$i = \frac{75}{1 + 3.322 \times \text{Log of } 108}$$
$$= \frac{75}{1 + 3.322 \times 2.03342}$$
$$= \frac{75}{1 + 6.755} = \frac{75}{7.755}$$
$$= 9.7 \text{ or } 10.$$

HEIGHTS

$$i = \frac{20}{1 + 3.322 \text{ Log } 108}$$
$$= \frac{20}{1 + 3.322 \times 2.03342}$$
$$= \frac{20}{1 + 6.755} = \frac{20}{7.755}$$
$$= 2.6 \text{ or } 3.$$

For the data on the Wages of 188 Oil Refinery Employees used in Chapters 6 and 7 the results are:

$$i = \frac{110}{1 + 3.322 \text{ Log } 188}$$
$$= \frac{100}{1 + 3.322 \times 2.27416}$$
$$= \frac{110}{1 + 7.56} = \frac{110}{8.56}$$
$$= 12.8$$

Ten classes were used for convenience since this number falls within the 7 to 11 range of Table 19.

Selecting Mid-points of Classes

The next step in an analysis of data after the class intervals have been set up and the frequency distribution completed in the tally sheet, is the determination of the *mid-point* of each class. There has been some confusion in textbooks on statistics as to the location and the means of locating the mid-point of a class interval. *The clearest, easiest, and most accurate method is to add the lower*

[1] From H. A. Sturges, *Journal of the American Statistical Association*, March 1926, p. 65.

limit of the first class to the lower limit of the second class and divide the difference by 2. This method clearly and accurately locates the mid-point of the first class. Since the classes are all the same width, the mid-points of the successive classes will be separated from the mid-point of the first class and from each other by the width of the class intervals. This relationship may be expressed and computed from the following formula.

FORMULA NO. 3

$$m_1 = \frac{L_1 + L_2}{2}$$

in which

m_1 = mid-point of first class
L_1 = lower limit of first class
L_2 = lower limit of second class

If we use the class intervals from Worksheet No. 5 as an example, the computations are,

$$m_1 = \frac{44.5 + 47.5}{2} = \frac{92.0}{2} = 46$$

Using the wider classes in Worksheet No. 4 as an example, we have,

$$m_1 = \frac{75.5 + 85.5}{2} = \frac{161.0}{2} = 80.5$$

Mid-points may also be determined from a graph as follows:

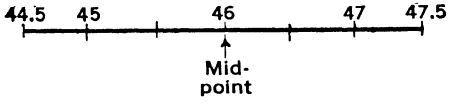


FIG. 35. Method of locating mid-point of a class interval by graphics

In Figure 35 the specific number, 44.5, does not occupy a space on the line but only a point. Likewise the number, 47.5, occupies only one point. All the line between these two points is filled by the intervening numbers and fractions such as, 44.6, 44.99, 45.0, 45.001, 45.999, 46.0, 47.499, etc. The point half way between the two extreme, or end, points is evidently 46.0, the mid-point of the class.

Class Limits

The data in Worksheet No. 1 are in whole numbers and can be accurate only within the half-unit deviation on either side of the specific datum, because age is a variate of the continuous type containing small fractional variations. Children whose exact heights in tenths of inches fall between 44.5 and 45.5 are listed as 45.0 inches tall.

WORKSHEET NO. 4

CLASS INTERVALS, MID-POINTS, TALLYS AND FREQUENCIES OF AGES OF 108 TULSA GRADE SCHOOL CHILDREN TAKEN FROM WORKSHEET NO. 1

1 Class Intervals	2 Mid-points	3 Tallies	4 Frequencies
75.5 to 85.5	80.5	𝍸 𝍸 𝍸 1	16
85.5 to 95.5	90.5	𝍸 𝍸 1	11
95.5 to 105.5	100.5	𝍸 𝍸 𝍸 1	16
105.5 to 115.5	110.5	𝍸 𝍸 𝍸	15
115.5 to 125.5	120.5	𝍸 𝍸 111	13
125.5 to 135.5	130.5	𝍸 𝍸 111	13
135.5 to 145.5	140.5	𝍸 𝍸 111	13
145.5 to 155.5	150.5	𝍸 𝍸 1	11
			Total 108

WORKSHEET NO. 5

FREQUENCY DISTRIBUTION OF HEIGHTS OF 108 TULSA, OKLAHOMA GRADE SCHOOL CHILDREN TAKEN FROM WORKSHEET NO. 1

Class Intervals	Mid-points	Tallies	Frequencies
44.5 to 47.5	46	𝍸 1	6
47.5 to 50.5	49	𝍸 𝍸 𝍸 𝍸	20
50.5 to 53.5	52	𝍸 𝍸 𝍸 111	18
53.5 to 56.5	55	𝍸 𝍸 𝍸 𝍸 𝍸 111	28
56.5 to 59.5	58	𝍸 𝍸 𝍸 11	17
59.5 to 62.5	61	𝍸 𝍸 111	13
62.5 to 65.5	64	𝍸 1	6
			Total 108

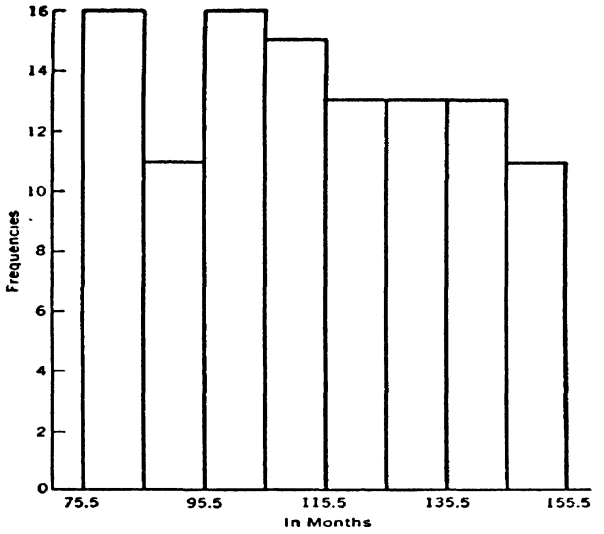


FIG. 36. Histogram of frequency distribution of ages of 108 Tulsa, Oklahoma grade school children, taken from Worksheet No. 4

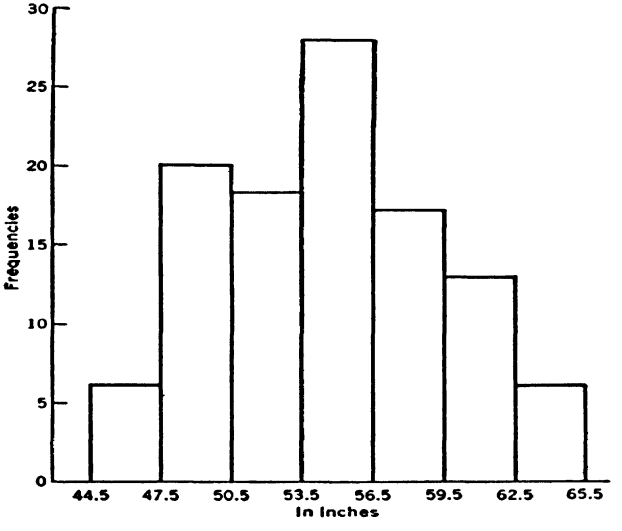


FIG. 37. Histogram of frequency distribution of heights of 108 Tulsa, Oklahoma grade school children, taken from Worksheet No. 5

Graphs of Frequency Distributions

When one has completed a tally sheet for a group of data and obtained a frequency distribution, he should make a picture of it in the form of a graph. This graph may be either a *histogram* or a *frequency polygon.* For comparative purposes the frequency distributions of Worksheets No. 4 and No. 5 are presented in both forms.

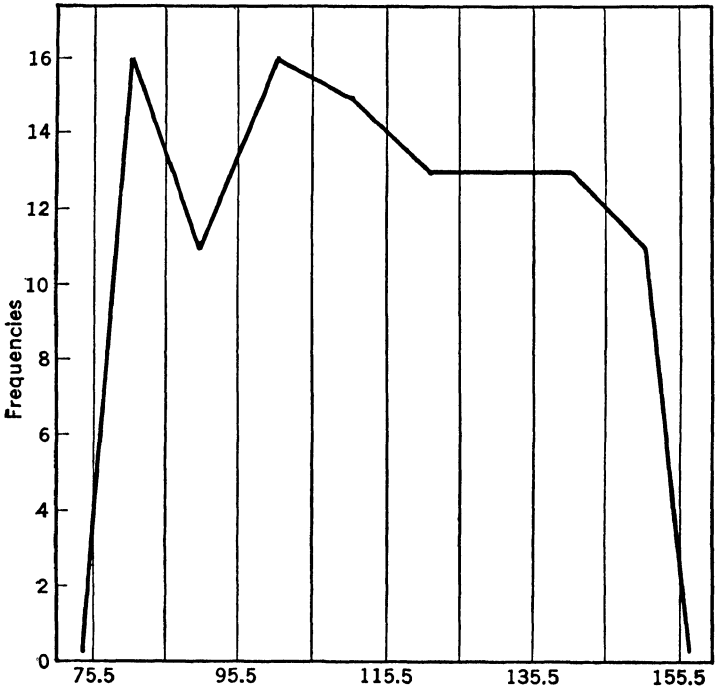


FIG. 38. Frequency polygon of frequency distribution of ages of 108 Tulsa, Oklahoma grade school children, taken from Worksheet No. 4

Types of Frequency Distributions

There are many types of frequency distributions, six of which are sufficiently common that the student should be introduced to them at the beginning of this study. Figure 40*A* is called the normal curve because it slopes equally on both sides from a central

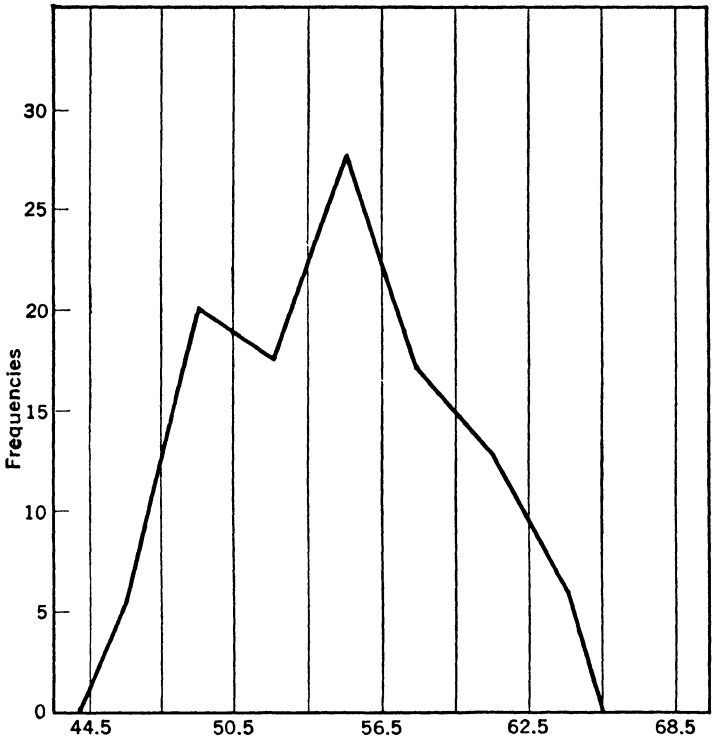


FIG. 39. Frequency polygon of frequency distribution of heights of 108 Tulsa, Oklahoma grade school children, taken from Worksheet No. 5

peak in ratios defined by the mathematical formula of the Normal Curve of Error. It is the basis of many of the generalizations of statistical theory. *B* is a slightly skewed, or lopsided, curve found frequently in social and business data. *C* is a highly skewed curve. *D* is a flat-topped curve illustrated by the frequency distribution of ages of grade school children in Worksheet No. 4. *E* is the low centered or "*U*" curve illustrated by the deaths of female pedestrians in automobile accidents, or the death rate per 1000 at various ages from 1 to 80 years of age. *F* is the "*I*" or "*J*" curve, depending on whether it slopes to the right or left and illustrated by the number of income tax returns for increasing sizes of corporations.

CURVES REPRESENTING PRINCIPAL VARIETIES OF FREQUENCY DISTRIBUTIONS

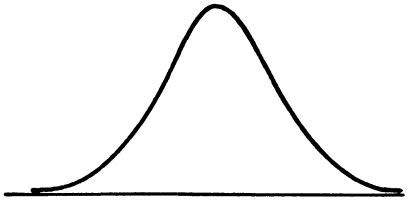


FIG. 40A. The Normal Curve

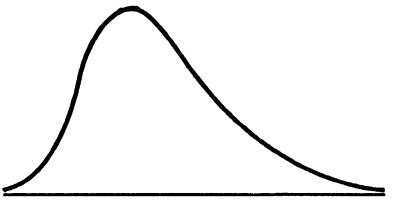


FIG. 40B. Slightly Skewed Curve

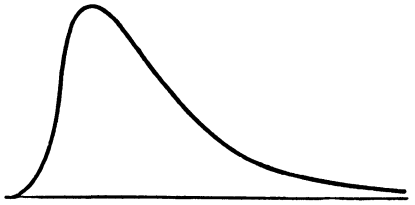


FIG. 40C. Highly Skewed Curve

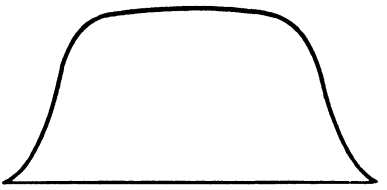


FIG. 40D. Flat-topped Curve

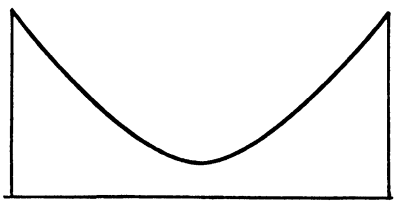


FIG. 40E. U-shaped Curve

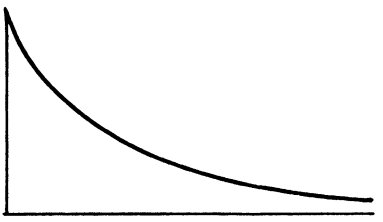


FIG. 40F. *I*-curve, if it slopes downward to left. *J*-curve, if it slopes downward to right

There are many degrees of variation for each of these common types of frequency distribution with which the student will become better acquainted as his knowledge of statistical sources expands.

SUMMARY

1. Data are raw measurements or numbers which are to be analyzed by statistical methods.
2. A statistical population or universe is a complete field of data.

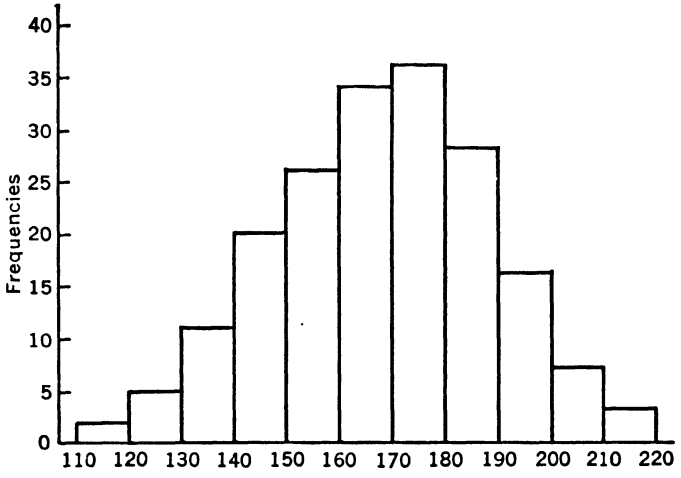


FIG. 41. Histogram of the frequency distribution of 188 oil refinery employees' eight semi-monthly wages, March 1–16, 1950. See Worksheet No. 7

3. Data which are measured in indivisible units, such as wife, child, or soldier, are said to be *discrete.*

4. Data which may be divided into infinitesimally small fractions, such as tons of coal, inches of wire, or acres of land, are said to be *continuous.*

5. An array is an arrangement of data in a schedule according to size, usually from the smallest item to the largest.

6. A class interval is a span of units of data, such as 4 to 7 inclusive, or 4–7.9, and is usually of equal or uniform width for the entire range.

7. A tally sheet is a device for arranging the raw data in a systematic order according to successive class intervals. The number of items in each class interval is the *frequency* of that class.

8. The class limits should be arranged as nearly as possible so as to place the mid-points of the classes as near as possible to the most frequent value of data for each class. For instance, in 10-cent-store sales the mid-points of the classes should preferably be 5, 10, 15, etc.

9. The mid-point of a class is easily located by subtracting the lower limit of the class from the lower limit of the succeeding class, dividing this difference by 2 and adding the quotient to the lower limit of the first class.

10. A frequency polygon is a graph of a frequency distribution in which the class intervals are measured on the x-axis and the class frequencies are measured on the y-axis.

11. The purpose of throwing data into frequency distributions is to economize time and effort in their analysis.

12. The use of class intervals introduces a slight error into the comparisons. Since this error tends to increase as the number of classes is reduced it is ordinarily advisable to have not less than eight or ten classes.

13. Since the amount of work increases as the number of classes increases it is ordinarily not desirable to have more than twenty or twenty-five classes. From twelve to fifteen is ordinarily a desirable number if the data lend themselves conveniently to such a classification.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics,* Chapters V to X, inclusive. Irwin, Inc., Chicago, 1941.

Chaddock, R. E., *Principles and Methods of Statistics,* Chapters IV, V. Houghton Mifflin Company, Boston, 1925.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics,* Chapters VII, XI. Prentice-Hall, Inc., New York, 1948.

Garret, H. C., *Statistics in Psychology and Education,* Chapter I. Longmans, Green and Company, New York, 1932.

Holzinger, K. J., *Statistical Methods for Students in Education,* Chapter II. Ginn and Company, Boston, 1928.

Mills, F. C., *Statistical Methods, Revised,* Chapter III. Henry Holt and Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agriculture,* Chapter I. John Wiley and Sons, New York, 1942.

Richardson, C. H., *An Introduction to Statistical Analysis,* Chapter II. Harcourt, Brace and Company, New York, 1935.

REVIEW QUESTIONS

1. What is a population or universe? Explain and name five examples.
2. What is a sample?
3. How is an array constructed? What is its use? What information will it yield?
4. Explain the use of a tally sheet.
5. What are the advantages and disadvantages of class intervals?
6. How wide should a class interval be? Explain fully.
7. How many class intervals should there be for a frequency distribution? Explain the advantages of a larger or smaller number.
8. Why should class intervals be of equal width?
9. When should the limits of class intervals be stated in fractions? Explain.

10. Locate the mid-points of the following class intervals: 10–13.9, 25–34.9, 12.5–17.4 and explain in detail how it is done.

11. How is a histogram constructed? Give example.

12. How is a frequency polygon drawn? Give example.

13. How does the "*J*" curve differ from the "normal" frequency distribution? Name three fields of data in which it may be found.

EXERCISES

1. Population of the counties of North Dakota, 1950, (in 100).

49, 168, 106, 18, 121, 40, 66, 253, 579, 117, 91, 60, 72, 54, 97, 53, 35, 391, 71, 54, 71, 62, 95, 63, 126, 76, 68, 188, 87, 192, 94, 81, 31, 139, 83, 143, 88, 54, 198, 111, 76, 52, 37, 23, 161, 51, 240, 63, 113, 205, 336, 104, 164.

(Select classes 50 wide beginning with 0–49.)

2. Iowa Total Score Achievement Test Grades of 120 High School Students Applying for Entrance to the University of Tulsa, 1950.

252, 116, 181, 120, 168, 111, 217, 160, 154, 119, 187, 110, 142, 116, 164, 176, 120, 84, 116, 111, 198, 156, 137, 215, 135, 200, 156, 228, 118, 84, 144, 131, 125, 158, 150, 130, 72, 135, 142, 130, 154, 163, 149, 114, 92, 90, 151, 129, 180, 134, 77, 112, 129, 90, 154, 106, 115, 204, 127, 118, 153, 103, 100, 158, 95, 157, 92, 154, 107, 130, 170, 148, 145, 128, 90, 147, 102, 154, 167, 162, 234, 152, 133, 206, 141, 96, 102, 118, 142, 124, 110, 137, 148, 165, 96, 99, 149, 142, 228, 57, 106, 157, 132, 139, 233, 92, 114, 153, 135, 95, 134, 139, 147, 138, 157, 70, 182, 102, 199, 162.

(Select classes 25 units wide beginning at 55.)

3. Average Weekly Wages of 70 Electrical Equipment Workers at a Chicago Plant, 1951.

56.75, 58.00, 58.62, 57.80, 56.15, 57.55, 58.82, 58.32, 57.64, 57.39, 58.71, 58.95, 57.92, 58.48, 58.10, 58.07, 58.31, 57.94, 57.83, 58.18, 58.29, 57.78, 58.12, 58.65, 56.30, 56.82, 57.29, 57.74, 58.99, 56.37, 56.78, 57.41, 57.28, 56.93, 56.88, 58.45, 56.91, 57.16, 57.22, 57.39, 57.55, 57.78, 57.17, 56.84, 57.90, 57.44, 56.31, 57.79, 57.21, 55.92, 57.31, 56.10, 58.80, 57.66, 57.13, 57.34, 56.66, 56.92, 57.89, 57.71, 58.87, 57.85, 57.90, 57.57, 58.23, 58.18, 57.21, 57.78, 58.09, 57.12.

(Select classes .50 wide beginning at 56.00–56.49.)

4. Lengths in millimeters of 100 Crêpe Myrtle leaves, Tulsa, Oklahoma, August, 1951.

59, 46, 53, 54, 45, 47, 67, 58, 63, 54, 60, 73, 52, 51, 64, 48, 54, 60, 52, 58, 57, 49, 52, 36, 39, 59, 52, 56, 63, 50, 45, 54, 53, 64, 44, 47, 50, 59, 55, 56, 51, 53, 55, 58, 42, 44, 45, 52, 57, 61, 68, 66, 65, 53, 55, 48, 49, 52, 51, 54,

60, 64, 59, 58, 48, 47, 57, 63, 46, 52, 55, 50, 67, 70, 42, 48, 67, 68, 70, 55, 63, 52, 51, 45, 49, 50, 43, 64, 52, 65, 63, 57, 38, 59, 64, 57, 63, 55, 44, 71.

(Select classes 5 wide beginning at 35 to 40.)

5. Dun & Bradstreet's "Estimates of Total Retail Dollar Sales of 52 American Cities with Populations between 100,000 and 200,000, 1949." (Sales in 1,000,000.)

223, 93, 187, 221, 134, 222, 227, 134, 104, 251, 250, 209, 221, 199, 198, 196, 202, 133, 122, 183, 155, 101, 115, 104, 190, 191, 163, 145, 105, 197, 126, 183, 175, 236, 183, 162, 163, 130, 136, 157, 156, 156, 108, 173, 123, 157, 117, 137, 167, 157, 119, 117.

(Select classes 25 wide beginning at 90 to 115.)

CHAPTER 6

AVERAGES

An average is a representative or typical number which may be used to indicate the value of a large group of numbers. It is a device to aid the mind in grasping the central or true significance of a large aggregate of facts or measurements while freeing oneself from the confusing burden of details. If the mind could grasp all the details of a large volume of data, understand all their interrelations, and retain all the information at once, statistics would be unnecessary. If we could keep in mind all the details of the ages, heights, and weights of a hundred children, the yields of ten thousand acres of corn, the wages of a million men, or prices of a thousand stocks, and the relationships which exist among them, we could dispense with averages and other statistics. The difficulty of keeping all details of the data in mind at once obscures our understanding of the central and logical relationships among the data. The selection of an average, or a central, typical representative number for the whole group removes these difficulties.

The reason an average is valid as a representative figure for a group of related measurements is that most samples of such data tend to pile up, or be concentrated, near the center of the group. In such a frequency distribution, items that differ to a *small* degree from the middle items are much more numerous than those which differ to a *large* degree. Among a thousand men there might be one or two only four feet tall and one or two six and a half feet tall. There would be, however, perhaps eight hundred men between five and six feet tall, and probably five hundred between sixty-five and seventy inches tall. There may be only one adult

in a million two feet tall, and one in ten million eight feet tall. *As the number of deviations from the mean increases, the number of items decreases; as the number of deviations from the mean decreases, the number of items increases.* It is this general tendency for related data to cluster or pile up about the normal that makes it possible to use one central, typical number to represent the entire group. Worksheet No. 7 and Figs. 39 and 41 illustrate this tendency. It is basic in statistical analysis.

It would be very difficult for us to reason abstractly, or to be at home in the modern world without averages. We think of the average man, the average student, average costs, income, expenses, height, weight, age, intelligence quotient, life expectancy, prices, wages, houses, yields, etc. All modern sciences from astronomy and physics to psychology and economics are based on averages. A valid comparison of two or more groups can be made only by means of the comparison of their averages.

There are five kinds of averages in common use. Why have more than one kind? For the same reason that the automobile mechanic has a variety of wrenches for different cars or various parts of the same car. For the same reason that a dentist has various forceps, drills, and other equipment for teeth that are radically different in shape and size. Some types of averages are better for one kind of data; others are better for another, or the same data may be analyzed by various methods for different purposes. The five averages are the arithmetic mean, the geometric mean, the harmonic mean, the median, and the mode.

ARITHMETIC MEAN

The most widely known and used average is the arithmetic mean. In popular language it is always referred to as "the average." It is a computed average and takes into consideration the two main characteristics of data, (a) the number of items, and (b) the size of the items. It is, therefore, a complete measure of data. Formulas number 4, 5, 6, and 7 provide for the calculation or the arithmetic mean under four slightly different conditions.

Simple Summation

Meaning of symbols:

X or Y = individual items of data
$\bar{X}$ or $\bar{Y}$ = arithmetic mean
N = number of items of data
Σ (Sigma) = The sum of, or total

FORMULA NO. 4

$$\bar{X} = \frac{\Sigma X}{N}$$

WORKSHEET NO. 6

ARITHMETIC MEAN OF THE CIRCULATION OF TEN LEADING AMERICAN MAGAZINES JANUARY TO SEPTEMBER, INCLUSIVE, 1950* IN 100.000.

	Data in (100.000)
Life	53.5
Ladies' Home Journal	45.4
Woman's Home Companion	40.6
Saturday Evening Post	40.4
McCall's	38.2
Woman's Day	37.6
Better Homes and Gardens	34.7
Look	31.8
Collier's	31.4
American Legion	30.3
	383.9

$$\bar{X} = \frac{\Sigma X}{N} = \frac{383.9}{10} = 38.39$$ hundreds of thousands of copies

or 3,839,000

* A.B.C., New York.

In this method the mean is found by simply adding up the values of the several individual items of data and then dividing that total by the number of items. It is especially suited to small samples. In samples of more than 100 items, and especially for

WORKSHEET NO. 7

COMPUTATION OF ARITHMETIC MEAN OF SEMI-MONTHLY EARNINGS OF 188 OIL REFINERY EMPLOYEES, MARCH 1–16, 1951

(Name and location of plant withheld)

130.05	157.84	160.80	195.04	181.44
173.28	195.58	178.88	165.76	168.63
184.32	163.28	176.80	191.52	166.32
172.16	167.37	168.87	194.60	168.96
166.00	183.52	160.94	133.76	187.12
140.80	168.16	154.48	174.88	157.28
180.80	175.12	137.89	176.72	205.21
163.87	167.20	175.51	164.56	157.20
169.36	169.84	180.46	190.88	121.76
173.28	184.92	175.52	207.58	187.12
171.12	178.00	127.52	186.88	174.16
145.20	173.39	170.88	196.00	176.88
130.40	156.64	180.48	139.04	160.16
195.00	144.32	179.28	196.00	176.32
204.04	186.88	189.76	160.16	140.80
174.58	136.36	140.80	167.04	146.88
159.60	189.28	134.72	205.12	190.50
179.20	162.23	158.92	130.40	145.28
197.24	143.46	142.86	147.84	165.60
192.16	154.32	152.32	202.40	157.44
171.36	171.20	201.18	161.36	151.09
188.64	198.24	146.96	165.50	184.63
157.84	211.44	188.64	142.08	129.20
144.32	176.30	172.56	198.92	181.44
159.52	161.92	187.68	177.28	142.17
163.36	139.20	166.32	145.53	114.09
155.76	166.53	176.35	171.04	127.12
145.20	144.32	180.72	167.04	208.41
151.20	170.24	173.60	155.84	125.74
142.56	162.48	178.96	182.05	136.19
151.74	188.94	189.92	159.68	158.12
191.52	155.20	173.36	197.60	182.97
159.92	152.90	156.64	168.96	166.97
151.84	196.44	157.60	172.48	153.94
170.64	133.76	118.56	172.48	186.68
180.32	165.76	165.14	146.56	173.28
180.92	146.97	160.31	170.88	
157.98	217.08	189.92	211.28	31557.01

$$\bar{X} = \frac{\Sigma X}{N}$$

$$= \frac{\$31{,}557.01}{188}$$

$$= \$167.856$$

samples that run into many hundreds or thousands of items, it is preferable to use class intervals. The student will recall that the principles and methods for setting up class intervals and constructing tally sheets are explained in Chapter 5.

The arithmetic mean is suited to the averaging of absolute numbers. The mean is always stated in the same units as the original data, such as bushels, inches, tons, dollars, etc.

Long Method with Class Intervals

Meaning of symbols:

i = size or width of class intervals
m = mid-point of class intervals, or class mark
f = frequency, or number of items in each class interval
N = total number of items in data, or total of all f's

FORMULA NO. 5

$$\bar{X} = \frac{\Sigma fm}{N}$$

WORKSHEET NO. 8

TALLY SHEET AND FREQUENCY DISTRIBUTION OF SEMI-MONTHLY EARNINGS OF 188 OIL REFINERY EMPLOYEES, MARCH 1–16, 1951
(Name and location of plant withheld)

Class Intervals $	Tally Marks There is one tally mark for each item of original data	Frequency Distribution
110 to 120	11	2
120 to 130	𝍸	5
130 to 140	𝍸 𝍸 1	11
140 to 150	𝍸 𝍸 𝍸 𝍸	20
150 to 160	𝍸 𝍸 𝍸 𝍸 𝍸 1	26
160 to 170	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 1111	34
170 to 180	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 1	36
180 to 190	𝍸 𝍸 𝍸 𝍸 𝍸 111	28
190 to 200	𝍸 𝍸 𝍸 1	16
200 to 210	𝍸 11	7
210 to 220	111	3
		N = 188

Eleven class intervals were set up in conformance with Formulas No. 1 and No. 2. Formula No. 1 indicates from 7 to 11 class intervals for 2^8 or 256 items of data. Formula No. 2 indicates 12 or 13 class intervals. The fact that these data fall in a continuous series of dollars and cents from \$114.09 to \$217.08 makes \$10.00 class interval most convenient. The data in Worksheet No. 7 were tallied in Worksheet No. 8 and the totals accumulated in the frequency column. This frequency distribution is quite smooth and nearly normal. It shows a heavy concentration of data in the five middle classes between \$140.00 and \$190.00 which include 144 of the 188 items. The arithmetic mean for these data is computed by the three methods of (1) individual items in Worksheet No. 7 (2) long method with frequency distribution, Worksheet No. 9 and (3) short method with frequency distribution, Worksheet No. 10.

WORKSHEET NO. 9

COMPUTATION OF ARITHMETIC MEAN OF SEMI-MONTHLY EARNINGS OF 188 OIL REFINERY EMPLOYEES MARCH 1–16, 1951 BY LONG METHOD WITH CLASS INTERVALS

Class Intervals Dollars of Earnings i	Mid-points m	Frequencies f	Products of Mid-points and Frequencies fm
\$110 to \$120	115	2	230
120 to 130	125	5	625
130 to 140	135	11	1485
140 to 150	145	20	2900
150 to 160	155	26	4030
160 to 170	165	34	5610
170 to 180	175	36	6300
180 to 190	185	28	5180
190 to 200	195	16	3120
200 to 210	205	7	1435
210 to 220	215	3	645
		188	\$31,560

$$\overline{X} = \frac{\Sigma fm}{N} = \frac{\$31{,}560}{188} = \$167.872$$

The computation illustrated by Worksheet No. 9 is called the long method because it usually runs into large numbers and totals. Since the mid-point of each class is multiplied by its frequency, the product is certain to be a large number, frequently running into the thousands, or even millions. It is also called the long method in contrast with the short method illustrated in Worksheet No. 10.

ARITHMETIC MEAN BY SHORT METHOD

Meaning of symbols:

A = assumed mean, mid-point opposite zero (0) deviation.

x' = deviation from an assumed mean. It shortens the process to take these deviations in class intervals.

c = correction for assumed mean.

$$c = \frac{\Sigma fx'}{N} i$$

WORKSHEET NO. 10

COMPUTATION OF ARITHMETIC MEAN OF SEMI-MONTHLY EARNINGS OF 188 OIL REFINERY EMPLOYEES MARCH 1–16, 1951 BY SHORT METHOD WITH CLASS INTERVALS

Class Intervals Dollars of Earnings	Mid-points	Frequencies	Example 1		Example 2	
i	m	f	x'	fx'	x'	fx'
$110 to $120	115	2	− 5	− 10	− 8	− 16
120 to 130	125	5	− 4	− 20	− 7	− 35
130 to 140	135	11	− 3	− 33	− 6	− 66
140 to 150	145	20	− 2	− 40	− 5	− 100
150 to 160	155	26	− 1	− 26	− 4	− 104
160 to 170	165	34	0		− 3	− 102
170 to 180	175	36	+ 1	+ 36	− 2	− 72
180 to 190	185	28	+ 2	+ 56	− 1	− 28
190 to 200	195	16	+ 3	+ 48	0	
200 to 210	205	7	+ 4	+ 28	+ 1	+ 7
210 to 220	215	3	+ 5	+ 15	+ 2	+ 6
				+ 183		− 523
				− 129		+ 13
				+ 54		− 510

FORMULA NO. 6

$$\bar{X} = A + c, \text{ or } \bar{X} = A + \frac{\Sigma f(x')}{N} i$$

This short method is likely to seem longer and more complicated to the beginner, but in making a complete statistical analysis of a large sample, it is a great timesaver as will be clearly indicated in later chapters. It eliminates almost all decimal fractions and large totals.

Example 1

$$\bar{X} = A + \frac{\Sigma fx'}{N} i$$

$$= \$165 + \frac{54}{188} 10$$

$$= 165 + \frac{540}{188}$$

$$= 165 + 2.872$$

$$= \$167.872$$

Example 2

$$\bar{X} = A + \frac{\Sigma fx'}{N} i$$

$$= \$195 + \frac{-510}{188} 10$$

$$= 195 + \frac{5100}{188}$$

$$= 195 - 27.128$$

$$= \$167.872$$

In Worksheet No. 10 the arithmetic mean is computed from two assumed means to prove that no matter at what point the assumed mean is placed, the results will be the same. In Example 1, the assumed mean is 165, the mid-point of class 160–170. This is slightly smaller than the true mean. We, therefore, have a plus correction, + 54, in the fx' column, which signifies that the assumed mean must be increased by the amount of this correction $c = \left(\frac{\Sigma f(x')}{N} i = \frac{54}{188} 10 = 2.872\right)$, which when added to the assumed mean (165 + 2.872 = 167.872) gives the correct mean. In Example 2, the assumed mean, 195, the mid-point opposite zero (0) in the deviation column is too large, and, therefore, gives a correction value of − 510 in the fx' column, which means that the correction must be subtracted from the assumed mean in this case, $195 - c = 195 - \frac{\Sigma f(x')}{N} i = 195 - \frac{510}{188} 10 = 195 - 27.128 = 167.872.$ This is identical with the mean in Example 1. The same result

would be obtained by this method regardless of the class in which the zero (0) deviation is placed. In actual practice, it is better to place the zero (0) as near the middle of the frequency distribution as possible. The middle location reduces the size of the correction figures and shortens the work. This is one of the most important worksheets in all statistics. It should be mastered.

It should be observed that the mean computed by the long method is identical with that obtained from the short method. It is also worth noting that all these means obtained from frequency distributions are slightly different from that given in Worksheet No. 7 from the individual items of data. This will be the case in any frequency distribution of any data. In this case the difference is small, 167.872 against 167.856 dollars, or .016. In many series it will be larger. The reason for this variation is that in frequency distributions each item of data in each class is imputed or assumed to be worth the mid-point of its class. The frequent deviation of the value of data from the class mid-point makes this assumption slightly incorrect. This small error, however, is not sufficient to make any material difference in the practical or theoretical results. Since the short method has so many other advantages it should ordinarily be used for all samples of 100 or more items and may be used for much smaller samples.

In Worksheet No. 11 the arithmetic mean of the annual incomes of Federal employees subject to the Civil Service Retirement Act as of September 30, 1947 is found to be $2,986.87.

It should be noted that in all of these samples the mean is stated in units of the original data.

Worksheet No. 11 includes the long method and the short method for the sake of contrast and comparison. Ordinarily in computing a mean one would not use both methods, and need make only one of the solutions, which usually would be the short one. The contrast is clearly shown in Worksheet No. 11, in which the total of the short method is a −774,425 while that of the long method is 4,507,855,500. The greatest advantages of the short method are revealed in the next three chapters in which it

is used to compute from two to five or six statistics from one worksheet. It is a means of great economy in time and labor.

WORKSHEET NO. 11

COMPUTATION OF ARITHMETIC MEAN OF THE ANNUAL INCOMES OF FEDERAL EMPLOYEES SUBJECT TO THE CIVIL SERVICE RETIREMENT ACT AS OF SEPTEMBER 30, 1947 FOR SALARIES OF $10,000 OR LESS *

Short Method						Long Method
i		m	f	x'	fx'	fm
$ 0 to	$ 1000	$ 500	11,472	− 3	34,416	5,736,000
1000 to	2000	1500	133,645	− 2	267,290	200,467,500
2000 to	3000	2500	783,657	− 1	783,657	1,959,142,500
3000 to	4000	3500	405,331	0		1,418,658,500
4000 to	5000	4500	103,193	+ 1	103,193	464,368,500
5000 to	6000	5500	37,042	+ 2	74,084	203,731,000
6000 to	7000	6500	15,352	+ 3	46,056	99,788,000
7000 to	8000	7500	11,969	+ 4	47,876	89,767,500
8000 to	9000	8500	5,643	+ 5	28,215	47,965,500
9000 to	10000	9500	1,919	+ 6	11,514	18,230,500
			1,509,223		− 1,085,363	4,507,855,500
					+ 310,938	
					− 774,425	

* Source: *Statistical Abstract of U.S., 1949*, Table No. 235, p. 213.

Short Method

$$\bar{X} = A + \frac{\Sigma fx'}{N} i$$

$$= \$3{,}500 + \frac{-774{,}425}{1{,}509{,}223} 1{,}000$$

$$= \$3{,}500 - \frac{774{,}425{,}000}{1{,}509{,}223}$$

$$= \$3{,}500 - \$513.13$$

$$= \$2{,}986.87$$

Long Method

$$\bar{X} = \frac{\Sigma fm}{N}$$

$$= \frac{\$4{,}507{,}855{,}500}{1{,}509{,}223}$$

$$= \$2{,}986.87$$

The short method is also much more flexible than the long method and may be easily expanded to include other computations with a minimum of labor.

WEIGHTED ARITHMETIC MEAN

In the computation of index numbers and in averaging ratios and products or other means, it is necessary to weight the items in the computation of the average, if a correct result is to be obtained. As was shown on pages 81 and 82, Chapter 4, in averaging ratios it is necessary to weight each ratio by the number of items included in it. The same principle applies to percentages. From one point of view all arithmetic means are weighted. If no special specific weight is assigned to the several items included, each item has a weight of one. All are weighted equally. In computing averages from class intervals, the frequency distribution may be considered as a series of weights for the various mid-points. In many cases, however, specific and varying weights are assigned to the several items. For instance, in averaging the wages of different classes of workmen such as masons, carpenters, plasterers, painters, decorators, etc., drawing different rates of wages per hour, it would be necessary to weight the various wage rates with the number of persons receiving the several wages, or the number employed in each kind of work, in order to determine the correct average wage payment. If common labor received \$.30 per hour and skilled labor received \$1.20 per hour, the unweighted average would be:

$$\$.30 + \$1.20 = \$1.50 \div 2 = \$.75 \text{ per hour.}$$

But if there are four common laborers for each skilled worker, the true weighted average wage would be:

$$4 \times \$.30 + \$1.20 = \$2.40 \div 5 = \$.48 \text{ per hour.}$$

Meaning of symbols:

X = individual item of data.

W = weight, or number of times an item of data is counted.

FORMULA NO. 7

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$$

WORKSHEET NO. 12

WEIGHTED AVERAGE WAGE RATE OF 60 BUILDING TRADES WORKERS

Kind of Work	Hourly Wage Rates	Weights Number Employed	
	X	W	WX
Painters	$2.25	4	$ 9.00
Plasterers	2.50	3	7.50
Carpenters	2.00	8	16.00
Helpers	1.50	15	22.50
Laborers	1.00	30	30.00
	$9.25	60	$85.00

$$\bar{X} = \frac{\Sigma X}{N} = \frac{\$9.25}{5} = \$1.85 \qquad \bar{X}_w = \frac{\Sigma WX}{\Sigma W} = \frac{\$85.00}{60} = \$1.417$$

The unweighted mean is $1.85 per hour, while the weighted mean is only $1.417. The reason that the weighted mean in this case is so much below the high wages of painters, plasterers, and carpenters is that there are so few of the highly paid workers and so many common laborers. If the weights were reversed and there were a larger number of skilled workers the weighted mean would be larger. This average is illustrated more fully in Chapter 18 on the making of the index numbers. In such computations a weighted mean is essential.

AVERAGING A SERIES OF MEANS

It is frequently desired to average a series of averages. In such cases it is necessary to weight each average with the number of items included in the computation of that separate mean. This may be illustrated by the average of per capita incomes of ten Oklahoma Counties, 1951.

This type of average is required quite frequently in department stores, multiple plant industries, chain stores, and similar business activities to convert simple single means into an overall measure for the total industry or plant.

WORKSHEET NO. 13

COMPARISON OF WEIGHTED AND UNWEIGHTED MEANS OF PER CAPITA INCOME OF THE TEN OKLAHOMA COUNTIES WITH THE HIGHEST PER CAPITA INCOMES, 1949

County	Per Capita Incomes Per County	Weights Total Population Per County	Total County Income or Weighted County Averages
Washington	$1,776	32,880	58,394,880
Tulsa	1,705	251,686	429,124,630
Oklahoma	1,602	325,352	521,213,904
Alfalfa	1,329	10,699	14,218,971
Grant	1,283	10,461	13,421,463
Kay	1,205	48,892	58,914,860
Garfield	1,196	52,820	63,172,720
Harper	1,189	5,977	7,106,653
Stephens	1,126	34,071	38,363,946
Woods	1,080	14,526	15,688,050
	$13,491	787,363	$1,219,620,107

$$\bar{X} = \frac{\Sigma X}{N} = \frac{\$13,491}{10} = \$1,349.10$$ Unweighted average of averages (incorrect).

The weighted mean of per capita income of the ten Oklahoma Counties is:

$$\bar{X}_W = \frac{\Sigma WX}{\Sigma W} = \frac{\$1,219,620,107}{787,363} = \$1,549.$$

Weighted Mean = $1,549.00
Unweighted Mean = 1,349.10
Error of Unweighted Mean = $ 199.90

To give the correct average of the average or per capita income by counties for counties, it is necessary to weight the county average per capita income by the population of the county. In fact this method of weighting amounts to changing average per capita incomes back to the original total incomes.

SUMMARY OF ARITHMETIC MEAN

1. The arithmetic mean may be computed from the original unorganized data by dividing the sum of the items by their number.

2. It may be computed from data organized into class intervals by multiplying the mid-point of each class by the class frequency, summing these products for all classes and dividing this sum by the number of items. This process is called the *long method.*

3. It may be computed by assuming the mid-point of some class to be the mean and then computing the algebraic sum of the class products of the class deviations times the class frequency, dividing the sum by N and adding algebraically this correction sum to the assumed mean. This is called the *short method* because it saves time and labor.

4. The weighted arithmetic mean is computed by multiplying each item, or each mid-point, by a number, called a weight, which includes the item in the sum as many times as there are units in the weight, and dividing the summed weighted items by the sum of the weights.

5. The arithmetic mean is a complete measure of central tendency because it includes both the *number* and the *size* of the averaged items.

6. It is the point which exactly balances the frequency distribution of absolute numbers, so that there is equal weight (number $\times$ size) on either side of it.

7. It is sound algebraically and may be easily manipulated in equations.

8. It is the most useful and widely used average in statistics for normal distributions of absolute numbers.

GEOMETRIC MEAN

The geometric mean is the nth root of the product of n factors. It is useful for averaging many kinds of ratios and percentage changes. It is especially useful in making price relative index numbers. If we wish to average the percentage change in the price of cotton when it falls from \$.10 a pound to \$.05 a pound,

and later rises to \$.10 a pound, only the geometric mean will give us the correct answer.

	Price of Cotton	*Average of Price Relatives by Arithmetic Mean*
		%
1931.....	\$.10	50, fell to 50% of 1931 price
1932.....	.05	200, rose to 200% of 1932 price
1933.....	.10	$\frac{250}{2}$ = 125% change (incorrect)

Average of Price Relatives by Geometric Mean

$$\sqrt[2]{50 \times 200} = \sqrt[2]{10{,}000} = 100\% \text{ (correct)}$$

It is quite evident that if the price of cotton was \$.10 in 1931 and \$.10 in 1933, that the latter price is 100% of the former price, even if there had been a fall and a rise in the meantime. The arithmetic mean makes the 1933 price of \$.10 125% of the 1931 price of \$.10, which is clearly incorrect. The geometric mean alone will follow through these price changes with the correct result. In the vast economic field of price changes, its use is necessary and satisfactory. Many index numbers are computed by the use of the geometric mean.

When the data are manipulated by single items and not by class intervals, the formula is:

FORMULA NO. 8

$$G = \sqrt[N]{X_1 \cdot X_2 \cdots X_n}$$

The geometric mean of 2 and 8 is 4. $\sqrt[2]{2 \times 8} = \sqrt[2]{16} = 4$. That of 3, 6, and 12 is 6. $\sqrt[3]{3 \times 6 \times 12} = \sqrt[3]{216} = 6$. The difficulty of extracting high roots is obviated by the use of logarithms.

FORMULA NO. 9

$$\text{Log } G = \frac{\Sigma(\text{Log } X_1 + \text{Log } X_2 + \cdots + \text{Log } X_n)}{N}$$

or abbreviated,

$$\text{Log } G = \frac{\Sigma(\text{Log } X)}{N}$$

WORKSHEET NO. 14

X	Log X
104	2.01703
95	1.97772
112	2.04922
97	1.98677
82	1.91381
	9.94455

$$\text{Log } G = \frac{\Sigma(\text{Log } X)}{N}$$

$$= \frac{9.94455}{5}$$

$$= 1.98891$$

G = antilog of 1.98891

$G = 97.48$

For the weighted geometric mean so widely used in price index numbers and percentage changes, the formula is:

Weighted Geometric Mean

FORMULA NO. 10

$$\text{Log } G_W = \frac{\Sigma(W_1 \text{ Log } X_1 + W_2 \text{ Log } X_2 + \cdots + W_n \text{ Log } X_n)}{\Sigma W}$$

WORKSHEET NO. 15

X	Log X	W	W Log X
104	2.01703	12	24.20436
95	1.97772	7	13.84404
112	2.04922	10	20.49220
97	1.98677	2	3.97354
82	1.91381	1	1.91381
		32	64.42795

$$\text{Log } G_W = \frac{\Sigma(W \text{ Log } X)}{\Sigma W} = \frac{64.42795}{32} = 2.013373$$

G_W = antilog of 2.013373, or 103.13

The reason the weighted geometric mean in this particular case is so much larger than the unweighted geometric mean from the same data is that the larger numbers bear the heavy weights, 104 is weighted with 12, and 112 with 10, while 97 has a weight of 2 and 82 only 1. If the heavy weights had been placed on the small numbers, the results would have been reversed.

Geometric Mean with Class Intervals

FORMULA NO. 11

$$\text{Log } G = \frac{\Sigma f \text{ Log } m}{N}$$

WORKSHEET NO. 16

COMPUTATION OF GEOMETRIC MEAN OF SALES RELATIVES SHOWING PERCENTAGE OF CHANGE IN TOTAL SALES OF DEPARTMENT STORES OF 135 U.S. CITIES FROM SEPTEMBER 1949 TO SEPTEMBER 1950 *

Sales Relatives	Mid-points	Frequencies	Logarithm of Mid-points	Frequencies × Logarithm of Mid-points
i	m	f	Log m	f Log m
— .5– 5.4	2.5	20	0.39794	7.95880
5.5–11.4	8.5	49	0.92942	45.54158
11.5–17.4	14.5	40	1.16137	46.45480
17.5–23.4	20.5	16	1.31175	20.98800
23.5–29.4	26.5	7	1.42325	9.96275
29.5–35.4	32.5	2	1.51188	3.02376
35.5–41.4	38.5	1	1.58546	1.58546
		135		135.51515

* Source: *Federal Reserve Bulletin*, November 1950, p. 1528.

$$\text{Log } G = \frac{\Sigma f \text{ Log } m}{N} = \frac{135.51515}{135}$$
$$= 1.003816$$
$$\text{Antilogarithm} = 10.088$$
$$= 10.09\%$$

When thirteen classes are used the geometric mean of these data is 10.01% or only .078 of one percent less, but the amount of labor in computation is doubled. The number of classes used here, seven, falls within the range indicated by Formula No. 1 in Table No. 19, page 100.

The arithmetic mean for these data would be 12.32% which would be an incorrect average for these price relative changes.

Since no price relative can fall below zero (0) but may rise to unlimited heights, the correct middle balance between these opposite movements can be computed only by a relative or geometric average. In this case the price relatives range between −.5 and 41.5. The absolute lower limit is 0, the upper limit is infinity. Prices cannot fall more than 100% but may rise hundreds of thousands or even millions of percent. To strike a correct balance between these converse ratios only the *n*th root of their products will suffice. This method reduces the large upward moving percentages changes to ratios equal to the smaller falling percentage changes. The arithmetic mean overemphasizes the large upward moving percentages by considering them as *absolute numbers* instead of *ratios* or *relative numbers* which they are.

This type of change is especially prevalent in the stock market in which prices sometimes rise from 100% to 200% or more, but rarely fall more than from 10% to 50%. The prices of commodities during World Wars I and II passed through the same degree of changes. Inflation and deflation are such frequent occurrences in modern business that the geometric mean is widely used.

CONTRAST OF GEOMETRIC AND ARITHMETIC MEANS

The relationship and contrast between the arithmetic mean and the geometric mean may be further illustrated by the following example.

Substitution Contrast

Arithmetic Mean		*Geometric Mean*
2	12.4	$2 \times 4 \times 8 \times 16 \times 32 = 32{,}768$
4	12.4	5th root of 32,768 $= 8$
8	12.4	$8 \times 8 \times 8 \times 8 \times 8 = 32{,}768$
16	12.4	
32	12.4	
62	62	

$$\bar{X} = \frac{\Sigma X}{N} = \frac{62}{5} = 12.4$$

1. The arithmetic mean of a series of items may be substituted for each item in the series and give the same *sum.*

The geometric mean of a series of items may be substituted for each item in the series and give the same *product.*

Deviation and Balance Contrast

$$\begin{aligned} 2 - 12.4 &= -\ 10.4 \\ 4 - 12.4 &= -\ 8.4 \\ 8 - 12.4 &= -\ 4.4 \\ &\quad -\ 23.2 \\ 16 - 12.4 &= +\ 3.6 \\ 32 - 12.4 &= +\ 19.6 \\ &\quad +\ 23.2 \\ &\quad 0 \end{aligned}$$

$$\frac{8}{2} \times \frac{8}{4} \times \frac{8}{8} \times \frac{16}{8} \times \frac{32}{8}$$

$$4 \times 2 = 2 \times 4$$

$$8 = 8$$

2. The deviations above and below an arithmetic mean are equal and total zero (0).

The product of the ratios above and below a geometric mean are equal.

Absolute vs. Relative Contrast

3. The arithmetic mean should be used to average absolute numbers.

The geometric mean should be used to average ratios, percentages, or relative numbers.

Simple and Compound Interest Contrast

4. The accumulation of compound interest is another relationship which can be measured correctly only by the method of the geometric mean.

Symbols

Po = amount of principal at beginning of period
Pn = amount of principal at end of period
r = rate of interest
N = number of years in the period.

If the interest is compounded annually the equation is:

$$Pn = Po\ (1 + r)^n$$

If we solve for r we have:

$$\frac{Pn}{Po} = (1+r)^n \quad \text{or} \quad \sqrt[n]{\frac{Pn}{Po}} = (1+r) \quad \text{or}$$

$$r = \sqrt[n]{\frac{Pn}{Po}} - 1$$

At what percent would \$500 double in 10 years?

$$r = \sqrt[10]{\frac{\$1000}{\$500}} - 1 = \sqrt[10]{2} - 1$$

The logarithm of 2 is .30103. The logarithm of the 10th root of 2 is $\frac{.30103}{10}$ or .030103. The antilogarithm of .030103 is 1.0718.

Therefore $r = \sqrt[10]{2} - 1 = 1.0718 - 1 = .0718$.
The rate of interest at which \$500 would become \$1,000 in ten years compounded annually is 7.18%.

Simple or Arithmetic Interest Rate

\$500	Principal
.0718	Rate
4000	
500	
3500	
\$35.90	interest for one year
10	10 years
\$359.00	= simple interest for 10 years
\$500.00	= compound interest for 10 years.

HARMONIC MEAN

The harmonic mean is based on the reciprocals of the numbers averaged. It is the reciprocal of the arithmetic mean of the reciprocals of the numbers averaged. It is restricted in its field of usefulness, but is necessary in averaging different rates of speed for equal given distances. One is no doubt familiar with the old problem in eighth-grade arithmetic, "If James can dig a ditch in two days, and John can dig it in three days, how long

will it take the two boys working together to dig it?" Such problems are based on the principle of the harmonic mean.

The following illustration will indicate the assumption which conditions its proper use. Two boys A and B ride bicycles. A travels 10 miles per hour. B travels 15 miles per hour. The two elements of the problem are time (T) and distance (D). Assumption I, *Both boys ride for two hours.* The speeds are weighted with *equal periods of time.* This assumption requires the arithmetic mean. The solution is as follows:

A travels 10 miles per hour
B travels 15 miles per hour
$\frac{25}{2} = 12.5$ miles per hour,

or in two hours A travels 20 miles and B travels 30 miles. The total time traveled is 4 hours, the total distance is 50 miles. $\frac{50}{4} = 12.5$ miles per hour. For the assumption of equal time periods the arithmetic mean is required.

Assumption II, *Both boys ride 30 miles.* The speeds are weighted in this case with *equal distances.* This assumption requires the harmonic mean. The solution is as follows:

A travels 30 miles in 3 hours
B travels 30 miles in 2 hours

Together they travel 60 miles in 5 hours. $\frac{60}{5} = 12$ miles per hour, or

A travels one mile in $\frac{1}{10}$ of an hour
B travels one mile in $\frac{1}{15}$ of an hour
$\frac{1}{10} + \frac{1}{15} = \frac{5}{30} \div \frac{2}{1} = \frac{5}{60}$.

They average a mile in $\frac{5}{60}$ of an hour or 12 miles per hour.

In averaging time rates when the various rates of speed are weighted by equal distances only the harmonic mean gives the correct result.[1]

During World War I, the United States Government statisticians had to use this mean to average the rates of speed of ships

[1] See "The Nature and Use of the Harmonic Mean," by Wirth F. Ferger, *Journal of the American Statistical Association*, March, 1931, p. 36.

carrying supplies to our armies in France. In World War II, we again find it necessary to measure the average load of supplies we can maintain in Ireland, Egypt, Australia, Alaska, and the East Indies. It is useful in averaging the supplies of iron ore that can be kept moving on our Great Lakes. Modern industry involves so much of speeds in work and transportation that the harmonic mean has a relatively more important place in present-day business than in the past.

FORMULA NO. 12, used with individual items

$$\frac{1}{H} = \frac{\Sigma\left(\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}\right)}{N}$$

FORMULA NO. 13, used with class intervals

$$\frac{1}{H} = \frac{\Sigma\left(f\frac{1}{m_1} + f\frac{1}{m_2} + \cdots + f\frac{1}{m_n}\right)}{N}$$

in which $\frac{1}{m}$ is the reciprocal of the mid-point of the class interval.

WORKSHEET NO. 17

COMPUTATION OF HARMONIC MEAN OF SPEED OF PLANETS AROUND THE SUN SHOWING DIFFERENCE BETWEEN AVERAGE SPEED OF MINOR PLANETS AND MAJOR PLANETS

Planets	Orbital Velocity Miles per Second	Reciprocal of Orbital Velocity	
Minor Planets			
Mercury	24.0	.041666667	
Venus	21.9	.045662100	
Earth	18.5	.054054054	
Mars	15.0	.066666667	
			.208050488
Major Planets			
Jupiter	8.1	.123456790	
Saturn	6.0	.166666667	
Uranus	4.2	.238095240	
Neptune	3.8	.263157890	
			.791376587

Mean Reciprocal for Minor Planets $\frac{.208050488}{4} = .052012622$

Harmonic Mean of Minor Planet's Velocity

$$\frac{1}{.052012622} = 19.22 \text{ miles per second.}$$

Mean Reciprocal for Major Planets

$$\frac{.791376587}{4} = .197844146$$

Harmonic Mean of Major Planet's Velocity

$$\frac{1}{.197844146} = 5.1 \text{ miles per second.}$$

$$\frac{\text{Harmonic Mean Velocity of Minor Planets } 19.22}{\text{Harmonic Mean Velocity of Major Planets } 5.1} = 3.77$$

Ratio of Mean Velocities 3.77 : 1.0

Contrast With Arithmetic Mean

The arithmetic mean of the velocity of the minor planets is 19.85 miles per second.

The arithmetic mean of the velocity of the major planets is 5.52 miles per second.

The ratio between the two mean speeds is $\frac{19.85}{5.52} = 3.58$ as compared with 3.77 for the harmonic mean.

Harmonic Mean of Speed of Santa Fé Trains

Mean Reciprocal of speeds of 6 fastest special service trains is

$$\frac{.81559717}{6} = .135932862$$

Harmonic Mean of speeds of 6 fastest special service trains is

$$\frac{1}{.135932862} = 7.36 \text{ hours} \left\{ \begin{array}{l} \text{Kansas City} \\ \text{to Chicago} \end{array} \right.$$

Mean Reciprocal of speeds of 14 slower accommodation trains is

$$\frac{1.33588490}{14} = .095420356$$

Harmonic Mean of speeds of 14 slower accommodation trains is

$$\frac{1}{.095420356} = 10.48 \text{ hours.} \left\{ \begin{array}{l} \text{Kansas City} \\ \text{to Chicago} \end{array} \right.$$

Ratio of mean schedules $\frac{10.48}{7.36} = 1.42$ 1.42 : 1.0 or 142 : 100

The 6 fastest special service trains on an average move one and one-half times as fast as the 14 slower accommodation trains.

Santa Fé Trains	*Arithmetic Mean*
Six fast trains	7.37 hours
Fourteen slower trains	10.53 hours

WORKSHEET NO. 18

COMPUTATION OF HARMONIC MEAN OF SPEED OF ATCHISON, TOPEKA AND SANTA FÉ PASSENGER TRAINS EAST AND WEST FROM CHICAGO TO KANSAS CITY, MO.

Train Number	Schedule Time in Hours	Reciprocals of Train Time	
1	6.9	.14492754	
2	7.1	.14084507	
3	7.5	.13333333	
4	7.5	.13333333	
5	7.6	.13157895	
6	7.6	.13157895	
			.81559717
7	9.4	.10638298	
8	9.5	.10526316	
9	9.5	.10526316	
10	9.8	.10204082	
11	10.3	.09708738	
12	10.5	.09523810	
13	10.6	.09433962	
14	10.8	.09259259	
15	10.8	.09259259	
16	10.9	.09174312	
17	11.1	.09009009	
18	11.2	.08928571	
19	11.3	.08849558	
20	11.7	.08547009	
			1.33588499

The Harmonic Mean can be computed from class intervals with Formula No. 13 by using reciprocals instead of logs in Worksheet No. 16.

MEDIAN

The median for a discrete series may be defined roughly as the middle item of a series or an array of data. For a continuous series, it is that *point* in the range of the data that divides the data into two equal parts so that there are just as many items on one side of it as on the other. As the name implies, the median is the middle of the data. It is a less complete measure than the mean. It takes into consideration only the number of the items, disregarding their size completely. The arithmetic mean of the seven items, 5, 6, 6, 7, 9, 12 and 200 = 35. The median of these same seven items is the middle item, or 7. The large item, 200, has no more effect on the median than the smallest item of 5.

The steps necessary to locate the median in unorganized data without class intervals are as follows:

1. Arrange the data in an array from the smallest item to the largest. This method is revealed in Worksheet No. 19.
2. Select the middle item, or if there is an even number, the simple mean of the two middle items.

Median from Ungrouped Data

Data: 7, 12, 8, 5, 9, 10, 9
Array: 5, 7, 8, 9, 9, 10, 12

Middle item = median = 9.

Median from Data in Class Intervals

FORMULA NO. 14

$$Me = L + \frac{p}{f} i, \qquad \frac{N}{2} = \text{one-half of items, the number on each side of the median.}$$

Meaning of symbols:

Me = median
L = lower limit of class interval in which median falls
f = frequency of median class interval
p = part of median class frequency used to make up half of items
N = total number of items in sample

WORKSHEET NO. 19

COMPUTATION OF MEDIAN QUARTILES, QUINTILES, DECILES, AND PERCENTILES FROM FREQUENCY DISTRIBUTION OF SEMI-MONTHLY INCOMES OF 188 OIL REFINERY EMPLOYEES, MARCH, 1951.

Class Intervals in Dollars of Income i	Frequencies f	Cumulated Frequencies	
		Forward	Backward
\$110 to \$120	2	2	188
120 to 130	5	7	186
130 to 140	11	18	181
140 to 150	20	38	170
150 to 160	26	64	150
160 to 170	34	98	124
170 to 180	36	134	90
180 to 190	28	162	54
190 to 200	16	178	26
200 to 210	7	185	10
210 to 220	3	188	3
	188		

$\frac{N}{2} = \frac{188}{94}$

Lower limit of Median class = 160

$Me = L + \frac{P}{f} i$

$= \$160 + \frac{30}{34} 10$

$= \$160 + \frac{300}{34}$

$= \$160 + 8.82$

$= \$168.82$

94 = one-half of items
64 = sum of frequencies up to median class
30 = P = part of median class frequency required to make $\frac{1}{2}N$.

The median in this continuous series of data is that point on the range of \$110. to \$220. which divides the 188 items into two groups of 94 items each, on each side of the median group. The method of computing it from class intervals is as follows:

1. Divide N by 2 to determine the number of items on each side of the median point.

2. Accumulate the frequencies from the top of the frequency

column up to but not including the median frequency, as in this case, $2 + 5 + 11 + 20 + 26 = 64$. At this point the median has not been reached, but when the next frequency is added the median has been passed. The median, therefore, in this case is in the class, 160 to 170, and the frequency in which the median is located is 34.

3. Subtract 64 from 94. This leaves 30, or 94–64, to be taken out of the frequency of the median class of 34.

4. Up to 160 on the range scales there is a total of 64 items. To this value must be added $\frac{30}{34}$ of the width of the median class of 10, which is $\frac{30}{34}10$, or $\frac{300}{34}$, or 8.82.

5. The lower limit of the median class, 160, plus 8.82, or \$168.82 locates the median point on the range scale or X-axis.

First and Third Quartiles

FORMULA NO. 15

$$Q_1 = L + \frac{P}{f}i, \qquad \frac{N}{4} = \frac{188}{4} = 47$$

$Q_1 = \$150 + \frac{9}{26}10$
$= \$150 + \frac{90}{26}$
$= \$150 + 3.46$
$= \$153.46$

47 = one-fourth of items
38 = sum of frequencies up to first quartile class
9 = P = part of first quartile class frequency required to complete $\frac{1}{4}N$.
L = lower limit of first quartile class = \$150.

FORMULA NO. 16

$$Q_3 = L + \frac{P}{f}i, \qquad \frac{3N}{4} = \frac{3 \times 188}{4} = \frac{564}{4} = 141$$

$Q_3 = \$180 + \frac{7}{28}10$
$= \$180 + \frac{70}{28}$
$= \$180 + 2.50$
$= \$182.50$

141 = three-fourths of items
134 = sum of frequencies up to third quartile class.
7 = P = part of third quartile class frequency required to make $\frac{3}{4}N$.
L = lower limit of third quartile class = 180

Quintiles

FORMULA NO. 17

$$Q_{n_1} = L + \frac{P}{f} i \quad , \quad \frac{N}{5} = \frac{188}{5} = 37.6$$

$Q_V = \$140 + \frac{19.6}{20}10$ — 37.6 = one-fifth of items

$= \$140 + \frac{19.6}{20}$ — 18.0 = sum of frequencies up to first quintile class.

$= \$140 + 9.80$ — 19.6 = P = part of first quintile class frequency required to make $\frac{1}{5}N$.

$= \$149.80$ — L = lower limit of first quintile class = 140

The second, third, and fourth quintiles may be computed from the same general formula with the exception that the second quintile requires $\frac{2N}{5}$, the third $\frac{3N}{5}$, and the fourth $\frac{4N}{5}$ for the division of the sample items. Since the quintiles divide the sample into five sections they make possible a more detailed and minute analysis of the data than do the quartiles. They are frequently used in reducing educational and social data but are of equal value in sales and personnel management studies.

Sextiles, Septiles, Octiles, Deciles, and Percentiles

A frequency distribution may be divided into as many sections or segments as the type of analysis desired requires. All of the above measures may be obtained from the same general formula used to compute the median, $Me = L + \frac{p}{f} i$.

They are:

$$\text{Sextiles, Sex.} = L + \frac{p}{f} i, \frac{N}{6}$$

$$\text{Septiles, Sep.} = L + \frac{p}{f} i, \frac{N}{7}$$

$$\text{Octiles, Oct.} = L + \frac{p}{f} i, \frac{N}{8}$$

$$\text{Deciles, Dec.} = L + \frac{p}{f}\, i, \frac{N}{10}$$

$$\text{Percentiles, P.} = L + \frac{p}{f}\, i, \frac{N}{100}$$

Deciles which divide the data into tenths, and percentiles which split a series into hundredths are more widely used than the others. They are frequently employed in making a detailed analysis of sales distributions, personnel data, or in population, or in educational and psychological analyses and measuring scales. For the earnings of the Oil Refinery employees these measures are:

$$\text{Sex}_1 = \$140 + \frac{13.67}{20}\,10 \qquad \frac{N}{6} = \frac{188}{6} = 31.67$$
$$= \$140 + 6.84$$
$$= \$146.84$$

$$\text{Sep}_2 = \$150 + \frac{15.71}{26.00}\,10 \qquad \frac{2N}{7} = \frac{376}{7} = 53.71$$
$$= \$150 + 6.04$$
$$= \$156.04$$

$$\text{Oct}_3 = \$160 + \frac{6.5}{34}\,10 \qquad \frac{3N}{8} = \frac{564}{8} = 70.50$$
$$= \$160 + 1.91$$
$$= \$161.91$$

$$\text{Dec}_7 = \$170. + \frac{33.6}{36}\,10 \qquad \frac{7N}{10} = \frac{1316}{10} = 131.6$$
$$= \$170 + 9.33$$
$$= \$179.33$$

$$P_{33} = \$150 + \frac{24.04}{26}\,10 \qquad \frac{33XN}{100} = \frac{6204}{100} = 62.04$$
$$= \$150 + 9.25$$
$$= \$159.25$$

Percentiles give a very minute division of a frequency distribution and usually are used to distribute the items of a relatively large sample either in educational test grades, sales budget quotas, or social or economic measuring scales. Percentiles serve as a super measuring scale for sample of 500 to 1000 items or more.

The median and all the related smaller measures are purely

locational values and entirely exclude the size of the several individual sample items. They are, therefore, all less complete measures than the means.

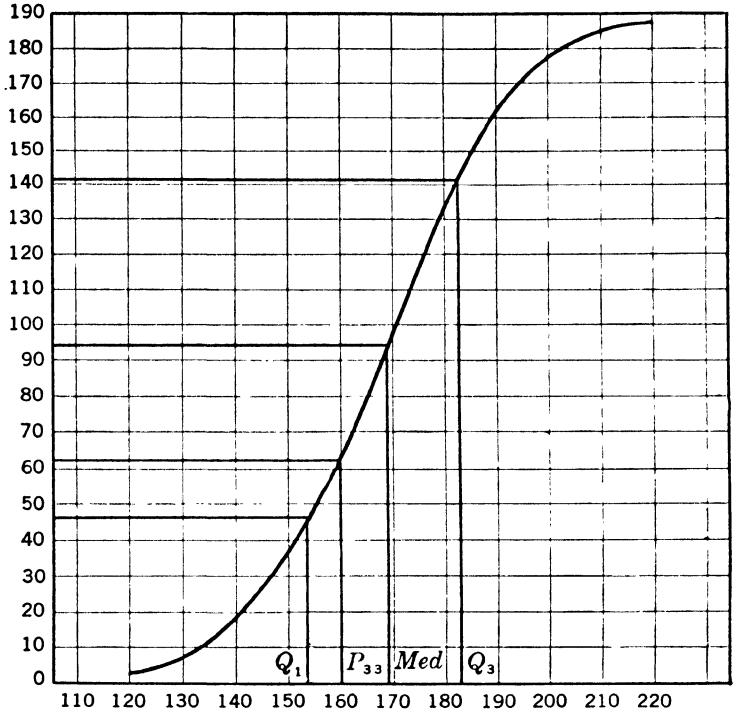


FIG. 42. Graphic method of computing locational measures. (See Worksheet No. 18)

Method of Constructing Median, Quartile, Decile, Percentile Chart

1. Set the class intervals on the X-axis, and the frequencies on the Y-axis.

2. Plot the accumulated frequencies in column 3, Worksheet No. 19, so that the accumulated frequency in each case falls at the right-hand end of the appropriate class interval. Two is plotted above 120, 7 is plotted above 130, 18 above 140, 38 above 150, and so on to the last figure, or total, 188, which is plotted at the end of the last class interval, or above 220.

3. The median, quartiles, deciles, and percentiles are located on the X-axis as follows:

a. For the median, follow up the Y-axis from zero to $\frac{N}{2}$, or 94 in this case. Place a square so that the point touches the diagonal frequency curve line and the top edge of the horizontal arm touches 94, the median frequency on the Y-axis. In this position the right edge of the vertical arm will cross the X-axis at the location of the median.

b. For the first quartile, move the square down toward zero on the Y-axis, being careful to keep the point on the frequency curve line. When the upper edge of the horizontal arm crosses the Y-axis at $\frac{N}{4}$, the vertical arm will cross the X-axis at the point of the first quartile.

c. To locate the third quartile, the square is moved up and to the right, keeping the point on the frequency curve, until it reaches $\frac{3N}{4}$ on the Y-axis. At this location the vertical bar will cross the X-axis at the point of the third quartile.

d. Any decile, percentile or other ratio may be located quickly by this method. When a frequency distribution is to be divided into many parts, this is an easy and sufficiently accurate method for doing it.

MODE

The mode is the least exact and least stable of all the averages. Its primary use is in quickly pointing out the value of the most numerous group of items. While the median takes into consideration the entire number of items, the mode considers only the most numerous group. It is especially unstable because a slight change in the sample may shift the mode a considerable distance as indicated below. The mode is 9, the most numerous group.

Mode in Ungrouped Data

Data: 7, 12, 8, 5, 9, 10, 9
Array: 5, 7, 8, 9, 9, 10, 12

If this sample were changed to 5, 5, 7, 8, 9, 10, 12, the mode would be 5. A shift of one item moved the mode from 9 to 5, decreasing its size almost 50 percent. If the distribution had been 5, 5, 7, 8, 9, 9, 10, 12, there would have been two modes a considerable distance apart, both contending for recognition. Such bi-modal or tri-modal distributions are frequent, and when they occur invalidate the use of the mode. For this reason the mode is employed only as a rough measure of the concentration of data.

Mode by Formula from Class Intervals

Symbols for interpolation formula:

Mo = mode

f_1 = frequency of class interval next smaller in size than the modal class

f_2 = frequency of class interval next larger in size than the modal class.

Mode by Interpolation Formula

FORMULA NO. 18

$$Mo = L + \frac{f_2}{f_2 + f_1} i$$

$$= 170 + \frac{28}{28 + 34} 10$$

$$= 170 + \frac{280}{62}$$

$$= 170 + 4.52$$

$$= 174.52$$

Mode by Locational Formula

FORMULA NO. 19

$$Mo = \bar{X} - 3(X - Me)$$

$$= 167.87 - 3(167.87 - 168.82)$$

$$= 167.87 - 3(-.95)$$

$$= 167.87 + 2.85$$

$$= 170.72$$

The mode is of such small practical consequence in most statistical studies that the student need spend little time on it. Most modern texts give it little space. It is of little use in scientific measurements or in practical business administration. We speak of the "mode" in the fashions of women's clothes, and of new automobiles. Otherwise it is rarely used.

GENERAL SUMMARY OF AVERAGES

Averages, or measures of central tendency, as they are often called, are quite different in methods of computation, purpose of use and characteristics and limitations. The following characterization of each one should be carefully studied by the student.

1. *Arithmetic Mean*

(a) The arithmetic mean is the most widely used and the most useful measure of central tendency in statistics.

(b) The mean is the most complete and adequate average, because it takes into consideration both the total number of items and their size. It is the point of balance in the data, the center of gravity.

(c) It is a computed or mathematical measure and can be adequately manipulated in algebraic equations.

(d) Its greatest weakness is that in extremely skewed distributions it tends to be pulled too far toward the extreme items, out of the center of the distribution, and ceases accurately to represent the large body of the data. In such cases the median or even the mode is more representative of the data.

(e) The arithmetic mean is used primarily to measure natural or absolute numbers, such as pounds, tons, inches, feet, yards, miles, gallons, barrels, meters, dollars, or other common denominate numbers. It is the logical measure for comparing two or more samples in units of such numbers. Since the mean is a logical representative value of a group of data, means are widely used in ranking samples or groups of values such as production in two or more plants, sales of two or more stores, and similar comparisons.

2. *Geometric Mean*

(a) The geometric mean has all the strong points and weaknesses of the arithmetic mean, and in addition,

(b) The geometric mean is used to average ratio or percentage changes. It is, therefore, of great value in making averages of price relative index

numbers. It gives less emphasis to extreme items than does the arithmetic mean.

(c) By the use of logarithms it is quite easy to compute and may be manipulated algebraically.

(d) It cannot be computed by logarithms for minus numbers or series containing zero.

3. *Harmonic Mean*

(a) The harmonic mean is more limited in its field of usefulness than the geometric mean, but in the domain of time, rate changes, or speeds, it is essential in order to obtain correct results.

(b) The harmonic mean gives still less emphasis to extreme items than does the geometric mean.

(c) It is easy to compute by means of reciprocals, but unless the reciprocals are carried out to at least seven or eight places, some inaccuracy will result.

4. *Median*

(a) The median is an incomplete average, because it takes into consideration only the number of items and not their size. In locating the median the smallest item of data counts for just as much as the largest one. It is based on a counting of noses and not on a weighing of minds or bodies.

(b) Since it falls in the middle of the data, in extremely skewed distributions, it is more representative of the large body of the data than is the mean.

(c) It is a locational and not a computed average and cannot be manipulated algebraically.

5. *Mode*

(a) The mode is the least complete and least dependable of all averages.

(b) It is a locational and not a computed average and cannot be manipulated algebraically.

(c) It is primarily useful only for representing the large central group, or most numerous frequency of the data.

(d) In some distributions there may be two or more modes, bi-modal or tri-modal centers of equal concentration, which make the mode largely useless in such cases.

(e) In extremely skewed distributions it is usually too near one end of the data to be very representative.

Selecting an Average

In analyzing data, the statistician should first study the nature of his data carefully and the use he wishes to make of it, and then select that

TABLE 20

SUMMARY TABLE OF ALL MEASURES COMPUTED ON THE WHEAT YIELDS OF 160 FARMS

Measure	Formula	Results
Arithmetic Mean, Ungrouped Data	$\bar{X} = \frac{\Sigma X}{N}$	167.86
Arithmetic Mean, Class Intervals Long Method	$\bar{X} = \frac{\Sigma fm}{N}$	167.87
Arithmetic Mean, Class Intervals Short Method	$\bar{X} = A + \frac{\Sigma fx'}{N} i$	167.87
Median, Class Intervals	$Me = L + \frac{p}{f} i$	168.82
First Quartile	$Q_1 = L + \frac{p}{f} i$	153.46
Third Quartile	$Q_3 = L + \frac{p}{f} i$	182.50
Mode Interpolation	$Mo = L + \frac{f_2}{f_2 + f_1} i$	174.52
Mode Locational	$Mo = \bar{X} - 3(\bar{X} - Me)$	170.72

average which will give the best results as to accuracy of representation and needs for further manipulation. Any technician, to get good results, must be well acquainted with the material on which he works. It is not alone sufficient to be an expert with tools. One must know the field in which those tools are to be applied. A cabinet maker, to succeed well, must know the nature and structure of the various woods with which he must work as well as know the exact use of his tools. Statistics is a method of quantitative analysis. The statistician is a technician. To compute an adequate, a good, average, or other statistic, he must be well informed in the field of data in which he works. A biologist who knows statistics can do much better statistical work in the field of biology than a statistician who is not acquainted with that science. The same principle holds for chemistry, economics, sociology, psychology, business administration, and all other fields of knowledge. A good research man must know both his statistical tools and the field of data in which he works. Of all statistical tools, the average is the most basic, essential, and fundamental.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapters XVI, XVII. Irwin, Inc., Chicago, 1942.

Chaddock, R. E., *Principles and Methods of Statistics*, Chapters VI, VII, VIII. Houghton Mifflin Company, Boston, 1925.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter IX. Prentice-Hall, Inc., New York, 1948.

Garrett, H. E., *Statistics in Psychology and Education*, Chapter I. Longmans, Green & Company, New York, 1926.

Mills, F. C., *Statistical Methods, Revised*, Chapter IV. Henry Holt and Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapter II. John Wiley & Sons, New York, 1942.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chapter III. Harcourt, Brace & Company, New York, 1935.

Waugh, A. E., *Elements of Statistical Method*, Chapter IV. McGraw-Hill Book Company, New York, 1938.

REVIEW QUESTIONS

1. Give three reasons for using averages.
2. Write four formulas for computing the arithmetic mean, and explain the differences between them.
3. Are all arithmetic means weighted or not? Why?
4. Explain the difference between the geometric mean and the arithmetic mean.
5. Explain why and how the geometric mean is adapted to averaging percentage changes.
6. What is the principal use of the harmonic mean and why does it meet this requirement?
7. What are the differences between the harmonic and geometric means and the methods of computing them?
8. What are the relative sizes of the arithmetic, harmonic, and geometric means?
9. Which gives the more complete measure of data, the median or the mean? Why?
10. What is the formula for the second quartile? How does it differ from the first and third quartiles?
11. Explain the use and defects of the mode.
12. For what kind of sample might the median be a more truly representative measure than the mean? Why?
13. What is a decile? What is a percentile?

14. What is a logarithm, an antilog, a reciprocal, a frequency?
15. On what bases does one decide what particular average to use for any specific sample?
16. Of what use is an average in research work?

EXERCISES

1. Use all the frequency distributions computed for Chapter 5.

2. Yields of Wheat, bu.

Class Intervals	Frequencies (60 Farms)
7.5 to 10.5	5
10.5 to 13.5	9
13.5 to 16.5	18
16.5 to 19.5	15
19.5 to 22.5	8
22.5 to 25.5	3
25.5 to 28.5	2

3. Number of Rooms in Urban Homes, 1947

Class Intervals	Frequencies (10,000)
1 Room	47.
2 Rooms	175.
3 Rooms	337.
4 Rooms	461.
5 Rooms	547.
6 Rooms	463.
7 Rooms	186.

4. Number of Rooms in Rural Farm Homes

Class Intervals	Frequencies (10,000)
1 Room	7.5
2 Rooms	33.3
3 Rooms	77.9
4 Rooms	131.8
5 Rooms	119.7
6 Rooms	114.4
7 Rooms	77.5

5. Federal Male Employees Subject to Civil Service Retirement

Class Intervals (Salary)	Frequencies (in 1000)
$ 0. to 1000.	5.7
1000. to 2000.	86.7
2000. to 3000.	506.8
3000. to 4000.	364.3
4000. to 5000.	94.9
5000. to 6000.	34.7
6000. to 7000.	14.8
7000. to 8000.	11.7
8000. to 9000.	5.6
9000. to 10000.	1.9

6. Family's Purchase of Groceries in 26 Weeks, 1951

Class Intervals	Frequencies
$ 0 to 2.00	13
2.00 to 4.00	18
4.00 to 6.00	27
6.00 to 8.00	15
8.00 to 10.00	4
10.00 to 12.00	1

7. Grades of 226 Students in Elementary Economics, University of Tulsa, May 1951

Class Intervals (Points)	Frequencies
400 to 361	27
360 to 321	45
320 to 281	86
280 to 241	38
240 to 201	21
200 to 161	9

CHAPTER 7

DISPERSION, SKEWNESS, AND VARIATION

We have now learned to compute an average, a representative figure which measures the central value of a frequency distribution and may be used to represent it in comparison with other frequency distributions. By this method we can compare the standing of Class A with Class B, or the wages of carpenters and plumbers, or the annual income of physicians and lawyers. It is, however, the nature of the ordinary frequency distribution to scatter about its mid-point or average. In some distributions the items fall very close together, and, therefore, close to the mean. In such cases the mean represents the group quite well because it is the same size or nearly the same size of the other items in the group. In other distributions the items are widely scattered. Some of them are very much smaller than the mean, while others are very much larger. In such cases the average does not represent most of the individual items very well. This may be illustrated as follows:

	Items	*Mean*
Group 1	150, 150, 150, 150, 150, 150, 150, 150, 150	150
Group 2	146, 147, 148, 149, 150, 151, 152, 153, 154	150
Group 3	16, 25, 50, 112, 150, 188, 255, 370, 384	150

In Group 1 the mean is 150 which is exactly the same as all the other items in the distribution. It, therefore, represents all the items in the group perfectly.

In Group 2 the mean is 150. Only one sample item is this size, but the others cluster about it. The mean represents all these items.

In Group 3 a quite different condition exists. The mean is 150 again, but only one item in the group is anywhere near its size. It is nearly ten times as large as 16, the smallest number, and less than one-half the size of the largest figure, 384. It is evident at a glance that this average does not represent accurately the individual items in its group. The reason for this poor representation is the wide scatter of this group, from 16 to 384. In Group 1 the scatter about the mean was *zero*. In the second group it was *eight*, 146–154, inclusive. In Group 3 the scatter is 368. This illustration indicates the need for some measure of the scatter of the items in a sample about the sample average. The name given to this scatter is dispersion.

WORKSHEET NO. 20

DISPERSION OF GRADES OF TWO CLASSES IN STATISTICS, UNIVERSITY OF TULSA, SEMESTER ENDING MAY 28, 1951

Class A — 16 students	Class B — 13 students
Grades (540 points possible)	Grades (540 points possible)
503	532 Less Representative
500 More Representative	519 Average
500 Average	508
488	498
482	468
477←475.9 mean	467
473	458←458 median
472 } ←472 median	455←455.5 mean
472 }	425
471 Grade Range	419 Grade Range
470 503	401 532
469 553	398 372
466 50	373 160
460 Homogeneous Data	Heterogeneous Data
457 Narrow Range	Wide Range
454	
The middle one-half (8) of these students fall in the narrow range of 482–469, inclusive.	The middle one-half (7) of these students fall in a wide range of 498–419, inclusive.

RANGE

The simplest and largest measure of dispersion is the range. It is the total inclusive distance from the smallest to the largest items. If the smallest item is 9 and the largest item is 31, the range is $31 - 9 = 22$. The 9, smallest item is on the inside of the range as well as the largest item. If the data are thrown into class intervals as follows, 8.5–10.4 and so on to 30.5–32.4 the range is:

$$32.4 - (8.5 - .1) = 32.4 - 8.4 = 24.$$

By either method both the largest and smallest items must be included. Frequently the range for the same data in grouped and ungrouped forms will be slightly different. This difference is due to the fact that the class intervals may extend slightly beyond the actual data.

The range is somewhat more widely used at present than formerly as a measure of dispersion, especially in tabular analysis. Its principal defect is that it is based on only the two most extreme items and is in no way a measure of the scatter of the many intervening items. At the extremes it passes over several class intervals which have such small frequencies that they are negligible in value. They are as wide as other classes but their content is thin. Most of the items in most distributions tend to fall in a few classes near the center. The range tends to overemphasize scatter. A narrower weighted measure is more truly descriptive of the distribution.

The two statistics classes shown in Worksheet No. 20 illustrate this point. If we drop off three items from each end of Class A, the range is reduced from 51 to 22, more than one-half. By reducing the sample 37.5% the range is reduced 56.9%. In Class B, dropping the two extreme items reduces the range from 159 to 121. Near the centers of the arrays the concentration of items is so great that dropping a few more extreme items has little effect on the range.

The range is a locational measure depending on the two extreme items of the sample alone. It is, therefore, entirely unweighted and as such is inadequate for accurate measurements of dispersion.

QUARTILE DEVIATION

The quartile deviation is one-half the distance between the third and first quartiles.

FORMULA NO. 20

$$Q_d = \frac{Q_3 - Q_1}{2}$$

For the wage of the Oil Refinery employees the computation is:

$$Q_d = \frac{182.50 - 153.46}{2} = \frac{29.04}{2} = 14.52$$

Such a deviation is used plus and minus, or measured on both sides of the mid-point between the quartiles. In this case ± 14.52, or the full distance between the two quartiles, is 29.04, which is the *inter-quartile range.* While the full range of the wages of the Oil Refinery employees is 110, the quartile range is only 29.04, or only a little over one-fourth of the total range. Since fifty percent of any frequency distribution is included between the two quartiles, in this case 94 of the 188 items fall within the narrow $29.04 quartile range while the other 94 items are scattered over $110–$29.04, or remaining $80.96 of the range.

The quartile range is an excellent measure of the dispersion of the dense middle half of the data. The main objection to it is that it is *too* small. Just outside of it on either edge is a rather dense area of items, sufficiently numerous to be of real importance, which it excludes. Although the total range is too broad to be a good measure of dispersion, the quartile range is too narrow. Both measures are purely locational and are not weighted by the other items in the group.

MEAN DEVIATION

Another measure of dispersion is the *mean deviation,* based on the arithmetic mean of the deviation of all the items in a distribution from their median. The mean deviation may be computed from the arithmetic mean as a base, but the median is preferred

because the sum of the deviations of the items from the median is a minimum and is likely to be more stable and representative in comparing frequency distributions with various amounts of dispersion. The deviations are averaged without regard to sign. From ungrouped data it is computed as follows:

Meaning of symbols:

M_d = mean deviation

$|d| = |(X - Me)|$, the difference between the datum and the sample median. The bars | | mean that the signs are neglected.

FORMULA NO. 21

$$M_d = \frac{\Sigma|d|}{N}$$

WORKSHEET NO. 21

MEAN DEVIATION OF THE CIRCULATION OF TEN NATIONAL MAGAZINES, SEPTEMBER, 1950

Circulation (in 100,000) X	$-$	M_d	$=$	d
53.5	−	37.9	=	15.6
45.4	−	37.9	=	7.5
40.6	−	37.9	=	2.7
40.4	−	37.9	=	2.5
38.2	−	37.9	=	.3
37.6	−	37.9	=	.3
34.7	−	37.9	=	3.2
31.8	−	37.9	=	6.1
31.4	−	37.9	=	6.5
30.3	−	37.9	=	7.6
				52.3

$$M_d = \frac{\Sigma|d|}{N} = \frac{52.3}{10} = 5.23$$

or 523,000 copies

This formula is easy to use, but for large samples with a median involving decimal fractions it is time-consuming and tedious. Under such conditions it is more convenient to use class intervals and

FORMULA NO. 22

$$M_d = \frac{\Sigma f|(m - Me)|}{N}$$

For the wage data of Oil Refinery employees the mean deviation is $16.76 and when measured plus and minus from the median includes 106.27 of the 188 items or 56.5% of them.

The mathematical peculiarity of the mean deviation must be noted. It violates the basic principles of algebra in that while the deviations from all items or mid-points smaller than the median are minus in sign and the deviations from all items or mid-points larger than the median are plus, these significant algebraic signs are totally disregarded. They have to be disregarded if the mean deviation is to be used or even to exist. If the pluses were to be subtracted from the minuses or if their algebraic sum were taken, the net result would be zero or approximately zero in any symmetrical distribution. Instead of getting a full measure of the dispersion one would secure only a small residual of the dispersion. Since the mean deviation is mathematically unsound it does not lend itself to further analysis. This mathematical handicap of the mean deviation is one of the reasons that it is not widely used in statistical studies. It may be used as a basis of correlation, but for this and other purposes there is a superior measure. One of the chief reasons for teaching the theory and method of the mean deviation is that it is a necessary step in the development and understanding of this superior measure which is called *standard deviation.*

STANDARD DEVIATION

The mean deviation falls between the broad total range and the narrow quartile range. Measured ± from the median it includes 57.5% of the items in a normal distribution. Although it is a more

truly representative measure of dispersion than either of the others, it has the two defects that (1) it is unsound algebraically and (2) it still leaves outside its limits a considerable area of the dense portion of the frequency distribution. The standard deviation has none of these limitations and is generally considered by statisticians as by far the best measure of dispersion. It is based on squared deviations from the mean.

Meaning of symbols:

σ = standard deviation
x = deviation between an item of data or a mid-point and the arithmetic mean
N = number of items in sample

Formula No. 23, used only for individual items

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

WORKSHEET NO. 22

STANDARD DEVIATION OF THE CIRCULATION OF TEN NATIONAL MAGAZINES, SEPTEMBER, 1950

Circulation in 100,000 X		$\bar{X}$	=	x	x^2
53.5	−	38.39		+ 15.11	228.3121
45.4	−	38.39		+ 7.01	49.1401
40.6	−	38.39		+ 2.21	4.8841
40.4	−	38.39		+ 2.01	4.0401
38.2	−	38.39		− .19	.0361
37.6	−	38.39		− .79	.6241
34.7	−	38.39		− 3.69	13.6161
31.8	−	38.39		− 6.59	43.4281
31.4	−	38.39		− 6.99	48.8601
30.3	−	38.39		− 8.09	65.4481
					458.3890

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{458.3890}{10}} = \sqrt{45.8389}$$

= 6.77 or 677,000 copies

The mean deviation for these same data in Worksheet No. 21 is only 523,000 copies while the standard deviation is 677,000 copies. The greater size of the standard deviation is due to the fact that in computing σ, the x's are squared. This process weights the small deviations such as 2.01, .19 and .79 by small numbers 2.01, .19 and .79, but weights the large numbers 15.11, 7.01 and 8.09, by large numbers, 15.11, 7.01 and 8.09. This squaring of the deviations gives the items which are more distant from the mean the most importance, or weight, and thus increases the size of the standard deviation beyond that of the mean deviation.

Next to the arithmetic mean, the standard deviation is the most useful and most widely used measure in statistical analysis. These two measures are to the statistician what the ax and cross-cut saw are to the woodsman — the basic tools for working up his raw materials. In data composed of absolute numbers the arithmetic mean measures the central tendency of the group. The standard deviation measures the scatter of the items around this representative number.

One point should be emphasized at this place. The standard deviation should always be computed from deviations from the arithmetic mean instead of the median. When taken from the mean it is a minimum; that is, a standard deviation computed from the true arithmetic mean is smaller than one computed from any other average, and it is more stable and dependable as between various frequency distributions. The method illustrated in Worksheet No. 23 reveals every detail of the computation and all the principles and methods involved, but it is too long and laborious to use in large series of data. For such problems the data should always be thrown into class intervals and the methods used in Worksheets Nos. 10 or 24 employed.

Long Method with Class Intervals

FORMULA NO. 24

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}}$$

WORKSHEET NO. 23

COMPUTATION OF STANDARD DEVIATION FROM CLASS INTERVALS BY LONG METHOD FOR SEMI-MONTHLY WAGES OF OIL REFINERY EMPLOYEES, MARCH, 1951

Class Intervals Income in Dollars i	Mid-points m	Frequencies f	$\bar{X}$ = \$167.87 x $(m - \bar{X})$	Deviations Squared x^2	Frequencies Times Squared Deviations fx^2
\$110 to \$120	115	2	− 52.87	2,795.2369	5,590.4738
120 to 130	125	5	− 42.87	1,837.8369	9,189.1845
130 to 140	135	11	− 32.87	1,080.4369	11,884.8059
140 to 150	145	20	− 22.87	523.0369	10,460.7380
150 to 160	155	26	− 12.87	165.6369	4,306.5594
160 to 170	165	34	− 2.87	8.2369	280.0546
170 to 180	175	36	+ 7.13	50.8369	1,830.1284
180 to 190	185	28	+ 17.13	293.4369	8,216.2332
190 to 200	195	16	+ 27.13	736.0369	11,776.5904
200 to 210	205	7	+ 37.13	1,378.6369	9,650.4583
210 to 220	215	3	+ 47.13	2,221.2369	6,663.7107
		188			79,848.9372

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{79,848.9372}{188}} = \sqrt{424.728} = \$20.6$$

The mean deviation for these data is only \$16.76, but the standard deviation is \$20.60. Measured plus and minus from the mean, \$167.87, it reaches on the class-interval, x-axis, scale from (\$167.87 ± \$20.60) \$147.27 to \$188.47 and includes \$125.176 items out of the 188 in the sample, or 66.6% of the data. In a perfectly normal frequency distribution of large size, a ± σ includes 68.27% of the items in the distribution. This distribution of the wages of Oil Refinery employees deviates slightly from normal in both skewness and kurtosis which accounts for the fact that the percentage of items in the one ± σ is a little smaller than is to be expected in a normal distribution. See Appendix Table II.

The above method of computing the standard deviation from the true mean is long and tedious because it usually involves long decimal fractions, and runs into large numbers. The following method is much shorter and should be thoroughly mastered by the student. It is accurate, easy to compute and widely used.

Short Method with Class Intervals

If one will duplicate Worksheet No. 10 (page 118) for computing the arithmetic mean by the short method and add one more column, $f(x')^2$, the arithmetic mean and standard deviation both may be computed at the same time with a minimum of labor.

Meaning of symbols:

σ = standard deviation computed from the true arithmetic mean, or the true standard deviation
S = standard deviation computed from an assumed mean
c = correction for the standard deviation computed from an assumed mean
x' = deviation from an assumed mean

Formulas:

$$\sigma^2 = s^2 - c^2 \qquad \text{SUMMARY FORMULA NO. 25}$$

$$S^2 = \frac{\Sigma f(x')^2}{N} \qquad \text{FORMULA NO. 26}$$

$$c^2 = \left(\frac{\Sigma f(x')}{N}\right)^2 \qquad \text{FORMULA NO. 27}$$

$$\sigma = i\sqrt{\frac{\Sigma f(x')^2}{N} - \left(\frac{\Sigma f(x')}{N}\right)^2} \qquad \text{FORMULA NO. 28}$$

The first five columns of Worksheet No. 10 are identical with those in Worksheet No. 24, which provides the arithmetic mean by the short method. By the simple device of adding a sixth column, which is the product of the fourth and fifth columns, the standard deviation is secured.

WORKSHEET NO. 24

COMPUTATION OF STANDARD DEVIATION FROM CLASS INTERVALS BY SHORT METHOD FOR SEMI-MONTHLY WAGES OF OIL REFINERY EMPLOYEES, MARCH, 1951

Class Intervals Income in Dollars	Mid-points	Frequencies	Deviations from Assumed Mean	Frequencies Times Deviations from Assumed Mean	Frequencies Times Squared Deviations from Assumed Mean
i	m	f	x	fx'	$f(x')^2$
\$110 to \$120	115	2	− 5	− 10	50
120 to 130	125	5	− 4	− 20	80
130 to 140	135	11	− 3	− 33	99
140 to 150	145	20	− 2	− 40	80
150 to 160	155	26	− 1	− 26	26
160 to 170	165	34	0		
170 to 180	175	36	+ 1	+ 36	36
180 to 190	185	28	+ 2	+ 56	112
190 to 200	195	16	+ 3	+ 48	144
200 to 210	205	7	+ 4	+ 28	112
210 to 220	215	3	+ 5	+ 15	75
		188		+ 183	814
				− 129	
				+ 54	

$$\sigma = i\sqrt{\frac{\Sigma f(x')^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2}$$

$$= 10\sqrt{\frac{814}{188} - \frac{54}{188}} = 10\sqrt{4.3298 - .0825}$$

$$= 10\sqrt{4.2473} = 10. \times 2.06 = \$20.6$$

$$\overline{X} = A + \frac{\Sigma fx'}{N} i = 165 + \frac{54}{188} 10 = 165\frac{540}{188}$$

$$= 165 + 2.87 = \$167.87$$

Since the x's in Worksheet No. 24 on the wages of Oil Refinery employees are taken in class intervals, the standard deviation 2.06 is also in class intervals; $\sigma = 2.06$ class intervals. To change this

number to units of the original data, which are in class intervals ($10) ten dollars wide, the 2.06 must be multiplied by $10, which gives the true standard deviation as $20.60.

Worksheet No. 24 and the two formulas for the arithmetic mean,

$$\bar{X} = A + \frac{\Sigma fx'}{N} i, \text{ and the standard deviation,}$$

$$\sigma = i\sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2} \quad \text{are}$$

among the most useful and widely used in all statistics. With a minimum of numbers and computations and with these of small size even for large samples, both $\bar{X}$ and σ may be quickly computed. The student should master them fully.

The relative width of the four measures of dispersion are clearly shown in Fig. 43 on page 162. For the distribution of wages for the Oil Refinery employees, they are:

Range		$110.00
Quartile Deviation	± $14.52	$29.04
Mean Deviation	± $16.76	$33.52
Standard Deviation	± $20.60	$41.20

Of course these particular figures hold for only this sample of wage of the Oil Refinery employees, but the relative size of the measures remains about the same for any fairly normal frequency distribution. The ratio of the standard deviation to the range for various sizes of samples is approximately as follows:

Size of Sample	$\frac{Range}{\sigma}$
30	4.0
100	5.0
200	5.5
500	6.0
1000 and over	6.5

The reason for this variation is that in small samples there is little likelihood of getting any of the extremely small or extremely large items of the population. The small sample is almost certain to be composed of items from the dense middle portion of the universe. This limitation would give it a relatively short range. In

a large sample, however, some of the more extreme items would likely be included. This wider distribution would lengthen the range.

In a perfectly normal distribution $Q_d = 0.6745\sigma$, and $M_d = 0.7979\sigma$. These ratios will vary slightly as the frequency curve departs from normal.

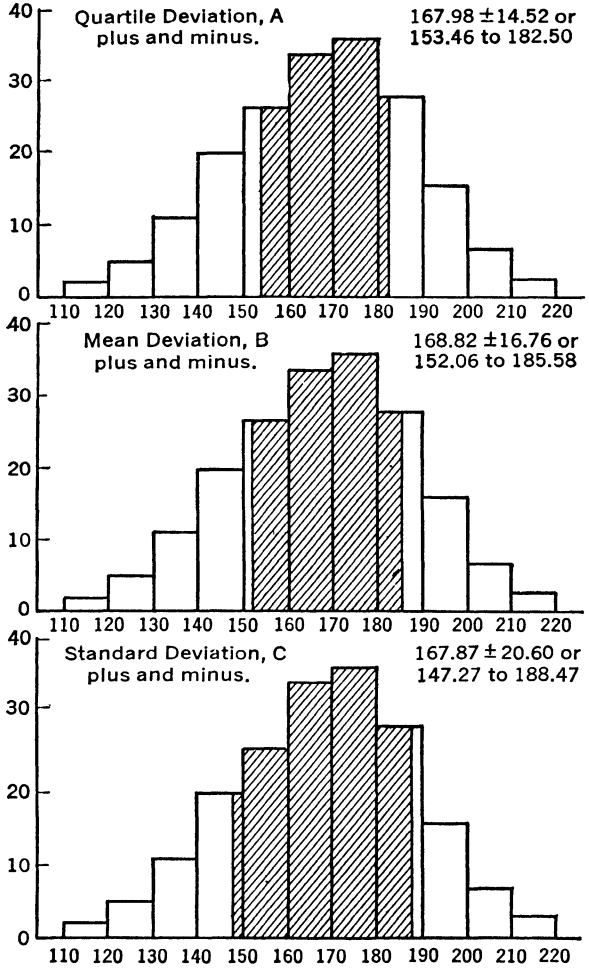


FIG. 43. Relative sizes of (A) Quartile Deviation, (B) Mean Deviation, and (C) Standard Deviation for semi-monthly wages of 188 oil refinery workers

Like averages, all measures of dispersion are expressed in terms of the original data; i.e., in inches, pounds, tons, dollars, bales.

QUADRATIC MEAN

The quadratic mean often called the root-mean-square in some respects resembles a standard deviation in methods of computation but is only an average or mean. The standard deviation is always based on deviations from the mean. A quadratic mean may be computed for any type of measurement. It is the square root of the arithmetic mean of the squares of the values. Since, when an item is squared, it is weighted by itself, the quadratic mean gives the greater weight to the larger items. It is, therefore, larger than the simple arithmetic mean for the same data.

It is not widely used, but it may be employed to measure the scatter of items about a fixed point such as rifle shots around a "bull's eye." How small must the quadratic mean scatter of the shots be before one is rated a sharpshooter? The data shown in Worksheet No. 25 are very good illustrations of this measure.

WORKSHEET NO. 25

QUADRATIC MEAN OF RANDOM SAMPLE OF 100 RIFLE SHOTS AT 50 FEET TAKEN FROM PRACTICE TARGETS OF OKLAHOMA A. & M. COLLEGE R.O.T.C. BASIC MILITARY STUDENTS, FEBRUARY 12, 1942

Class Intervals in Millimeters	Mid-point of Class	Frequencies	Deviations from 0 in Class Units	Frequencies Times Deviation	Frequencies Times Squared Deviation from Zero
i	m	f	x	fx	fx^2
0– 3.9	2	38	.5	19.0	9.50
4– 7.9	6	23	1.5	34.5	51.75
8–11.9	10	14	2.5	35.0	87.50
12–15.9	14	10	3.5	35.0	122.50
16–19.9	18	7	4.5	31.5	141.75
20–23.9	22	4	5.5	22.0	121.00
24–27.9	26	3	5.6	19.5	126.75
28–31.9	30	1	7.5	7.5	56.25
		100			717.00

FORMULA NO. 29

$$Mq = i\sqrt{\frac{\Sigma fx^2}{N}} = 4\sqrt{\frac{717}{100}} = 4\sqrt{7.17} = 4 \times 2.68$$
$$= 10.72 \text{ millimeters.}$$

Two characteristics of the data have now been measured, its central point of concentration, or average, and its tendency to scatter about that average, or its dispersion. Two other measures are necessary to complete the analysis and description of a frequency distribution.

SKEWNESS

Skewness is the tendency of a distribution to depart from a symmetrical balance of its two sides, or to be lopsided. Perfectly normal distributions are rare except in perfectly random and fair games of chance and in the pure physical sciences. In the field of biology and related sciences, distributions tend to approach normal, but in economics, sociology, political science, and wherever human activities and institutions are concerned, distributions tend to be skewed.

Meaning of symbols:

S_k = degree of skewness

FORMULA NO. 30

$$S_k = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

This theory is based on the fact that in a perfectly smooth or balanced distribution the numerator would be zero, or no skewness. The relationship in a perfectly normal distribution is that the distance of either quartile from the median is the same. That is:

$$Q_3 - Me = Me - Q_1$$

transposing

$$Q_3 + Q_1 - Me - Me = 0$$
$$Q_3 + Q_1 - 2Me = 0$$

The denominator is the inter-quartile range. For the data on wages of the Oil Refinery employees the computation is:

$$S_k = \frac{182.50 + 153.46 - 2 \times 168.82}{182.50 - 153.46}$$

$$= \frac{1.68}{29.04} = .058$$

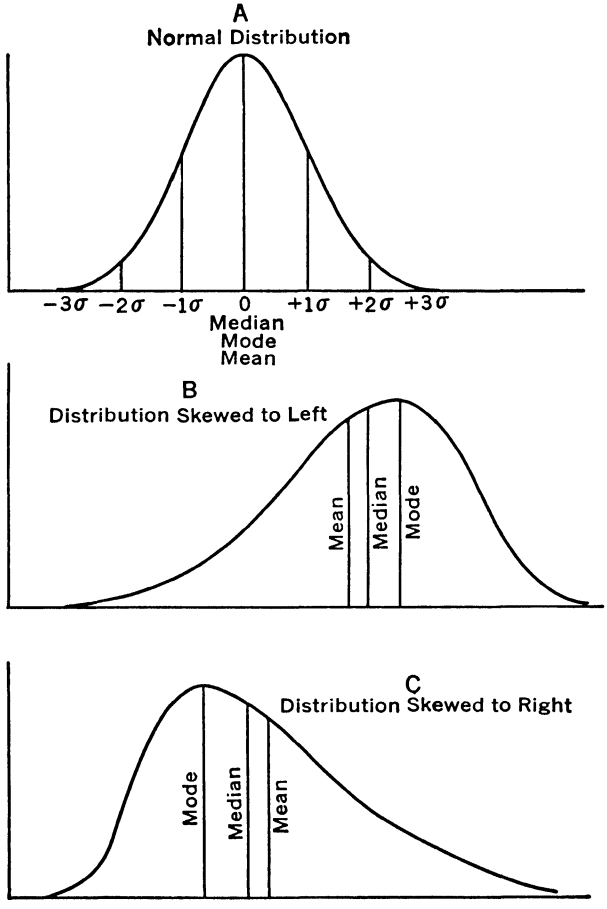


FIG. 44. Forms of normal and skewed distributions showing locations and relative shifts of (1) Arithmetic means, (2) Medians, and (3) Modes

With this formula a figure of .20 or less is a small or moderate amount of skewness. A figure of .40 or more is a large degree of skewness.

FORMULA NO. 31

$$S_k = \frac{3(\bar{X} - Me)}{\sigma}$$

This formula is based on the fact that when a distribution departs from normal, the mean, median, and mode are pulled apart in the proportion of 1 to 3; that is, the distance between the mean and median is one-third the distance between the mean and mode. The mean always remains nearest the tail or skew of the distribution. The mode remains under the highest frequency and the median still divides the area of the distribution into two equal parts. The standard deviation is used as the base of variation.

For the data on wages of the Oil Refinery employees, the computations are:

$$S_k = \frac{3(167.87 - 168.82)}{20.60}$$

$$= \frac{3 \times .95}{20.60} = \frac{2.85}{20.60} = .138$$

In comparing a group of frequency distributions the same formula for skewness, whichever it is, should be employed throughout the comparison because the several formulas give somewhat different results. To change from one formula to another as among several distributions will be confusing, although the formulas may be equally good in themselves.

THE MOMENTS OF A FREQUENCY DISTRIBUTION

A third measure of skewness is that based on the third moment of the deviations of a frequency distribution about its arithmetic mean. This method is much more complex than those given above, but it is so closely related to the arithmetic mean, standard deviation and other statistical measures that its fuller development is justified.

The word "moment" is a term used in mechanics which includes the combined ideas of "force" and "distance," as a force applied to a lever a distance from a fulcrum. Its use in statistics carries the same implication in which the force or weight is the frequency of the class and the distance is the deviation of that class from the arithmetic mean, or zero (0) origin.

Meaning of symbols:

π (pi) = a moment of distribution about the true mean

ν (nu) = a moment of distribution about an assumed or arbitrary mean.

MOMENTS AROUND AN ARITHMETIC MEAN

First moment about the mean = $\pi_1 = 0$

Second moment about the mean = $\pi_2 = \nu_2 - \nu_1^2$

Third moment about the mean = $\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$

Fourth moment about the mean = $\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$

MOMENTS AROUND AN ARBITRARY MEAN

$$\text{First moment} = \nu_1 = \frac{\Sigma fx'}{N}$$

$$\text{Second moment} = \nu_2 = \frac{\Sigma f(x')^2}{N}$$

$$\text{Third moment} = \nu_3 = \frac{\Sigma f(x')^3}{N}$$

$$\text{Fourth moment} = \nu_4 = \frac{\Sigma f(x')^4}{N}$$

The student will readily recognize the first moment, $\nu_1 = \frac{\Sigma fx'}{N}$, as the correction we have been using in the computation of the arithmetic mean from an assumed mean, Worksheet No. 10 and Formula No. 6.

$$\bar{X} = A + \left(\frac{\Sigma fx'}{N}\right)i$$

The second moment is equally evident in Worksheet No. 24 and Formula No. 28, $\sigma = i\sqrt{\frac{\Sigma f(x')^2}{N} - \left(\frac{\Sigma f(x')}{N}\right)^2}$, is used in computing the standard deviation from an assumed origin.

MEASURE OF SKEWNESS BY THE THIRD MOMENT

The third moment is used as a measure of skewness and the fourth moment will be presented near the end of this chapter as a measure of kurtosis. The development of these four moments is, therefore, a device for clarifying more fully the development of the four measures of a frequency distribution and the interrelations among these four measures, the mean, the standard deviation, skewness and kurtosis.

It is evident that for a normal curve such as Fig. 40A the sum of the x column of deviations computed from a true mean would be zero (0), the minus values exactly canceling the plus values. In such a normal distribution the sum of the third moment, Σx^3, would also be zero, the minuses canceling the pluses. But if the distribution is not symmetrical, but is skewed as Figs. 40B and C the pluses and minuses will not cancel and the residual value may be used to measure the degree of skewness.

WORKSHEET NO. 26

Normal

X		$\bar{X}$		x	x^2	x^3
1	−	3	=	− 2	4	− 8
2	−	3	=	− 1	1	− 1
3	−	3	=	0	0	0
4	−	3	=	+ 1	1	+ 1
5	−	3	=	+ 2	4	+ 8
15				0	10	0

First and third moments = 0.

$$\pi_1 = \frac{\Sigma x}{N} = \frac{0}{5} = 0$$

$$\pi_3 = \frac{\Sigma x^3}{N} = \frac{0}{5} = 0$$

Skewed

X		$\bar{X}$		x	x^2	x^3
1	−	4	=	− 3	9	− 27
2	−	4	=	− 2	4	− 8
3	−	4	=	− 1	1	− 1
5	−	4	=	+ 1	1	+ 1
9	−	4	=	+ 5	25	+ 125
20				0	40	+ 90

First moment = 0
Third moment = + 90

$$\pi_1 = \frac{\Sigma x}{N} = \frac{0}{5} = 0$$

$$\pi_3 = \frac{\Sigma x^3}{5} = \frac{90}{5} = 18$$

If the distribution is skewed at all the sum of the cubed deviations will not equal zero (0). The size of the sum and its algebraic sign indicate the amount and direction of the skewness. A number of less than 5 indicates moderate or small skewness.

WORKSHEET NO. 27

COMPUTATION OF FIRST FOUR MOMENTS FOR SEMI-MONTHLY WAGES OF 188 OIL REFINERY EMPLOYEES, MARCH, 1951

Mid-points i	Frequencies f	Deviations from Assumed Mean x'	fx'	$f(x')^2$	$f(x')^3$	$f(x')^4$
115	2	− 5	− 10	50	− 250	1250
125	5	− 4	− 20	80	− 320	1280
135	11	− 3	− 33	99	− 297	891
145	20	− 2	− 40	80	− 160	320
155	26	− 1	− 26	26	− 26	26
165	34	0				
175	36	+ 1	+ 36	36	+ 36	36
185	28	+ 2	+ 56	112	+ 224	448
195	16	+ 3	+ 48	144	+ 432	1296
205	7	+ 4	+ 28	112	+ 448	1792
215	3	+ 5	+ 15	75	+ 375	1875
	188		+ 54	+ 814	+ 462	+ 9214

COMPUTATION OF SKEWNESS BY THIRD MOMENT

$$\nu_1 = \frac{\Sigma f(x')}{N} = \frac{+54}{188} = +.2872$$

$$\nu_2 = \frac{\Sigma f(x')^2}{N} = \frac{+814}{188} = +4.3298$$

$$\nu_3 = \Sigma f(x')^3 = \frac{+462}{188} = +2.4574$$

$$\nu_4 = \Sigma f(x')^4 = \frac{+9214}{188} = +49.0106$$

$$\pi_1 = 0$$

$$\pi_2 = \nu_2 - \nu_1^2 = 4.3298 - (+.2872)^2 = 4.3298 + .0825 = 4.2473$$

FORMULA NO. 32

$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$$
$$= 2.4574 - 3(+.2872 \times 4.3298) + 2 \times (+.2872)^3$$
$$= +2.4574 - 3.7305 + .0474 = -1.2257,$$ the third moment measure of skewness.

The reason that measures of skewness are important in statistical studies is that the theory, formulas and inferences of frequency distribution analysis are based on the normal curve of distribution and when a sample departs too far from normal, the theory, analysis and conclusions must be modified to correct the errors arising from abnormality. For moderate or small skewness no compensations are required.

KURTOSIS

Some frequency distributions depart from normal by either rising to a tall slender peak as indicated in Fig. 45 or by being flat or truncated as indicated in Fig. 46.

A distribution like that in Fig. 45 is said to be *leptokurtic*, or peaked. One like that in Fig. 46 is said to be *platykurtic*, or flat-topped. If either abnormality runs to an extreme it invalidates to some extent the theoretical inferences drawn from the normal curve.

The degree of kurtosis may be measured by the fourth moment of the distribution about its mean.

FORMULA NO. 33

$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$$

The computations for the wages of the Oil Refinery employees in Worksheet No. 27 are as follows:

$$\pi_4 = 49.0106 - 4(+.2872 \times +2.4574) + 6(+.2872)^2(4.3298) - 3(-.2872)^4$$
$$= 49.0106 - 2.8231 + 2.1432 - .0204$$
$$= 49.3103$$

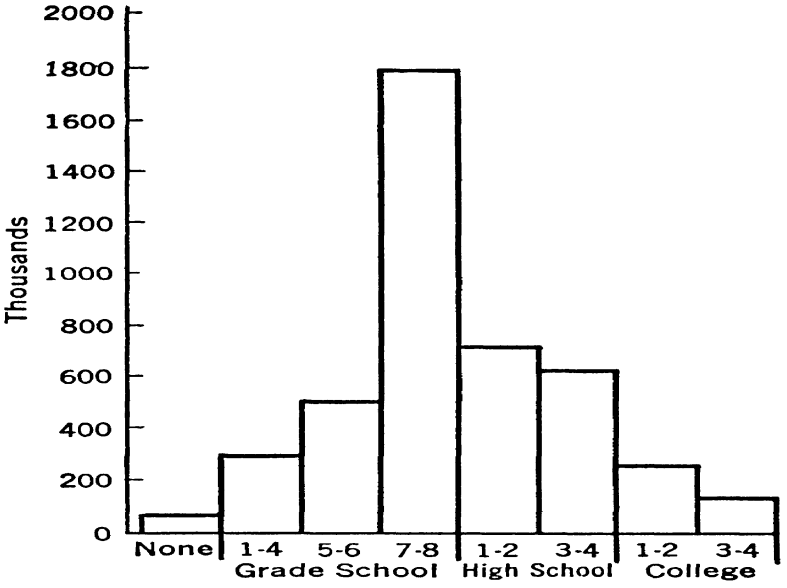


FIG. 45. Leptokurtic, or peaked distribution. Years of schooling completed by married women in the U.S. Sample C. (*Statistical Abstract of U.S.*, 1949, p. 27)

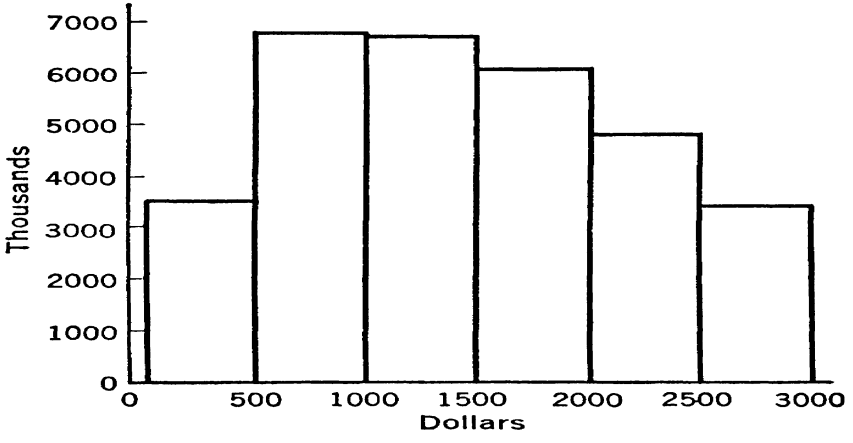


FIG. 46. Platykurtic, or flat-topped distribution. Distribution of families and single consumers by money income level of $3,000 and below, in U.S., 1942. (*Statistical Abstract of U.S.*, 1949, p. 289)

This result is an absolute measure of kurtosis and is too large a number to be easily manageable in comparative studies. A more useful measure is the alpha (α) ratio.

FORMULA NO. 34

$$\alpha_4 = \frac{\pi_4}{\pi_2{}^2} = \frac{49.3103}{18.0396} = 2.733$$

In the normal curve the α_4 ratio = 3. For a flat-topped distribution it is $\alpha_4 < 3$. For a peaked distribution it is $\alpha_4 > 3$.

COEFFICIENT OF VARIATION

It is frequently useful to compare the relative dispersion of two or more frequency distributions. Since the standard deviation of any distribution is the best measure of its variation, σ becomes the logical numerator for measures of variation among a series of distributions.

FORMULA NO. 35

$$V = \frac{\sigma}{\bar{X}} 100$$

For the three frequency distributions on Tulsa grade school children the computations are:

	$\bar{X}$	σ	V
Height (Inches)	54.6	4.8	8.8%
Age (Months)	113.8	22.6	19.8%
Weight (Pounds)	76.8	18.5	24.1%

These children vary twice as much in age as they do in height and three times as much in weight as in height.

For the wages of employees in the Oil Refinery the computations are:

$$V = \frac{\sigma}{\bar{X}} 100 = \frac{20.60}{167.87} 100 = 12.3\%$$

This measure is useful for comparing departments, stores, plants, and other factors in business.

SUMMARY

1. The *range* includes all the data from the smallest to the largest item. It tends to be smaller in small samples (25 to 30 items) than in large samples (100 or more items), but in any case it is not a good measure of dispersion, because on its extremities it covers a long distance of very sparse data. It is too large to be an accurate measure of dispersion. It is also subject to a large variation from chance sampling. It is a locational value.

2. The *quartile range* includes the middle one-half of the sample items. The chance of an item being inside or outside is 50–50. It is too small to be a good measure of dispersion. It leaves out too much of the densely populated area of the normal curve. It is a locational value.

3. The *average deviation* measured plus and minus is a wider measure of dispersion than the quartile range, but it has the disadvantage of being algebraically unsound in that its computation requires the disregarding of all plus and minus signs. It may be used as a basis of correlation but because of other limitations it is not widely used. It is a computed value.

4. The *standard deviation* is the square root of the sum of the squared deviations from the mean divided by the number of items. It is algebraically sound and is the most stable and widely used measure of dispersion. It is a computed measure. It is used as the basis for correlation and most computations of statistical error. The student should master it thoroughly.

5. *Skewness* is the measure of the lopsidedness of a frequency distribution. They may be skewed either to the right or to the left. The skew is in the direction of the long narrow end or tail of the distribution. Distributions in the physical and biological universe tend to be more nearly normal than among social data which are usually skewed to the right. If the skewness is quite large it invalidates the use of the arithmetic mean and standard deviations as accurate measures of the distribution. In such cases the median or the mode and the quartile range are more representative of the large body of the data.

6. *Kurtosis* is the departure of a frequency distribution from normality by its being either peaked or flat at the top. If the departure from normal is extreme it tends to make the standard deviation abnormal and to reduce its value for further computations. It may be measured by the fourth moment about the mean.

7. The *coefficient of variation* measures the relative scatter of the items of a sample about their mean as compared with other distributions. Some distributions are closely confined about their mean; others are widely scattered. The narrower the scatter of the items about the mean, the more representative of the items is the mean.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics,* Chapter XVIII. Irwin, Inc., Chicago, 1941.

Chaddock, R. E., *Principles and Methods of Statistics,* Chapter IX. Houghton Mifflin Company, Boston, 1925.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics,* Chapter XIII. Prentice-Hall, Inc., New York, 1948.

Davies, G. R., and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences,* Chapter IV. John Wiley and Sons, New York, 1933.

Holzinger, K. J., *Statistical Methods for Students in Education,* Chapter VII. Ginn and Company, Boston, 1928.

Mills, F. C., *Statistical Methods, Revised,* Chapter V. Henry Holt and Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics,* Chapter III. John Wiley and Sons, New York, 1942.

Richardson, C. H., *An Introduction to Statistical Analysis,* Chapter IV. Harcourt, Brace and Company, New York, 1935.

REVIEW QUESTIONS

1. Define dispersion. Why is it necessary to measure dispersion in order to make complete comparisons of frequency distributions?
2. What is the range? What are its advantages and disadvantages as a measure of dispersion?
3. What is the difference between the quartile range and quartile deviation? How much and what portion of the distribution do they measure?
4. Compare the length of the quartile range with the total range.
5. What are the advantages and disadvantages of the mean deviation?
6. Explain the difference in the methods of computing the mean deviation and the quartile deviation.
7. What is the difference between the mean deviation and the standard deviation?
8. Explain the difference between the long method and the short method of computing the standard deviation.
9. Compare the standard deviation with the other measures of dispersion as to ease of computation, adequacy of measurement of dispersion, and algebraic manipulation.
10. What is skewness? In what ways may it be measured? Of what practical use is its measurement?
11. What is the coefficient of variation and for what is it used?
12. In what units is the standard deviation always measured?

13. If a distribution is highly skewed, what average is often better than the mean?

14. What does "skewed to the left" and "skewed to the right" mean?

15. What is kurtosis? How may it be measured?

16. What is a "moment" as the term is used in statistics? Explain in detail.

EXERCISES

1. Use the exercises at the end of Chapter 6.

CHAPTER 8

LINEAR REGRESSION

Regression involves a comparison between *two* or *more* variables. In all the statistics computed up to this point one variable only has been measured at a time. Means have been computed for height of school children, cost of housing units, wheat yields, train speeds, price relatives and other data, but in no case, as yet, have there been any comparisons made between *two* series. Each one has stood absolutely alone. Standard deviations have been computed for several variables, but no comparisons have been made between any of these series of data. In fact, the computation of averages and dispersions is limited to the measurement of single variables. Fundamental and important as are all of those methods, they do not extend to the measurement of the relations among two or more series of related data. One of the more common methods of statistical analysis used to measure relations between two variables is called *regression.*

Immediately after one has computed the mean and standard deviations of the weight and height of children, other questions arise. Is there any relationship between the height of a child and his weight? Is a tall child likely to weigh more than a short one? How accurately can one estimate the weight of a child if his height is known? How accurate, or dependable, would such an estimate be? How much error on the average is there in such estimates? Is there any scientific basis for these age-height-weight charts by which mothers measure the growth of their babies from birth to school age and which teachers use to determine whether school children are *normal* or are *underweight* or *overweight?* When is a child underweight or overweight? What

percentage of his weight may be accounted for by his height? What percentage of children's weights may be explained by their heights, and what part of their weights may have to be accounted for by age, or heredity, or family characteristics, or by proper and sufficient food, or undernourishment? These and many other questions must be answered before our analysis of these data is complete or even sufficient for ordinary purposes.

Many questions arise concerning the yields of wheat and farm income. What is the relationship between the two? Is a yield of one bushel or five bushels as likely to be associated with high income as would twenty bushels? How much does it cost to produce a bushel of wheat? Does the cost per bushel decline as the yield increases? What is the regression between these variables?

What is the relationship between demand for a good and its price? or between its supply and price? How many more potatoes would be sold at $.50 a bushel than at $1.00 a bushel? What is the effect on consumption of doubling the tax on cigarettes? These and thousands of other variables must be compared to answer the problems of management and science.

Regression measures are the means of answering many of these questions. *Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.* It is generally recognized as the most important single field of statistical analysis. It is the basis of correlation analysis, co-variance analysis, and of many important measures of error. It is the stuff that science is made of. It furnishes an accurate basis for the comparison of variables, of series of data, of significant relationships. It ferrets out and expresses in clear, measurable terms the hidden and inner meanings of great masses of numbers which otherwise would remain unknown. Comparisons are the basis of all organized knowledge and of all human institutions. Every day of our lives, we are forced to make comparisons and far-reaching decisions based on comparisons. Shall we go to town or to school if it rains? What will be the effect on my grades if I reduce my study time for social pleasures or athletic exercises or student activities? If I take eighteen hours instead of fifteen,

what will be the results on grade points? If I reduce my hours of sleep from eight to five, what will be the effect on my efficiency? What are the chances that marriages of persons under twenty years of age will be as satisfactory and enduring as those made after twenty-five? What is the relationship between the prices the farmer receives and the prices the consumer pays? What is the relationship between the speed of driving and the time and distance required to stop the car? In the hurry of daily life we answer these and a thousand other vital questions as best we can and go on with the job. The methods of regression in statistics give us more exact and dependable answers.

Science may be defined as a *statement of observed uniformities and the relationships among them.* Regression is the basic tool of analysis in all science. Chemistry, physics, astronomy, as well as psychology, economics, sociology, and political science are no more than statements of uniformities which have been observed concerning phenomena and the regressions which have been measured among them. In a more popular sense we may call regression a trend, a line which shows how many units of change in one variable are associated with one unit of change in another variable. If one boy is six inches taller than another, how many more pounds should he weigh? If he is two years older than the other boy, how tall should he be? Is there any reliable method by which weight can be estimated from height? Is it possible to estimate the incomes of farms from the yields of wheat or cotton or hay per acre? Such possibilities are always present whenever two variables are in close association. For instance, the weight of children might increase 3.5 pounds for each increase of 1.0 inch in height; or the income of wheat farms might increase 1.0% for each increase of 3.0 bushels in yield, or the price of potatoes in the general market might decline 1.0 cent for each increase of 10,000,000 bushels in supply.

INDEPENDENT AND DEPENDENT VARIABLES

When comparisons are made between two variables, one of them is called the *independent* variable and the other is designated the

dependent variable. The independent series is the one which varies or is thought of as varying independently as height, yields, supply, heat, or speed. The dependent series is the one which changes or is thought of as changing as a result of change in the independent series, such as weight, income, price, expansion, or distance traveled when associated with the variables listed in the previous sentence. It is *not* at all necessary that there be a *causal relationship* between the independent and dependent variables. All that is required is that the amount of change in the dependent be considered in relation to a unit of change in the independent variable. The dependent series is always measured on the vertical or Y-axis and the independent series on the horizontal or X-axis.

REGRESSION BASED ON AVERAGES

In measuring regression or trend relationship between two variables *the result is based on the average of all the pairs of data in the sample.* Regression cannot be computed on a single pair of items any more than a mean or standard deviation can be computed for one datum. As a mean is an average of one series of data so a regression line is an expression of the average relationship between the paired items of two related series of data. There can be no regression between the height and weight of one child. When the sample is increased to include two or more pairs of items the measure of their average relationship is regression.

METHODS OF COMPUTING REGRESSION LINES

In this chapter three methods of computing regression lines will be explained, (1) Freehand lines, (2) Lines through class averages, and (3) The method of least squares. The first two methods require little or no mathematics and are used largely for purely descriptive purposes or for preliminary analysis. For more complete and adequate analysis and inference the more mathematical least squares method is employed.

Freehand Curves

The first step in any case in fitting a regression line is to plot the data on coordinate graph paper, or on any paper on which

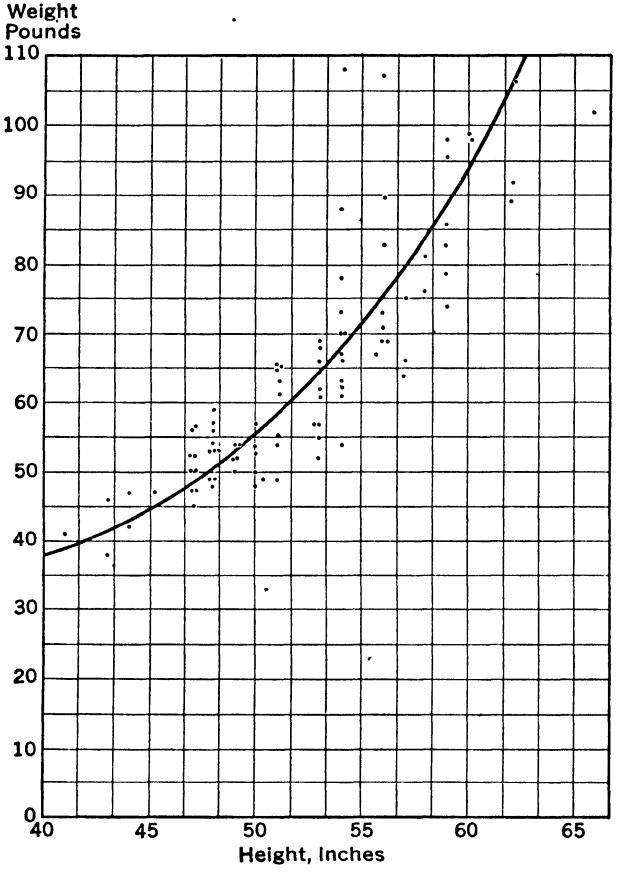


FIG. 47. Freehand curve showing regression between heights and weights of 106 Stillwater grade-school children. (Data From Ex. 5, Chap. 8)

one can set up the X-axis and Y-axis and measure abscissas and ordinates as indicated in Figs. 47, 48 and 49, which were plotted from the paired values of height and weight, age and height, and age and weight from the data in Worksheet No. 1.

Figure 47 shows a freehand curve which expresses the average

relationship between the height on X and the weight on Y of the 106 school children. This line is a smooth curve which is relatively horizontal for the smaller values of X but rises more rapidly for the larger X values. This indicates that increase in weight is

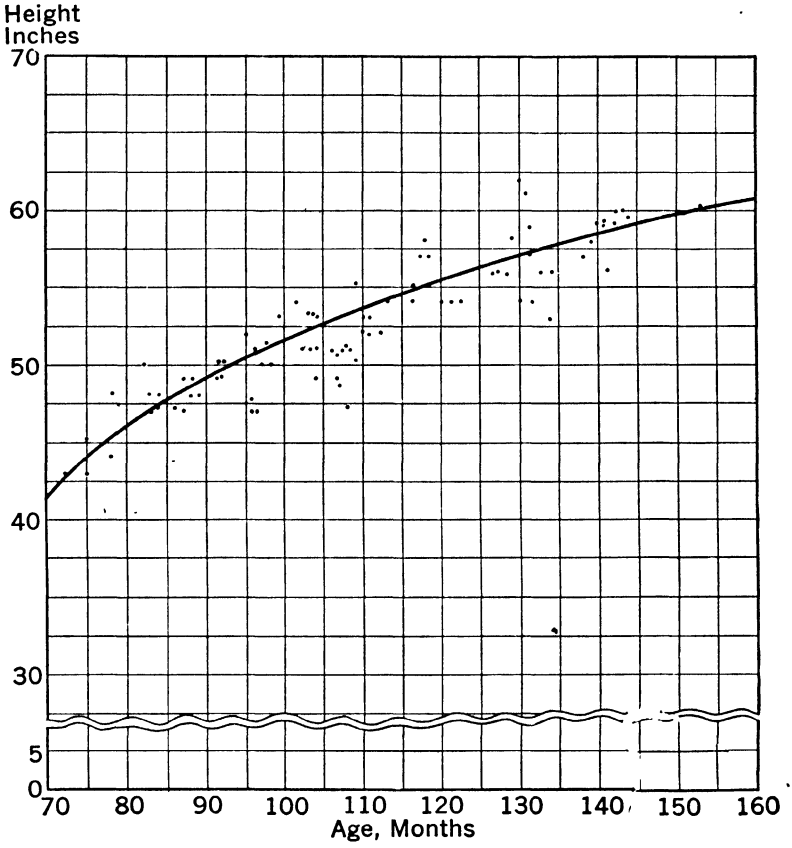


FIG. 48. Freehand curve showing regression between heights and ages of 106 Stillwater grade-school children. (Data Ex. 5, Chap. 8)

not a strictly proportional change for increase in height, but that as children grow taller they tend to increase in weight at a more rapid rate than they do in height. This would be expected since weight is a three-dimensional product while height is only one.

Figure 48 shows the relationship between age and height for

the same children. This curve is drawn on the graph freehand. Although it is almost straight, it rises more steeply for the earlier ages and more slowly as the children grow older. This shows that these children on the average increased in height more slowly as they approached maturity.

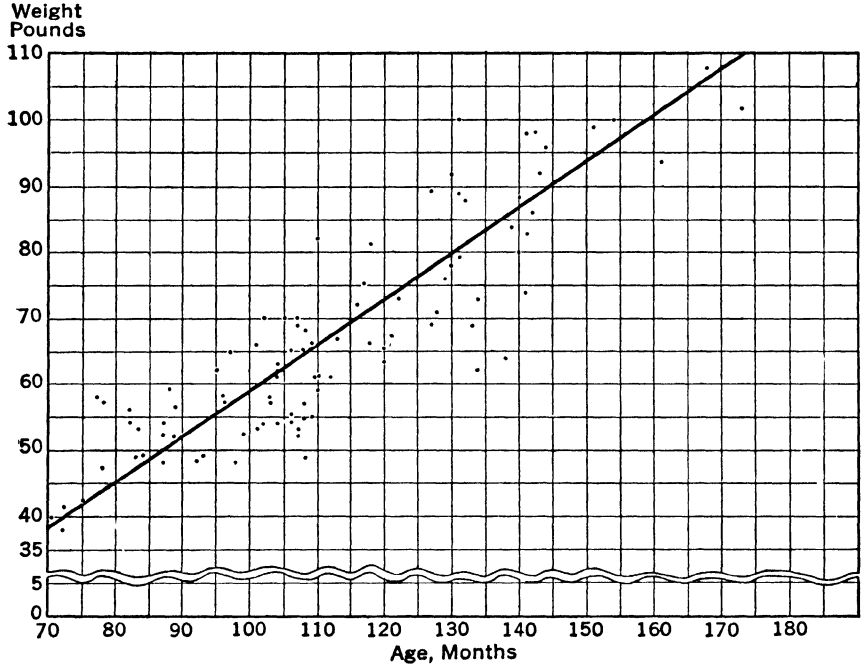


FIG. 49. Freehand curve showing regression between weights and ages of 106 Stillwater grade-school children. (Data from Ex. 5, Chap. 8)

Figure 49 shows the relationship between age and weight. This line was drawn in freehand. It is a perfectly straight line. In all three graphs there is no prejudgment or assumption of any particular relationship between the variables. The data are simply plotted on an accurate scale and a freehand line drawn in which as nearly as possible describes their general trend or movement. Such a method is purely descriptive, but it is quite useful in indicating the general type of mathematical line that is to be used if one is to be fitted. This is important because all mathematical lines rest on pre-judgments or assumptions of relation-

ship which may be either correct or incorrect, depending on the amount of information available before the decision is made. If properly done the freehand curve is an aid in making the decision as to what type of mathematical line to use.

Technique for Drawing Freehand Curves

1. Plot the data on coordinate graph paper with one variable measured on the X-axis and the other variable on the Y-axis.
2. Hold a tight string or a slender rubber band over the plotted points so as to split the points above and below the string or band into two equal regions or sections.
3. If a straight line does not correctly represent the relationship, shift the position of the string, turning it along the ends of the plotted data until you form a clear idea of the approximate degree of curvature to give the line.
4. First draw in a light line with pencil so that it splits the data in the middle from one end of the X-axis to the other. By gentle erasures and re-drawing, adjust the line until it is the best you can make. Then draw in a heavier line with ink.

The freehand curve has the following advantages:

1. It is easily and quickly obtained.
2. For purely descriptive purposes it is sufficiently accurate.

Its disadvantages are:

1. It is not represented by any algebraic equation.
2. Since it is freehand, the same person at another time, or another person, would not exactly reproduce the same line.

In spite of these limitations, the freehand curve is very useful in preliminary analysis. It is an easy means of determining the type of mathematical regression that should be worked out if a more complete analysis is attempted later. With a little practice many persons become exceedingly adept in drawing freehand curves.

Lines through Class Averages

By this method the data are divided into two approximately equal parts. The division does not have to be absolutely equal,

but a division near the median of the X-variable is preferable. The steps to be followed are:

1. Arrange the items of the independent or X-variable in an array in the left-hand column with the associated Y-value in an adjoining column. *The paired X- and Y-values cannot be separated.* This requirement will leave the Y-values in an irregular order.
2. Divide the columns somewhere near the median.
3. Compute simple means for each column of X-values and their associated Y-values.
4. Plot the original data on coordinate graph paper.
5. Locate the two sets of class averages on the graph of plotted data and draw a straight line through these two points with a rule.

This method will frequently give a quite accurate or representative regression or trend line as is indicated in Worksheet No. 28 and Fig. 50.

WORKSHEET NO. 28

COMPUTATION OF REGRESSION LINE THROUGH TWO CLASS AVERAGES FOR HEIGHTS AND WEIGHTS OF 106 CHILDREN

	X, 41–52		X, 53–66	
	Height X	Weight Y	Height X	Weight Y
	41	41	53	61
	43	38	53	57
	43	46	53	61
	44	42	53	82
	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .
	52	61	66	102
Sums	2523	2774	3035	4122
Items	52	52	54	54
Means	48.52	53.35	56.20	76.33

Using only two class interval averages, the class average regression line would be straight and almost identical with the least square line.

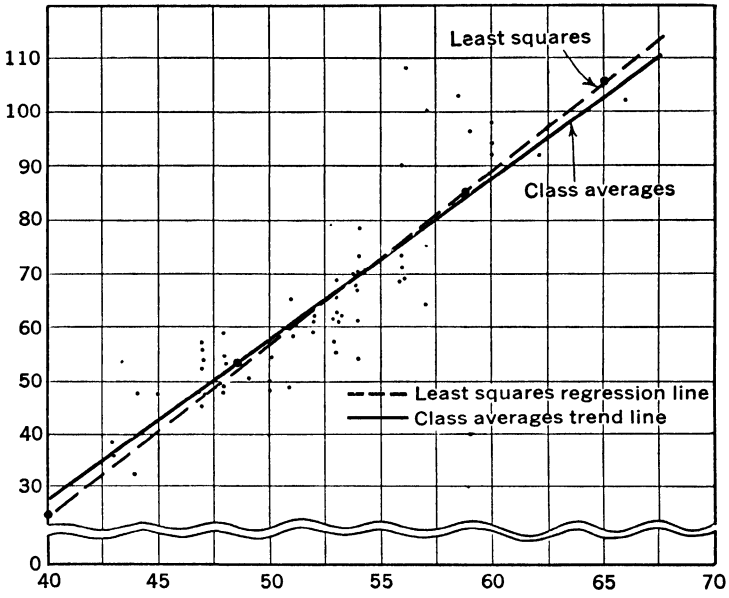


FIG. 50. (A) Regression line through class averages, (B) Least squares regression line for heights and weights of 106 school children, computed from Worksheet 30

Use of Class Average Lines

Many of the measurements required to clarify the basic relationships used in the management of business may be had by class average lines. In a department store even of large size the comparative growth trends of the several departments may be easily computed by this method. The relationship of costs to production, to price and to each other may be measured by class average lines. Advertising expenditures, salesmen's commissions, radio and television costs and other selling expenses may be related to sales by this device. In personnel management the qualifications of employees, such as. education, on the job training, experience, efficiency, productivity, or any other qualification which can be expressed in numbers may be compared by class average lines. Class average trends are so inexpensive to compute that they are especially useful as first approximations.

Interpretation of Freehand and Class Average Regression Lines

A freehand curve is largely a descriptive device. Just as the plotted points of the data on the coordinate graph picture to some extent the logical average relationships between the two variables so does the freehand curve through the center of those points further describe this relationship in a more concrete summary form. The only possible inferences which may logically be drawn from a freehand curve are that (1) to the extent to which the sample truly represents the population does the freehand curve represent the true regression in the population, provided (2) the freehand curve does correctly measure the actual regression in the sample. Error in either case will to the extent that it is present invalidate any conclusions drawn from the freehand curve. (1) If the sample does not adequately represent the population, the sample freehand curve cannot adequately represent the regression in the population. (2) If the freehand curve does not correctly measure the regression in the sample, it cannot correctly measure the regression in the population even though the sample is adequate. One or both of these errors will always be present. It is the purpose of the statistician to keep them as small as possible. The advantage of the freehand curve is that it is an inexpensive descriptive device which can often be made to give a good picture of the trend in the data.

Assumptions under Mathematical Curves

Most mathematical curves used in statistics are based on assumptions or pre-judgments as to their appropriateness to measure the relationship under consideration. It is possible to apply a dozen or even a hundred different mathematical curves to one sample of data. Since these mathematical curves are quite diverse in shape and functions, it is logical that if one of them does fit the data well the others cannot. Of course, none of the ones tried may fit. Outside of the fields of astronomy, atomic chemistry, quantum physics and electronics, data are usually much too complex to coincide with purely theoretical mathematical functions.

In the fields of biology, agriculture, education, economics, sociology, and political science no such basic theoretical mathematical functions have yet been discovered or developed. Whenever life and protoplasm form part of the data, and especially whenever human intellect, judgment and prejudices become part of the data, and still more especially whenever human institutions with all their social lag and traditions enter the data, the subject matter for measurement and analysis becomes so very complex that a workable, purely theoretical, mathematical attack on the problem becomes largely or entirely impractical and useless.

Under such circumstances the statistician enters an experimental and empirical field in which he often has to resort to methods of experimentation which in their initial stages may decline to the level of "cut and fit" or "successive approximations."

An elementary statistics text is not a proper place to carry any conflict between purely theoretical mathematics and the exigencies of empirical research to a final solution. The most that is desired is to point out clearly that in most statistical work, many mathematical regression lines could be applied to the same data and that perhaps few if any of them would fit well, and that it is often necessary to try several before a satisfactory fit is found. This does not mean that in many cases a basic functional relationship may not be discovered which a specific mathematical function measures closely or even perfectly, but it does mean that in most cases, in the social sciences at least, such functions are not known and that the likelihood of their early discovery declines with the increasing complexity of the data. The elementary student in social and even in biological statistics must, therefore, rely largely on empirical experimental methods in deciding what kind of regression line to employ for any given data. For the simpler economic changes such as variations of business activity with the time of year or of growth over a period of time, basic mathematical functions are more easily discovered than for the more complex business cycle or social evolution. Mathematical economists have made some progress in discovering and measuring the more simple economic functions. Further progress in this field is to be expected.

Assumptions under Class Average Lines

When a statistician decides to divide his data into only two classes he automatically pre-determines a straight trend line, because the two averages limit the result to a straight line. The postulate, "Two points determine a straight line," is well known. If, however, the data are divided into three or more classes with paired averages for each of the several classes, the regression line may be a smooth curve among the averages or a series of straight lines from one class mean to the next. In deciding upon more than two classes the statistician does not pre-determine a curved line of any particular function, but merely makes possible the revealing of the actual conditions and relations existing in the data. If the relationship is actually a straight line, the use of more than two classes will give the straight regression. If, however, the actual relationship is not straight, the use of more than two classes will give a more or less true picture of what the true relationship is. The use of two class averages presumes a straight-line regression. The use of more than two classes does not presume any particular relationship but merely indicates what does exist. This point is treated more fully in the chapter on curvilinear regression.

Since it is often quite difficult to determine beforehand what kind or type of mathematical line to fit to the data, it is usually best either to fit a freehand curve or the curve through class averages first, and by means of this curve decide the type of mathematical curve required. Such a procedure will make unnecessary the difficult task of experimenting with several expensive mathematical curves before the correct one is discovered and will secure a better fit. If all that is desired is to describe one set of data without any thought of setting up a generalization for all similar data, that is, if it is a special case, the class average curve is as good or better than any other for describing it. If, however, it is thought that the relationship existing in any sample is representative of a large group of data and is a permanent and universal principle having a wide application, the mathematical expression of this relationship in a formula is of great value. The

formulas of the physical and biological sciences and, to a somewhat less extent, those of the social sciences are good examples. Illustrations of this point are the law of gravitation, or the formula of falling bodies, Boyle's law of gasses, the laws of electricity, or the formula for capitalizing income as presented in economics.

Mathematical Regression Lines

Whenever the relationship between two or more variables can be adequately expressed by a mathematical regression line, it is preferable to any other type of line because it is more exact and measures the relationship in a formula which can be manipulated algebraically for further analysis. There are so many different types of mathematical lines that it is usually possible to find one that will express with sufficient accuracy almost any relationship.

Least Squares Method

The type of line most frequently used is the one fitted by the least squares method. The derivation of this formula requires the use of the calculus and higher mathematics. In an elementary course in statistics such as this, no effort is made to introduce the student to the derivation of formulas. That should be reserved for more advanced courses. It is fortunate that it is not necessary for the student of statistics to understand fully the derivation of formulas in order to use them effectively. The brevity and hurry of human life make it necessary every day for us to use much equipment which we as individuals could not create. Many of us use typewriters efficiently, but few of us could create one or even repair one. Most adults drive automobiles, but very few of us could make one, and our inability to repair them is responsible for the large number of garages. The same principle applies to fountain pens, electric lights, locks and keys, clocks, watches, time tables, calendars, and a thousand other items without which modern life would be impossible. The same principle applies to statistical formulas. Already you have computed easily several types of averages and measures of dispersion without being able to create the formulas used. This is not to say that a knowledge of the calculus and higher mathematics is not very helpful to the

statistician. It is, and every student who wishes to become an efficient statistician or go far in research work should have two or three years of mathematics. But even so, it is possible to use statistical formulas effectively without being able to create them.

By "least squares" is meant a regression line which is drawn through the paired items of the data so that the vertical distances between the points and the regression line when squared and summed make the smallest total possible. It is a line which on the average comes nearest to all the points or items of data. There is but one possible location for such a line. This one best position of total least squares having been located, we have the best fitted line possible for any one given type of line. To move this line ever so little, up or down, or to turn it about on its own center point would increase the total squared distances of the points of data to the line. This relationship may be indicated by Fig. 51.

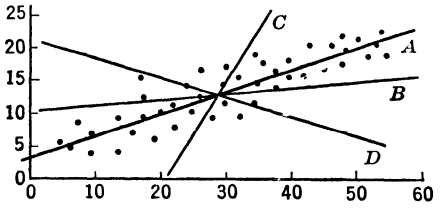


FIG. 51. Location of least squares regression line

In Fig. 51, line A is the best least squares line. It passes nearer all the points of data than any other line. If the regression line were shifted to the position of line B, some points would be closer to the B line than to the A line, but other points would be so much farther away that the total squared distances would be greater. The total squared distances would be even greater if the line were changed to the position of C or D. Not only the turning of the line but also the raising or lowering of the line will increase the sum of squares.

LOCATION AND DESCRIPTION OF STRAIGHT LINE

The mathematical formula for a straight line is

$$Y = a + bX$$

in which Y and X represent the measurements of pairs of data. Such a line is located in a plane as follows:

Meaning of symbols:

The Y-axis is the vertical line
The X-axis is the horizontal line
0 = point of origin, intersection of axes
$+ Y$ = measures on Y-axis above zero
$- Y$ = measures on Y-axis below zero
$+ X$ = measures on X-axis to right of zero
$- X$ = measures on X-axis to left of zero
Abscissa = a measurement along X
Ordinate = a measurement along Y

The upper right-hand quadrant where X and Y are both plus (+) is the part of the graph usually used in trend analysis, but the other quadrants are also required. Figure 52 illustrates these points.

In Fig. 52 three trend lines, MN, OP, and QR, are shown.

Meaning of symbols:

Y = dependent series of data measured on vertical axis
X = independent series of data measured on horizontal axis

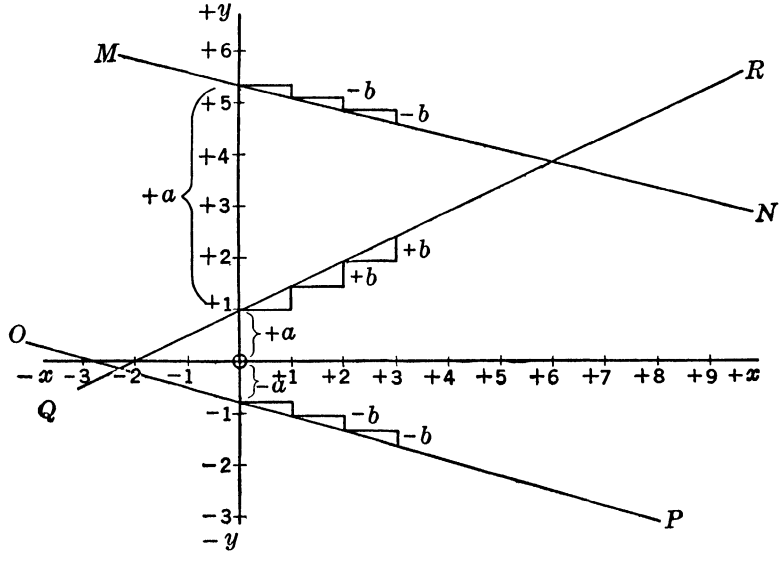


FIG. 52. Showing location and meaning of a and b in regression lines

a = Y-intercept, distance between point of origin, zero, and the point where the regression line crosses the Y-axis. It measures how high or low the line is on the graph. Below zero it is (−); above zero it is (+).

b = the slope of the line, how flat or steep the line is, how many units it changes on Y for the change of one unit on X. If values of Y decline as one moves to the right on X, b is minus (−). If values of Y increase as one moves to the right on X, b is plus (+).

X and Y are known. They are the paired items of the data.

a and b are the unknowns. Since there are two of them, they must be obtained by the solution of *two* simultaneous equations.

The worksheet and formulas for computing such a line are relatively simple and need give the student with a limited knowledge of algebra little difficulty. The two necessary equations are formed as follows:

Taking the basic equation, $Y = a + bX$, for each pair of items in the data, the relationships are expressed in Column 1.

Equations in Symbols

Column 1	Column 2 (Equations in Column 1 multiplied by coefficient of a)	Column 3 (Equations in Column 1 multiplied by coefficient of b)
$Y' = a + bX'$	$1Y' = 1a + b1X'$	$X'Y' = aX + bX^2$
$Y'' = a + bX''$	$1Y'' = 1a + b1X''$	$X''Y'' = aX + bX^2$
$Y''' = a + bX'''$	$1Y''' = 1a + b1X'''$	$X'''Y''' = aX + bX^2$
. . . . .	. . . . .	. . . . .
. . . . .	. . . . .	. . . . .
. . . . .	. . . . .	. . . . .
$Y_n = a + bX_n$	$1Y_n = 1a + b1X_n$	$X_nY_n = aX + bX^2$
	$\Sigma Y = Na + b\Sigma X$	$\Sigma XY = a\Sigma X + b\Sigma X^2$

In Column 2 we have each equation in Column 1 multiplied through by the coefficient of "a" in Column 1. This coefficient is unity and, therefore, does not change the size of the equation.

In Column 3 we have each equation in Column 1 multiplied

through by the coefficient of "b" which is X. The two equations derived by this method are called *The Normal Equations*, and are:

FORMULA No. 36

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

The solution of these two equations gives us values for a and for b.

Equations in Data

No. 1	No. 2	No. 3
$Y = a + bX$	(No. 1 multiplied by coefficient of a)	(No. 1 multiplied by coefficient of b)
$7 = a + b5$	$7 = 1a + 5b$	$35 = 5a + 25b$
$5 = a + b4$	$5 = 1a + 4b$	$20 = 4a + 16b$
$10 = a + b9$	$10 = 1a + 9b$	$90 = 9a + 81b$
$8 = a + b8$	$8 = 1a + 8b$	$64 = 8a + 64b$
$4 = a + b2$	$4 = 1a + 2b$	$8 = 2a + 4b$
$6 = a + b3$	$6 = 1a + 3b$	$18 = 3a + 9b$
$9 = a + b7$	$9 = 1a + 7b$	$63 = 7a + 49b$
	$49 = 7a + 38b$	$298 = 38a + 248b$

Solution of Simultaneous Equations

$298 = 38a + 248b$ (Divide equation by 38)
$49 = 7a + 38b$ (Divide equation by 7)

Divide each equation through by coefficient of a as follows:

No. 1 $7.842 = a + 6.526b$ Subtract No. 2 from No. 1
No. 2 $7.000 = a + 5.428b$
$.842 = 1.098b$

$$b = \frac{.842}{1.098} = .767$$

Substituting the numerical value of b in equation No. 2,

$$a = 7.0 - 5.428b = 7.0 - 4.163 = 2.837$$

For these data the equation, $Y = a + bX$, becomes

$$Y = 2.837 + .767X.$$

At the point on the X-axis where $X = 0$, the associated average value of $Y = a$, or 2.837. There is a specific Y-value for every specific value of X. The value for b, .767, indicates that for every change of one unit in X there is a corresponding change of .767 of one unit in Y.

Significance of Sign of b

The algebraic sign of b in the equation, $Y = a + bX$, is the indicator of the direction of the slope of the regression line. If b is plus (+) the relationship between X and Y is positive and values for Y increase in the ratio of $b : 1$ as values of X increase. On the graph the trend line rises to the right. If, however, the sign of b is minus (−) there is a negative or inverse relationship between X and Y, and values of Y decrease as values of X increase. The trend line declines as the X values increase.

The graph reveals that the actual values of the data do not fall exactly on the regression line. Since all values for Y estimated from the equation $Y = a + bX$ do fall exactly on the trend line, it is evident that there is a discrepancy or variation between the data and the computed average values. It is very important in statistics that this error be measured.

STANDARD ERROR OF ESTIMATE

Since one of the greatest values of the regression equation is its use in estimating average values for Y from specific values of X, and since these two quantities are usually not identical, the *standard error of estimate* was devised to measure this error. *The standard error of estimate is the square root of the mean of the squared deviations between the actual values of Y and the estimated values of Y.*

Symbols: S_y = standard error of estimate
Y' = estimated values of Y
$z = (Y - Y')$

FORMULA NO. 37

$$S_y = \sqrt{\frac{\Sigma(Y - Y')^2}{N}}$$

or

FORMULA NO. 38

$$S_y = \sqrt{\frac{\Sigma z^2}{N}}$$

The detailed worksheet for the computation of S_y from individual items of data is as follows:

WORKSHEET NO. 29

COMPUTATION OF STANDARD ERROR OF ESTIMATE FOR INDIVIDUAL ITEMS OF DATA

X	Y	$a + \underset{(.767X)}{(bX)} = Y'$	$(Y - Y')$ z	z^2
5	7	2.837 + 3.835 = 6.672	+ .328	.107584
4	5	2.837 + 3.068 = 5.905	− .905	.819025
9	10	2.837 + 6.903 = 9.740	+ .260	.067600
8	8	2.837 + 6.136 = 8.973	− .973	.946729
2	4	2.837 + 1.534 = 4.371	− .371	.137641
3	6	2.837 + 2.301 = 5.138	+ .862	.743044
7	9	2.837 + 5.369 = 8.206	+ .794	.630436
			− .005	3.452059

$$S_y = \sqrt{\frac{\Sigma z^2}{N}} = \sqrt{\frac{3.452059}{7}} = \sqrt{.49315} = .70$$

The standard error of estimate occupies the same relationship to a regression line that a standard deviation does to a mean. σ_y measures the scatter of items around a mean. S_y measures the scatter of the Y-data around the regression line. The S_y is always measured on the Y-axis vertically from the regression line. One S_y + and − from the regression line includes approximately .6827 of the items of data in a large sample, or about two-thirds of the data. The detailed method given in Worksheet No. 29 is presented not as a model to be followed in working problems of any considerable size but as a teaching device to hold all the relationships involved, as it were, under a methodological microscope so that the student could see every detail of the relationships and processes involved. For actual statistical work much shorter methods are available, but they tend to obscure the details of the processes for the beginning student. Formulas Nos. 43 and 44 supply a better method to follow in computing S_y for large samples. They can be worked from the deviations appearing in the correlation and regression worksheet.

MODEL LEAST SQUARES WORKSHEET

By using Worksheet No. 30 it is possible to reduce the volume of work required very materially. The first step is to change the form of the normal equations. These are transformed from

$$(1) \quad \Sigma Y = Na + b\Sigma X$$
$$(2) \quad \Sigma XY = a\Sigma X + b\Sigma X^2$$

in which only the total of the original data can be used, to the derived equations,

Formula No. 39		Formula No. 41
(1) $b_{yx} = \dfrac{\Sigma xy}{\Sigma x^2}$	or	$b_{xy} = \dfrac{\Sigma xy}{\Sigma y^2}$
Formula No. 40		**Formula No. 42**
(2) $a = \bar{Y} - b\bar{X}$	or	$a = \bar{X} - b\bar{Y}$

(1) Equation (1) of the normal equations is changed to equation (2) of the derived equations as follows:

$$\Sigma Y = N\bar{Y}, \Sigma X = N\bar{X},$$

therefore $\Sigma Y = Na + b\Sigma X$ equals $N\bar{Y} = Na + bN\bar{X}$

dividing by N the equation becomes

$$\bar{Y} = a + b\bar{X}, \quad \text{or} \quad a = \bar{Y} - b\bar{X}$$

(2) Equations (1) and (2) of the normal equations are changed to equation (1) of the derived equations as follows:

(1) Multiply normal equation (1) through by $\bar{X}$,

$$\Sigma Y\bar{X} = aN\bar{X} + b\Sigma X\bar{X}$$

(2) Change to $N\bar{Y}\bar{X} = aN\bar{X} + bN\bar{X}^2$

(3) Subtract from normal equation (2)

$$\Sigma XY = aN\bar{X} + b\Sigma X^2$$
$$N\bar{X}\bar{Y} = aN\bar{X} + bN\bar{X}^2$$
$$\Sigma XY - N\bar{X}\bar{Y} = b\Sigma X^2 - bN\bar{X}^2$$

Collect terms,

$$\Sigma XY - N\bar{X}\bar{Y} = b(\Sigma X^2 - N\bar{X}^2)$$

$$b = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma X^2 - N\bar{X}^2}$$

but	$(X - \bar{X}) = x$
square	$(X - \bar{X})^2 = x^2 = X^2 - 2X\bar{X} + \bar{X}^2 = x^2$
sum	$\Sigma X^2 - 2\Sigma X\bar{X} + N\bar{X}^2 = \Sigma x^2$
equals	$\Sigma X^2 - 2N\bar{X}^2 + N\bar{X}^2 = \Sigma x^2$
equals	$\Sigma X^2 - N\bar{X}^2 = \Sigma x^2$
and	$(X - \bar{X})(Y - \bar{Y}) = xy$
equals	$XY - X\bar{Y} - \bar{X}Y + \bar{X}\bar{Y} = xy$
sum	$\Sigma XY - \Sigma X\bar{Y} - \bar{X}\Sigma Y + N\bar{X}\bar{Y} = \Sigma xy$
and	$\Sigma XY - N\bar{X}\bar{Y} - N\bar{X}\bar{Y} + N\bar{X}\bar{Y} = \Sigma xy$
	$\Sigma XY - N\bar{X}\bar{Y} = \Sigma xy$

Therefore

$$b = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma X^2 - N\bar{X}^2} = \frac{\Sigma xy}{\Sigma x^2}$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

In the formula $b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$, in which b is written b_{yx}, the yx are subscripts of b and indicate that this particular b measures the relationship between X and Y when Y is the dependent variable on the vertical axis and X is the independent variable on the horizontal axis.

In the formula $b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$, in b_{xy}, xy are subscripts of b and indicate that this particular b measures the relationship between X and Y when X is the dependent variable on the vertical axis and Y is the independent variable on the horizontal axis. These two formulas for the same data will not give identical regression lines. Ordinarily only b_{yx} [read b (y on x)] is computed. In the equation $Y = a + bX$, the Y-values are estimated from the X-values, but in the equation $X = a + bY$, the X-values are estimated from the Y-values.

Method of Computing Deviations in Standard Worksheet No. 30

(1) The two paired variables of the original data are laid down in the Columns X and Y. Their squares and products are computed.

(2) The five columns are summed.

(3) Means are computed for X and Y.

(4) The corrections are computed as follows:

(a) ΣX times $\overline{X}$ = corrections for ΣX^2

(b) ΣY times $\overline{Y}$ = corrections for ΣY^2

(c) ΣX times $\overline{Y}$, or ΣY times $\overline{X}$, equals corrections for ΣXY.

(5) The deviations from the means are obtained by subtracting the corrections from the sums immediately above them.

Using the equations in terms of deviations, $b_{yx} = \dfrac{\Sigma xy}{\Sigma x^2}$ and $a = \overline{Y} - b\overline{X}$, the results for Worksheet No. 30 are:

WORKSHEET NO. 30

COMPUTATION OF LEAST SQUARES REGRESSION LINE BETWEEN HEIGHTS AND WEIGHTS OF 106 SCHOOL CHILDREN

Child No.	Height * X	Weight † Y	X^2	Y^2	XY
1	53	61	2809	3721	3233
2	46	40	2116	1600	1840
3	51	49	2601	2401	2499
. . .	. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
106	58	76	3364	5776	4408
Sums	5558	6896	293,595	475,714	368,650
Means	52.433	65.056			
Corrections			291,423	448,626	361,581
Deviations			2,172	27,088	7,069

* Inches. † Pounds.

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{7{,}069}{2{,}172} = 3.2546$$

$$a = \overline{Y} - b\overline{X} = 65.056 - 3.2546 \times 52.433$$
$$= 65.056 - 170.648$$
$$= -105.6$$

$Y = a + bX$ is $Y = -105.6 + 3.2546X$

From this equation it is possible to compute the expected or average estimated weight of a child, if his height is known. Since all such estimates are made from the average relationship in the sample there will be an error or deviation between the actual weight and the estimated weight in most cases. To measure this error or deviation the computation of the standard error of estimate is necessary. In any case in which the regression line passes through all the items of data, the standard error of estimate will be zero (0).

Short Method for Computing Standard Error of Estimate

It would involve too much cost of time and money to compute ordinarily S_y by the long method used in Worksheet No. 29. Fortunately a much shorter method is available.

FORMULA NO. 43

$$S_y = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N}}$$

or

FORMULA NO. 44[1]

$$S_y = \sqrt{\frac{\Sigma y^2 - b_{yx} \cdot \Sigma xy}{N}}$$

[1] $Y - Y' = z$, $(Y - b_{yx}X) = z$, $\left(Y - \frac{\Sigma xy}{\Sigma x^2} X\right) = z$, by shifting the point of origin from the intersections of the X and Y-axes to the intersection of the means of X and Y, the estimated Y becomes $y = bx$ in terms of deviations from the means.

Then
$$(y - bx)^2 = z^2$$
$$(y^2 - 2bxy + b^2x^2) = z^2$$
$$y^2 - 2\frac{\Sigma xy}{\Sigma x^2} \cdot \frac{xy}{1} + \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma x^2} \cdot \frac{x^2}{1} = z^2$$

Summed
$$\Sigma y^2 - 2\frac{\Sigma xy \cdot \Sigma xy}{\Sigma x^2} + \frac{(\Sigma xy)^2}{\Sigma x^2} = \Sigma z^2$$

For the weight of the school children S_y is 6.2 pounds.

$$S_y = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N}} \qquad \text{or} \qquad S_y = \sqrt{\frac{\Sigma y^2 - (b_{yx} \cdot \Sigma xy)}{N}}$$

$$= \sqrt{\frac{27{,}088 - \frac{(7{,}069)^2}{2{,}172}}{106}} \qquad = \sqrt{\frac{27{,}088 - 3.2546 \times 7{,}069}{106}}$$

$$= \sqrt{\frac{27{,}088 - 23{,}066.8}{106}} \qquad = \sqrt{\frac{27{,}088 - 23{,}006.8}{106}}$$

$$= \sqrt{\frac{4081.2}{106}} = \sqrt{38.5018} \qquad = \sqrt{\frac{4081.2}{106}} = \sqrt{38.5018}$$

$$= 6.2 \text{ pounds.} \qquad = 6.2 \text{ pounds.}$$

The standard error of estimate is expressed in terms of the units of the original data of the Y, or dependent variable.

The dependability of S_y, like all other statistics, rests on the adequacy of the sample from which it is computed. If the sample is too small it is subject to a large degree of variability and uncertainty.

---

and $$\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = \Sigma z^2$$

but $$S_y^2 = \frac{\Sigma z^2}{N}$$

Then $$S_y^2 = \frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N}$$

and $$S_y = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N}}$$

Since $$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

$$S_y = \sqrt{\frac{\Sigma y^2 - b_{yx}(\Sigma xy)}{N}}$$

WORKSHEET NO. 31

COMPUTATION OF LINEAR REGRESSION BETWEEN PER CAPITA BANK DEPOSITS DEC. 31, 1948 (X) AND PER CAPITA INCOME PAYMENTS FOR 1948 (Y), IN THE 25 LOWEST INCOME STATES*

State	Deposits (in $10.) X	Payments (in $10.) Y	X^2	Y^2	XY
Ala.	44	89	1,936	7,921	3,916
Ariz.	66	117	4,356	13,689	7,722
Ark.	42	86	1,764	7,396	3,612
Fla.	71	114	5,041	12,996	8,094
Ga.	53	97	2,809	9,409	5,141
Idaho	84	125	7,056	15,625	10,500
Kan.	86	129	7,396	16,641	11,094
Ky.	58	91	3,364	8,281	5,278
La.	67	100	4,489	10,000	6,700
Me.	78	122	6,084	14,884	9,516
Minn.	103	135	10,609	18,225	13,905
Miss.	39	76	1,521	5,776	2,964
Mo.	107	136	11,449	18,496	14,552
N.H.	100	126	10,000	15,876	12,600
N.Mex.	56	112	3,136	12,544	6,272
N.C.	49	93	2,401	8,649	4,557
Okla.	70	103	4,900	10,609	7,210
Ore.	87	130	7,569	16,900	11,310
S.C.	35	86	1,225	7,396	3,010
Tenn.	63	96	3,969	9,216	6,048
Texas	90	119	8,100	14,161	10,710
Utah	88	123	7,744	15,129	10,824
Vt.	90	123	8,100	15,129	11,070
Va.	60	116	3,600	13,456	6,960
W.Va.	51	113	2,601	12,769	5,763
Sums	1,737	2,757	131,219.00	311,173.00	199,328.00
Means	69.48	110.28			
Corrections			120,686.76	304,041.96	191,556.36
Deviations			10,532.24	7,131.04	7,771.64

* Source: *Statistical Abstract U.S. 1950*, pp. 266 and 390.

Worksheet No. 31 and the computations accompanying it represent a complete simple model for the computation of linear regression.

Coding the Data to Reduce the Labor. The original values of both X and Y are reduced in size by dividing both by 10. The original values for Alabama are X = \$443 and Y = \$891. Dividing by 10 and rounding we have X = \$44, Y = \$89. This short cut reduces the total volume of labor by more than one-half. The coding of the original data, however, requires a decoding of the results at the end of the computations.

"a" and "S_y" must be multiplied by the same number by which the original Y values were divided. "b" must be multiplied by the quotient obtained by dividing the Y code by the X code. In this case the number is 1, as $\frac{Y \text{ code} = 10}{X \text{ code} = 10} = 1$. Observe these adjustments at the end of the computation.

By using the deviations from the means instead of the original totals the X^2 are reduced from 131,219. to 10,532.24, the Y^2 from 311,173. to 7,131.04 and the XY from 199,328. to only 7,771.64, and also the use of simultaneous equations avoided. By these economical devices the labor required to compute a simple linear regression line is reduced to a small percentage of that otherwise required.

The computations are:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{7{,}771.64}{10{,}532.24} = .73789$$

$$a = \overline{Y} - b_{yx}\,\overline{X} = 110.28 - (.73789 \times 69.48)$$
$$= 110.28 - 51.27 = 59.01$$

$$\text{Decoding, } b_{yx} = .73789 \times \frac{Y \text{ code } 10}{X \text{ code } 10} = .73789 \times 1 = .73789$$

$$a = 59.01 \times 10 = \$590.10$$

Regression line, $Y = 590.10 + .73789X$

$$S_y = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N}} = \sqrt{\frac{7.131.04 - \frac{(7{,}771.64)^2}{10{,}532.24}}{25}}$$

$$= \sqrt{\frac{7{,}131.04 - \frac{60{,}398{,}388.29}{10{,}532.24}}{25}}$$

$$= \sqrt{\frac{7{,}131.04 - 5734.62}{25}} = \sqrt{\frac{1{,}396.42}{25}} = \sqrt{55.857} = 7.47$$

Also

$$S_y^2 = \frac{\Sigma y^2 - (b_{yx} \cdot \Sigma xy)}{N} = \frac{7{,}131.04 - (.73789 \times 7{,}771.64)}{25}$$

$$= \frac{1{,}396.43}{25}$$

$$= 55.857$$

$S_y = 7.47$. Decoded $S_y = 7.47 \times 10 = \$74.70$

b_{yx} = read b y on x which means that this is the regression coefficient for Y as the dependent variable and X as the independent variable. If the variables are switched the regression line changes and is b_{xy} in which case X is the dependent variable and Y the independent. This b_{xy} will be a different number and the new line will have a different slope.

The computations for b_{xy} are:

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{7{,}771.64}{7{,}131.04} = 1.0898$$

Decoding b_{xy} $1.0898 \frac{\text{Code } X \text{ } 10}{\text{Code } Y \text{ } 10} = 1.0898 \times 1 = 1.0898.$

$a = X - b_{yx}\,\overline{Y} = 69.48 - (1.0898 \times 110.28) = 69.48 - 120.18$
$= -50.70$. Decoded $a = -50.70 \times 10 = -507.0$

Regression line $X = a + b_{yx}Y = -507.0 + 1.0898Y$

$$Sx = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma xy)^2}{\Sigma y^2}}{N}} = \sqrt{\frac{10{,}532.24 - \frac{(7{,}771.64)}{7.131.04}}{25}}$$

$$= \sqrt{\frac{10{,}532.24 - \frac{60{,}398{,}388.29}{7{,}131.04}}{25}} = \sqrt{\frac{10{,}532.24 - 8{,}469.79}{25}}$$

$$= \frac{2{,}062.45}{25} = \sqrt{82.498}. \quad Sx = \$9.08$$

$$Sx^2 = \frac{\Sigma x^2 - (b_{yx} \cdot \Sigma xy)}{N} = \frac{10{,}532.24 - (1{,}0898 \cdot 7{,}771.64)}{25}$$

$$= \frac{10{,}532.24 - 8{,}469.53}{25} = \frac{2{,}062.71}{25} = 82.49. \quad Sx = \$9.08$$

Decoded $Sx = \$9.08 \times 10 = \$90.8$

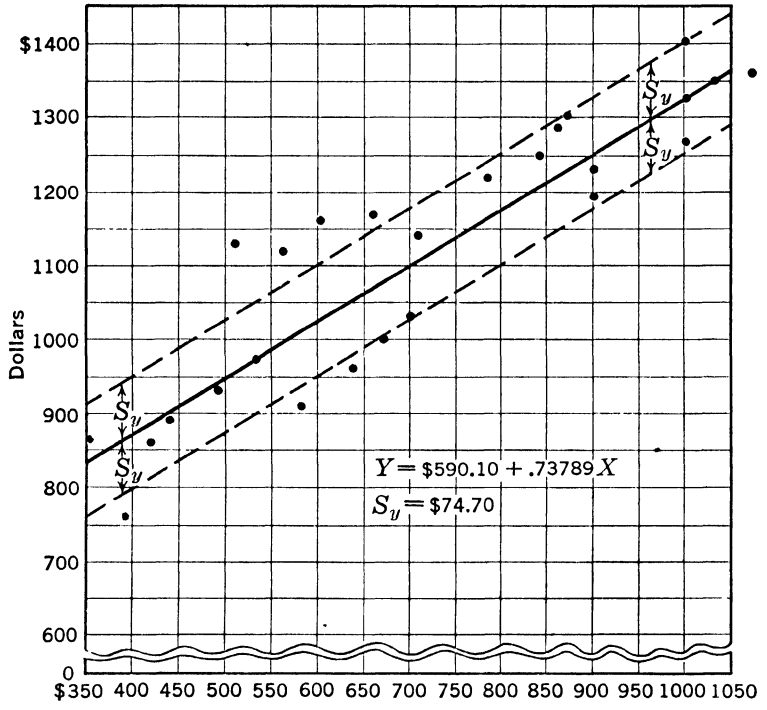


FIG. 53. Least squares regression line, and measures of standard error of estimate for per capita bank deposits (X), and per capita income payments (Y) for twenty-five states, 1948. (*Statistical Abstract of U.S.* See Worksheet No. 29)

The reason for using the data on only the twenty-five states with the lowest per capita income is to illustrate how one type of regression line may fit one portion of a population but be of no value for another portion or for the entire population. For the twenty-five states included in Worksheet No. 31 the regression line $Y = 590.10 + .73789X$ is quite accurate and from it one can

estimate per capita income payments (Y) from per capita bank deposits (X) with a standard error of estimate of only \$75.00.

If per capita bank deposits in a given state are \$1000, the chances are 68.27 times out of a hundred that per capita income is \$1328.00 $\pm$ \$75.00, which is a quite accurate estimate. But if this line is extrapolated to cover the other 23 states and the District of Columbia, the result will be a poor fit with a high degree of error. Such an estimate would miss the per capita income of Delaware by \$164.00 and that of New York by \$745.00.

The reason for this large error for the states with the large per capita incomes is that for these wealthier states per capita bank deposits increase much faster than does income. This result is due to the fact that in such states as California, Illinois, Massachusetts, New York and others the banking business of the nation is largely concentrated changing entirely the relation between per capita income and per capita bank deposits in these large industrial states.

When all forty-eight states and the District of Columbia are combined in one problem only a curvilinear line will fit well. For the larger states alone a more nearly flat line with a smaller b is required.

This problem is presented to show that a knowledge of statistical methods alone does not guarantee correct or useful results. Only a thorough knowledge of the field of data studied and the basic principles operating in it can lead a statistician to use his statistical tools properly or to obtain practical results. Like all complex tools statistical methods must be used with great care, which can come only from accurate thinking and long experience. Linear regression is quite limited because so many relationships in the physical, biological and social realms do not follow straight lines. This problem and Worksheet No. 32 indicate how important the material in Chapter 10 is. Since we can estimate the unknown income from the known bank deposits with a high degree of reliability, this regression line becomes valuable in advertising, marketing, production and business management. Such dependable estimates of the necessary unknown from the known are the basis of modern science and industry.

WORKSHEET NO. 32

COMPUTATION OF LEAST SQUARES STRAIGHT LINE REGRESSION BETWEEN COST OF WHEAT PER BUSHEL AND YIELD PER ACRE FOR 25 NORTH DAKOTA FARMS

	X *	Y †	XY	X^2	Y^2
	9	184	1656	81	33,856
	8	200	1600	64	40,000
	10	156	1560	100	24,336
	11	129	1419	121	16,641
	33	70	2310	1089	4,900
	20	67	1340	400	4,489
	34	64	2176	1156	4,096
	31	67	2077	961	4,489
	14	93	1302	196	8,649
	18	75	1350	324	5,625
	15	98	1470	225	9,604
	21	73	1533	441	5,329
	16	86	1376	256	7,396
	12	108	1296	144	11,664
	14	89	1246	196	7,921
	19	67	1273	361	4,489
	27	58	1566	729	3,364
	29	60	1740	841	3,600
	32	62	1984	1024	3,844
	17	86	1462	289	7,396
	24	50	1200	576	2,500
	22	54	1188	484	2,916
	26	48	1248	676	2,304
	28	49	1372	784	2,401
	30	60	1800	900	3,600
Sums	520	2,153	38,544	12,418	225,409
Means	20.8	86.12			
Corrections			44,782	10,816	185,416.4
Deviations			− 6,238	1,602	39,992.6

* Yields in bushels per acre.
† Costs in cents per bushel.

$$Y = a + bX$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-6{,}238}{1602} = -3.894$$

$$a = \bar{Y} - b\bar{X} = 86.12 - (-3.894)(20.8)$$
$$= 86.12 + 81.00 = 167.12$$
$$Y = 167.12 - 3.894X$$

$$S_y = \sqrt{\frac{\Sigma y^2 - b_{yx} \cdot \Sigma xy}{N}}$$

$$= \sqrt{\frac{39{,}992.6 - (-3.894)(-6{,}238)}{25}}$$

$$= \sqrt{\frac{39{,}992.6 - 24{,}290.8}{25}} = \sqrt{\frac{15{,}701.6}{25}}$$

$$= \sqrt{628.06} = 25.06 \text{ cents.}$$

A sufficient number of statistical methods have now been presented to make necessary a preliminary discussion of the three principal uses of statistics. Briefly, these may be listed as (1) description, (2) analysis, and (3) inference or estimation.

Description is the simplest use which can be made of statistics. Any sample or quantity of data may be thrown into an array and frequency distribution and have various averages and measures of dispersion computed and be depicted in charts and graphs with no further purpose than to give an objective mathematical and graphic description of these particular data. No further attempt may be made to analyze or point out further relationships which may be revealed or to search out and explain the inner and hitherto hidden meanings of the data. No effort may be made to use the data or this description of them to estimate, forecast, or project inferences as to other related data or to discover general principles or laws. Description of a bit of data is all that is sought and this having been accomplished no further labors are performed.

Description, however, is usually only a means to an end and is ordinarily done to bring to light obscured relationships which it is desired to analyze and explain. Analysis and explanation are always the legitimate ends of description. Usually a sample of

wheat yields is taken, described, and measured only to reveal relationships among wheat yields and among them and other variables such as rainfall, time of planting, kind of seed, condition of seed bed, types of winter weather, prevalence of pests, and cost in order to analyze complex farm management problems. An insight into relationships and an insight into hidden factors are the end of the study.

But analysis is usually not an end in itself. One analyzes in order that he may comprehend and forecast, infer, project, and apply his information for further practical or theoretical use. The great end of statistical description and analysis is *inference*. To make logical and accurate inferences on the basis of statistical analysis is a difficult procedure which requires both (1) a full and accurate knowledge of the subject matter studied and (2) a sufficient knowledge of mathematical methods, logical analysis, and wholesome caution. Regression is a basis for estimating. But since it rests on the average relationships among paired values of data, the smallness of the relationship or the smallness of the sample or both will usually bring an error into the computation which will limit the force of the inference drawn from the regression equation. Most situations are so complex, especially in the biological and social sciences, that no single independent variable accounts for all the change in any specific dependent variable. This fact means that there will be a considerable error in any estimate made from any regression equation which must be taken into consideration in any inference that is drawn.

The beginning student should be especially careful not to include more in his inference than is warranted by the careful analysis of the data. Any estimate of the weight of a child from his height will likely contain an error because the weight of children is also influenced by health, heredity, age, and adequate food. Per capita sales do not depend at all on magazine circulation, but both may depend to some degree on the common factor of purchasing power. If persons have more wealth they subscribe to more magazines and at the same time make more purchases of most other goods. Those with little money may buy few or no magazines and only small quantities of other merchandise.

SUMMARY

1. Regression is the measure of the average relationship between two variables in terms of the units of the original data. Linear regression is a straight line relationship.

2. The regression is positive when the number of units of the dependent variable increase with an increase in the number of units of the independent variable. The values on Y increase as the values on X increase.

3. The regression is negative when the number of units of the dependent variable decrease with an increase of the independent variable. The values on Y decrease as the values on X increase.

4. This relation may be described and measured by a freehand line drawn through the plotted points of data on a coordinate scale.

5. This relation may be described and measured by a straight line drawn through the class averages of the data divided into two approximately equal parts.

6. This relationship may be described and measured by a least squares straight line of the form $Y = a + bX$.

7. The estimated values of the dependent variable may be computed from this equation for all values of the independent variable.

8. The standard error of estimate is the square root of the averaged squared deviations of the data from the regression line. This measure is to a regression line what a standard deviation is to an arithmetic mean. One Sy, plus and minus, includes in a normal distribution about 68% of the items of the sample.

9. Regression lines and their standard errors of estimates are effective means for discovering scientific relationships and for expressing these relationships in accurate mathematical terms for scientific summary and inference.

SELECTED REFERENCES

Brumbaugh, M. A. and Kellogg, L. S., *Business Statistics*, Chapter XXII, Irwin, Inc. Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XV, Prentice-Hall, Inc., New York, 1948.

Ezekiel, M., *Methods of Correlation Analysis*, Chapter V, John Wiley & Sons, Inc., New York, 1941.

Mills, F. C., *Statistical Methods, Revised*, Chapter VII, Henry Holt & Co., New York, 1938.

Sasuly, Max, *Trend Analysis of Statistics, Theory and Technique*, The Brookings Institute, Washington, D.C., 1934.

Snedecor, G. W., *Statistical Methods*, Chapter VI, Collegiate Press, Inc , Ames, Iowa, 1938.

REVIEW QUESTIONS

1. What does the term "regression" as used in statistics mean?
2. Define "science" and show how regression is related to scientific knowledge and discoveries.
3. Explain the four steps required to make a "freehand curve."
4. What are the advantages and disadvantages of a freehand curve?
5. Explain the six steps required to make a regression line through class averages.
6. Compare the regression line through class averages with the freehand curve as to method and reliability.
7. Can the standard error of estimate and the coefficient of correlation be computed for either or both of these lines and if so, by what methods and formulas?
8. What is a "least squares" regression line?
9. What do "a" and "b" signify in the equation, $Y = a + bX$? Explain fully.
10. How are the "normal equations" obtained?
11. What is the difference between the long and short methods of computing a least squares regression line?
12. Show the complete algebraic transformation of the two normal equations

$$\begin{matrix}(\Sigma Y = Na + b\Sigma X) \\ (\Sigma XY = a\Sigma X + b\Sigma X^2)\end{matrix} \text{ into the equation } b = \frac{\Sigma xy}{\Sigma x^2}$$

13. Show that $(\Sigma Y = Na + bX) = (a = \overline{Y} - b\overline{X})$.
14. What is the standard error of estimate? How may it be computed? Why is it important in connection with regression lines? Explain fully.
15. What does extrapolation mean? To what extent is it dependable?
16. In what units is the standard error of estimate measured? Why?
17. What changes occur in the standard error of estimate as the "fit" of the regression improves? Why?
18. What is the difference between the equation, $Y = a + b_{yx}X$ and the equation, $X = a + b_{xy}Y$? Explain fully.
19. Explain why no regression line is complete without its standard error of estimate.
20. What advantages are there in a mathematical regression line over the free-hand type?

EXERCISES

1. Compute the regression line between per capita income (X) and life insurance in force (Y) for the 48 states and D.C. (1948 data).

State	Per Cap. Income	Per Cap. Insurance	State	Per Cap. Income	Per Cap. Insurance
Ala.	$ 891	$ 800	Neb.	$1,473	$1,200
Ariz.	1,168	800	Nev.	1,679	900
Ark.	863	500	N.H.	1,261	1,400
Calif.	1,651	1,200	N.J.	1,605	1,900
Colo.	1,429	1,400	N.Mex.	1,125	600
Conn.	1,700	2,000	N.Y.	1,891	2,100
Del.	1,741	2,000	N.C.	930	900
D.C.	1,691	2,100	N.D.	1,473	800
Fla.	1,137	1,100	Ohio	1,548	1,700
Ga.	971	1,000	Okla.	1,029	800
Ida.	1,252	800	Ore.	1,302	900
Ill.	1,817	1,800	Penn.	1,444	1,700
Ind.	1,403	1,400	R.I.	1,564	1,800
Iowa	1,491	1,200	S.C.	865	900
Kan.	1,291	1,100	S.D.	1,577	800
Ky.	909	800	Tenn.	955	900
La.	1,002	800	Tex.	1,192	1,000
Me.	1,219	1,100	Utah.	1,231	1,100
Md.	1,546	1,500	Vt.	1,229	1,200
Mass.	1,509	1,600	Va.	1,159	1,100
Mich.	1,484	1,500	Wash.	1,453	1,000
Minn.	1,353	1,200	W.Va.	1,133	1,000
Miss.	758	400	Wis.	1,443	1,300
Mo.	1,356	1,400	Wyo.	1,494	900
Mont.	1,791	1,100			

1. Source: *Life Insurance Fact Book*, and *U.S. Statistical Abstract*
2. Compute $Y = a + bX$, S_y.
3. These data may be coded by dividing by 10, 50, or 100 as the teacher may direct, reducing the size of the numbers.
4. Only a portion of the data may be used. Fifteen to twenty-five states are sufficient for one problem.
5. This exercise may be used to compute (1) freehand curves, (2) lines through class averages, and (3) least squares lines. They may also be used with Chapter 9.

2. Relationship between average annual relative humidity (X) and average annual precipitation (Y) for 50 American cities through 1945.

City	Humidity (%)	Precipitation (in.)	City	Humidity (%)	Precipitation (in.)
Mobile	72	62	Helena	56	12
Montgomery	67	52	North Platte	62	18
Phoenix	38	8	Omaha	67	28
Little Rock	66	48	Atlantic City	75	41
Fresno	52	10	Albuquerque	47	8
Los Angeles	62	16	Albany	72	37
San Francisco	73	22	New York	65	43
Denver	48	14	Asheville	68	38
Washington	67	42	Raleigh	72	47
Jacksonville	73	51	Bismarck	66	16
Miami	73	59	Cleveland	61	33
Atlanta	69	48	Okla. City	66	32
Boise	52	13	Portland, Ore.	70	42
Chicago	69	33	Harrisburg	64	38
Des Moines	71	32	Pittsburgh	67	36
Wichita	62	30	Charleston	72	49
Louisville	65	42	Huron	66	19
New Orleans	72	60	Nashville	69	46
Spokane	60	16	Amarillo	55	20
Boston	68	41	El Paso	38	9
Detroit	71	31	Ft. Worth	63	32
Minneapolis	64	27	Houston	73	46
Vicksburg	73	52	St. Lake	57	14
Kansas City	66	36	Richmond	68	42
St. Louis	64	39	Cheyenne	53	15

1. Twenty or twenty-five items are sufficient for one problem. By switching combinations of cities these data may be used for many semesters without much repetition.
2. Compute $y = a + bX$ and S_y. 15 to 20 items are sufficient for one problem.
3. These data may be used to compute (1) freehand curves, (2) lines through class means, and (3) least squares lines. They may also be used with Chapter 9 and Chapter 12.

3. The wholesale price of copper, cents per lb., N.Y. (X) and total imports of copper in 1000 of tons (Y). (1913–1940.)

Price	Imports	Price	Imports	Price	Imports
39	15.3	34	14.4	13	7.0
35	13.6	46	13.0	25	8.4
28	17.3	45	14.0	24	8.6
32	27.2	40	13.8	22	9.5
46	27.2	46	12.9	29	13.1
30	24.6	47	14.6	35	10.0
21	18.7	41	18.1	35	11.0
26	17.5	31	13.0	36	11.3
26	12.5	23	8.1		
31	13.4	12	5.6		

1. Source: *Survey of Current Business.*
2. Compute $Y = a + bX$ and S_y.

4. Price of crude oil at wells, Kansas & Oklahoma in dollars per bbl. (X), and number of wells completed (Y). (1913–1944.)

Price	Wells	Price	Wells	Price	Wells
$0.93	1,592	$1.44	1,351	$0.57	1,043
.80	1,359	1.45	1,216	.94	1,259
.58	763	1.67	1,380	.94	1,483
1.26	1,565	1.88	1,584	1.04	1,845
1.78	1,383	1.29	1,204	1.16	1,593
2.20	1,487	1.20	1,044	1.12	1,457
2.28	1,754	1.23	970	.95	1,594
3.40	2,023	1.11	566	.96	1,605
1.73	1,222	.57	870	1.06	877
1.80	1,444	.81	672	1.11	811

1. Source: *Survey of Current Business.*
2. Compute $Y = a + bX$ and S_y.
3. These data may be used to compute (1) freehand curves, (2) lines through class means, and (3) least squares lines. They may be used with Chapter 9, Chapter 10 and Chapter 12.

5. Height, weight, and age of 106 Stillwater, Oklahoma, School Children of grades 1 to 6 inclusive, 1939.

Height Inches	Weight Lbs.	Age Months	Height Inches	Weight Lbs.	Age Months	Height Inches	Weight Lbs.	Age Months
53	61	110	51	65	106	49	52	107
46	40	80	48	59	88	51	65	120
51	49	108	47	52	86	48	56	89
60	98	142	54	73	122	41	41	72
53	57	103	54	88	132	60	99	151
62	92	130	48	54	96	56	83	141
54	54	106	57	64	138	50	53	101
53	61	104	43	46	75	51	63	104
49	54	104	52	62	95	53	69	107
53	82	110	47	47	87	59	83	140
49	53	91	47	57	96	51	54	102
53	55	106	45	47	75	51	61	103
51	58	96	56	69	133	56	80	140
54	61	112	66	102	173	51	55	108
47	45	81	56	90	127	48	57	78
57	75	117	47	56	82	47	50	84
50	48	92	56	108	168	50	57	92
58	84	139	52	61	110	48	52	89
53	66	109	50	54	82	58	81	118
59	74	141	55	72	116	62	89	131
44	42	75	52	59	110	47	57	108
43	38	72	50	50	97	54	63	120
56	69	127	54	70	105	54	66	101
56	73	134	57	66	118	48	49	83
59	96	144	54	67	113	49	52	90
60	94	161	53	62	134	51	65	97
54	67	121	56	71	128	51	65	108
60	92	143	57	100	131	50	48	98
53	68	108	48	48	85	53	52	99
59	86	142	59	79	131	59	98	141
48	53	107	47	53	83	55	70	107
49	54	87	44	47	78	50	55	109
54	70	116	61	100	154	52	70	102
47	47	96	54	78	130	53	57	103
48	49	84	49	50	88	58	76	129
53	62	104						

CHAPTER 9

CORRELATION AND DETERMINATION

Statistical methods are devices for analyzing the relationships which exist within and among groups of related numbers. The more complete and varied our analysis is, the better we shall understand the meaning of the numbers. Up to this time we have learned to measure separate frequency distributions as to central tendency and dispersion and also the relationship between two related frequency distributions by means of regression lines. All of these measures are in terms of the units of the original data. If our original data are bushels of wheat, all our measures, averages, means, medians, modes, dispersions, standard deviation, regression and standard error of estimate — all will be so many bushels of wheat. If our original data are in tons, all our statistics so far calculated will be in tons. If our original data are in dollars, all the statistics we have learned to compute up to this time will also be in dollars, except the coefficient of variation.

There is, however, need for another type of measure, a ratio, or relative number, or percentage statement of the relationship between two variables. What per cent of the average weight of a group of persons is associated with their average height? To what degree is variation in the two variables associated?

This measure is not expressed in terms of the units of the original data, but is an abstract number based on 1, or 100, as unity. It is ordinarily called the coefficient of correlation. Cor-

relation is the square root of a percentage, and is, therefore, quite misleading to the beginner in statistics.

A more easily understood, and in many respects a better measure, is the coefficient of *determination*, which is a true percentage measuring the degree of similarity between the variations of two factors. Determination is the square of correlation, and is coming into general use as the more accurate and easily understood measure. Both will be explained and computed in this chapter. Correlation will be presented first because it is the older of the two.

The relationship between determination and correlation may be illustrated as follows:

Coefficient of Correlation $r = \sqrt{\%}$	*Coefficient of Determination* $r^2 = \%$
1.00	1.00
.90	.81
.80	.64
.70	.49
.60	.36
.50	.25
.40	.16
.30	.09
.20	.04
.10	.01

Determination is a true percentage and as such is easily understood and correctly used. For instance, one might think that $r = .3$ is one-half of $r = .6$, but it is in fact but one-fourth. One might think that $r = .2$ is one-fourth of $r = .8$, but actually it is only one-sixteenth of .8.

Correlation is a measure of the amount of variation of one variable that is associated with or accounted for by variation in another variable or variables.

In the temperate and frigid zones one observes that the directness and intensity of the sun's rays vary with the change in seasons of the year. The temperature of the earth's surface and the atmosphere also vary. Many factors may affect the temperature locally such as elevation above the sea, nearness to water, direc-

tion of winds, and the protection of mountains, but in spite of all these, on the average, the two variables of earth temperatures and directness of solar rays vary together. The more direct the rays the greater is the heat. The less direct the rays the less intense is the heat. These quantities rise and fall together. Correlation, or determination, is a method of measuring the similarity of the change in these two variables.

Another case of correlation is the similarity of the variations of the amount of rainfall and plant growth. If the rainfall is less than five inches a year, the earth's surface is almost bare. As rainfall increases, vegetation increases until dense growth or jungles appear when the precipitation is over 80 or 100 inches a year. Other factors are essential to vegetable growth besides moisture, such as temperature, soil fertility, and cultivation, but on the average, variation in rainfall and variation in plant growth do move closely together. The exact relationship between quantity of water and rate of plant growth has been worked out for certain crops in irrigated areas. Correlation and determination are measures of the degree of association in the movements of two or more variables.

The changes in the variation of two variables may be in the same direction, as in the two illustrations above, or they may be in the opposite direction, as in the case of the amount of cotton produced and its price. Plant growth increases as rainfall increases, but the price of a commodity tends to fall as the supply increases. When the changes in the variation of two or more variables move in the same direction, the correlation is said to be positive. When the changes in variation move in opposite directions, the correlation is said to be negative. Correlation may range from perfect positive correlation, + 1, through zero (0), or no correlation, to − 1, or perfect negative correlation.

Figure 54 shows the simplest explanation of correlation, the geometric or graphic presentation. The closer the plotted points of data come to the regression line, the higher is the correlation. The more widely they are scattered, the lower the correlation.

The first step in analyzing the correlation of data is the construction of a scatter diagram of the paired items, as shown above.

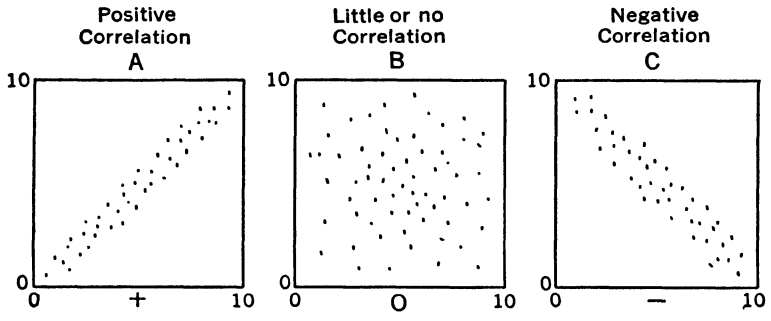


FIG. 54. Indicating the variation in correlation from high plus through 0 to high negative

Although the scatter diagram does not give a mathematical measure of the relationship, it does indicate quite clearly whether any significant correlation does exist between the variables.

In Fig. 54A, only low values of X are associated with low values of Y, and only high values of X are associated with high values of Y. In Fig. 54C, the reverse is true. Small X values accompany only large Y values, and large X values have only small Y values. In Fig. 54B, however, any value of X, large or small, is associated with any Y value, large or small. In such a case there is little or no correlation. *If the correlation is perfect, the plotted points of the data all fall exactly on the regression line.* Since all estimated values of the data, the values computed from the regression equation, $Y = a + bX$, or whatever it may be, always fall exactly on the regression line, it follows that *when correlation is perfect, the actual data and the estimated values are identical.* In such cases both fall exactly on the regression line. If the correlation is less than perfect, the plotted points will be scattered around the regression line. The smaller the correlation is, the wider is the scatter of dots, or data.

REGRESSION AND CORRELATION

For any specific data there is a close relationship between the regression line and the coefficient of correlation. The latter definitely depends upon the former. If the regression line does

not fit the data well, it is impossible to obtain the full correlation. If a straight line is fitted to data that are truly curvilinear, only a portion of the full correlation actually existing in the sample can be obtained. One can easily see from Fig. 55 that a curved

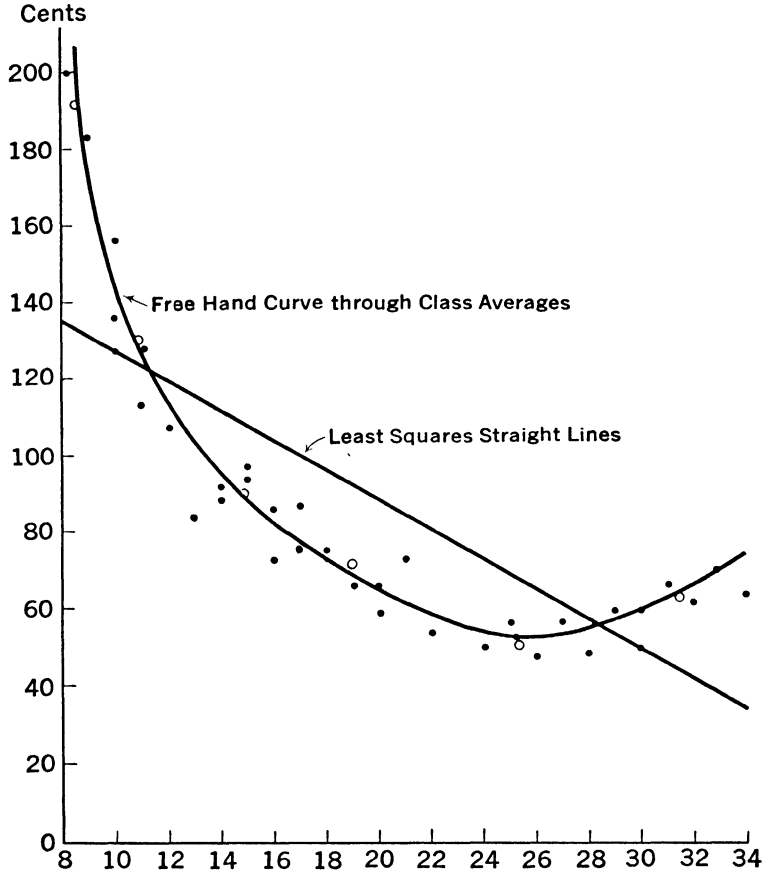


FIG. 55. Comparison of (*A*) freehand curve through class averages, and (*B* least squares straight line for (*y*) cost per bushel and (*x*) yield per acre of wheat. See Worksheet No. 30.

line would fit the data on cost of wheat per bushel and yield per acre much better than a straight line. For the straight line relationship the determination, r^2, is 57.2%. For the parabola the determination is 89.2%. Since the simple parabola is not a per-

fect regression fit for these data, it follows that the figure 89.2% is not all the determination actually in the sample. The freehand curve computed in Worksheet No. 38 and shown in Fig. 55 is a much better fit and gives a determination of about 98%. The student, therefore, should always keep clearly in mind the fact that *a full and adequate measure of correlation is dependent on a full and adequate measurement of the regression.* If you fail to get a correct regression line, you cannot get a full measure of correlation.

The relationship between regression and correlation is so close that some authors combine the two in a single treatment as though they were the same thing. Their separate treatment is justified by the fact that regression is measured in terms of the original units of the data, while correlation is expressed as a ratio or abstract number.

A simple problem will best illustrate this idea.

Meaning of symbols:

Y' = estimated values of Y, obtained from regression equation, $Y' = a + bX$, etc.

$\sigma_{y'}$ = standard deviation of estimated y''s.

σ_y = standard deviation of original Y's, the Y-values of the original data.

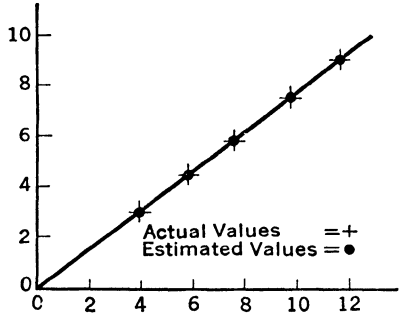


FIG. 56. Showing location of actual and estimated values as identical. Correlation is 1.

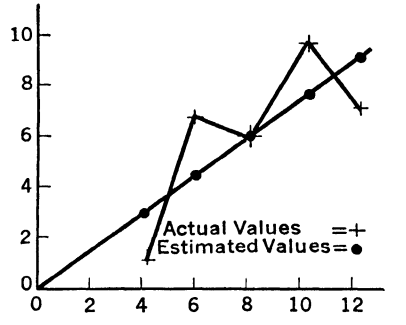


FIG. 57. Showing location of actual and estimated values not identical. Correlation less than 1.

Perfect Correlation

Regression for Fig. 56

	X	Y	XY	X^2
	4	3	12	16
	6	4.5	27	36
	8	6	48	64
	10	7.5	75	100
	12	9	108	144
Σ	40	30	270	360
Means	8	6		
Corrections			240	320
Deviations			30	40

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{30}{40} = .75$$

$$a = \bar{Y} - b\bar{X}$$
$$= 6 - .75 \cdot 8$$
$$= 6 - 6$$
$$= 0$$

$$Y = a + bX$$
$$Y = 0 + .75X$$

Standard Deviation of Y for Fig. 56

Y		$\bar{Y}$		y	y^2
3.0	−	6	=	− 3	9
4.5	−	6	=	− 1.5	2.25
6.0	−	6	=	0	
7.5	−	6	=	+ 1.5	2.25
9.0	−	6	=	+ 3	9
					22.50

$$\sigma_y^2 = \frac{\Sigma y^2}{N - 1} = \frac{22.5}{4}$$
$$= 5.625$$
$$\sigma_y = \sqrt{5.625}$$

Less than Perfect Correlation

Regression for Fig. 57

	X	Y	XY	X^2
	4	1	4	16
	6	6.5	39	36
	8	6	48	64
	10	9.5	95	100
	12	7	84	144
Σ	40	30	270	360
Means	8	6		
Corrections			240	320
Deviations			30	40

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{30}{40} = .75$$

$$a = Y - b\bar{X}$$
$$= 6 - .75 \cdot 8$$
$$= 6 - 6$$
$$= 0$$

$$Y = a + bX$$
$$Y = 0 + .75X$$

Standard Deviation of Y for Fig. 57

Y		$\bar{Y}$		y	y^2
1.0	−	6	=	− 5.0	25.0
6.5	−	6	=	+ 0.5	.25
6.0	−	6	=	0	
9.5	−	6	=	+ 3.5	12.25
7.0	−	6	=	+ 1.0	1.0
					38.50

$$\sigma_y^2 = \frac{\Sigma y^2}{N - 1} = \frac{38.50}{4}$$
$$= 9.625$$
$$\sigma_y = \sqrt{9.625}$$

In the above example the two regression lines are identical, but the deviations of the data from those lines are not identical in the two cases. The wider the data scatters about the line, the smaller the correlation.

Standard Error of Estimate on Y for Fig. 56

X	$Y - Y' = z$	z^2
4	3.0 − 3.0 = 0	0
6	4.5 − 4.5 = 0	0
8	6.0 − 6.0 = 0	0
10	7.5 − 7.5 = 0	0
12	9.0 − 9.0 = 0	0
		0

$$S_y^2 = \frac{\Sigma z^2}{N-1} = \frac{0}{4} = 0$$

Standard Error of Estimate on Y for Fig. 57

X	$Y - Y' = z$	z^2
4	1.0 − 3.0 = − 2	4
6	6.5 − 4.5 = + 2	4
8	6.0 − 6.0 = 0	0
10	9.5 − 7.5 = + 2	4
12	7.0 − 9.0 = − 2	4
		16

$$S_y^2 = \frac{\Sigma z^2}{N-1} = \frac{16}{4} = 4$$

In Fig. 56 all the actual data and all the estimated values fall on the regression line $Y = 0 + .75X$ and $S_y = 0$. In this case the standard deviation of the Y-values is $\sqrt{5.625}$. Since the estimated values of Y, or Y', are identical with the actual values of Y, the standard deviation of Y' is also $\sqrt{5.625}$. From this logical relationship the standard, or basic, formulas for determination and correlation may be derived as follows:

Determination

$$r^2 = \frac{\sigma_{y'}^2}{\sigma_y^2} = \frac{5.625}{5.625} = 1$$

Correlation

$$r = \frac{\sigma_{y'}}{\sigma_y} = \frac{\sqrt{5.625}}{\sqrt{5.625}} = 1$$

The determination will be one, however, only when $S_y = 0$ and the y and y'-values are identical. In Fig. 57 this is not the case. Here $\sigma_y^2 = 9.625$ and $\sigma_{y'}^2 = 5.625$. Therefore,

$$r^2 = \frac{\sigma_{y'}^2}{\sigma_y^2} = \frac{5.625}{9.625} = .584 = 58.4\%$$

FORMULA NO. 45

$r^2 = \dfrac{\sigma_{y'}^2}{\sigma_y^2}$ for determination

FORMULA NO. 46

$r = \dfrac{\sigma_{y'}}{\sigma_y}$ for correlation

These relationships may be restated as follows:

$$\frac{\sigma_y^2}{\sigma_y^2} = \frac{\sigma_y'^2}{\sigma_y^2} + \frac{S_y^2}{\sigma_y^2}$$

But

$$\frac{\sigma_y^2}{\sigma_y^2} = 1$$

and by definition

$$\frac{\sigma_y'^2}{\sigma_y^2} = r^2,$$

therefore, transposing terms and solving for r^2 we have,

$$\left(\frac{\sigma_y'^2}{\sigma_y^2}\right) = \left(\frac{\sigma_y^2}{\sigma_y^2}\right) - \left(\frac{S_y^2}{\sigma_y^2}\right) \quad \text{or} \quad r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

FORMULA NO. 47

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad \text{for determination}$$

FORMULA NO. 48

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} \quad \text{for correlation}$$

These formulas are based on the following facts and relationships:

σ_y^2 = the total amount of variation in Y that is *to be* explained.

$\sigma_y'^2$ = that portion of the total variation of Y that *is* explained by X.

S_y^2 = that portion of the total variation of Y that is *not* explained by X.

Since the amount of variation that *is* explained by X plus the amount that is *not* explained by X equals the total amount of variation in Y, it necessarily follows that $\frac{S_y^2}{\sigma_y^2}$, the squared standard error of estimate over the standard deviation of Y is the supplement of the coefficient of determination. In exact proportion as one becomes larger, the other becomes smaller. As r ap-

proaches 0, $\frac{S_y}{\sigma_y}$ approaches 1; as r approaches 1, $\frac{S_y}{\sigma_y}$ approaches 0.

$$r^2 = \left(\frac{\sigma_y^2}{\sigma_y^2}\right) = \text{Coefficient of determination}$$

$$1 - r^2 = \left(\frac{S_y^2}{\sigma_y^2}\right) = \text{Coefficient of non-determination}$$

$$\sqrt{r^2} = \text{Coefficient of correlation}$$

$$1 - r = \sqrt{\frac{S_y^2}{\sigma_y^2}} = \text{Coefficient of alienation}$$

The coefficient of alienation is the square root of the coefficient of non-determination, just as the coefficient of correlation is the square root of the coefficient of determination.

Formulas 45 and 46 are rarely used in actual statistical computations, although they are the actual basis of all correlation formulas. Formulas 47 and 48 are widely used because they are easily manipulated in connection with regression computations. No regression line is complete without its standard error of estimate, and S_y must always be computed. With this value at hand, the computation of

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

is almost no additonal work. Another relationship between the three values of σ_y, S_y, and r should be noted at this point.

FORMULA NO. 49

$$S_y = \sigma_y\sqrt{1 - r^2}$$

PEARSONIAN [1] COEFFICIENT OF CORRELATION

FORMULA NO. 50

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

[1] Named for Karl Pearson who originated it.

This formula is obtained from $r = \frac{\sigma_{y'}}{\sigma_y}$ by algebraic manipulation:

By shifting the measurement of the relationship between X and Y on a straight least squares regression line from the intersection of the X-axis with the Y-axis, or zero, to the intersection of the two means $\bar{Y}$ and $\bar{X}$, the equation $Y = a + bX$ is reduced to $y = bx$. Since a is the measure of the Y intercept, it disappears from the equation when we leave the Y-axis, or the point of origin. If from the equation

$$Y = a + bX$$

we subtract

$$\bar{Y} \qquad \bar{X}$$

the result is

$$Y - \bar{Y} = b(X - \bar{X})$$

which is

$$y = bx,$$

we have shifted from the large X and Y of the original data to the deviations from the mean as shown in Fig. 58.

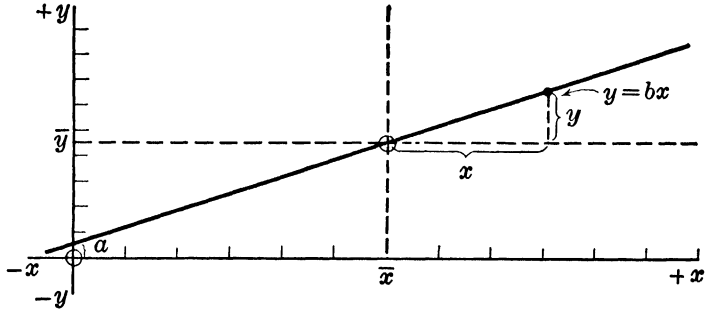


FIG. 58. Showing shift from $Y = a + bX$ to $y = bx$

In both equations $Y = a + bX$ and $y = bx$, the slope of the regression line, b, is the same. The only difference is that in the first case the relationship between X and Y is measured in terms of the original data, while in the second case it is measured in deviations from the means, $\bar{Y}$ and $\bar{X}$, and, therefore, in terms of $Y - \bar{Y} = y$ and $X - \bar{X} = x$. The estimated values for Y are no longer Y' but y'. Since $\sigma_{y'}$ is based on y', it follows that bx may be substituted for its equivalent y' in the equation $y' = bx$. Therefore, the equation

$r = \sqrt{\frac{\sigma_{y'}^2}{\sigma_y^2}}$ may be written $r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$

This formula is easy to compute, but is limited to straight regression lines. It is widely used in such relationships.

WORKSHEET NO. 33

DETAILED WORKSHEET FOR USING INDIVIDUAL ITEMS IN COMPUTATION OF THE PEARSONIAN COEFFICIENT OF CORRELATION

X		$\bar{X}$		x	x^2	Y		$\bar{Y}$		y	y^2	xy
12	−	10	=	+ 2	4	14	−	9	=	+ 5	25	+ 10
9	−	10	=	− 1	1	8	−	9	=	− 1	1	+ 1
8	−	10	=	− 2	4	6	−	9	=	− 3	9	+ 6
10	−	10	=	0		9	−	9	=	0		
11	−	10	=	+ 1	1	11	−	9	=	+ 2	4	+ 2
13	−	10	=	+ 3	9	12	−	9	=	+ 3	9	+ 9
7	−	10	=	− 3	9	3	−	9	=	− 6	36	+ 18
					28						84	46

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{28}{7}} = \sqrt{4} = 2$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{84}{7}} = \sqrt{12} = 3.464$$

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{46}{7 \times 2 \times 3.464} = \frac{46}{48.496} = .95$$

[1] The transformation is

$$\sigma_y' \sqrt{\frac{\Sigma(bx)^2}{N}} = \sqrt{\frac{\Sigma\left[\left(\frac{\Sigma xy}{\Sigma x^2}\right)x\right]^2}{N}} = \sqrt{\frac{\frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma x^2} \times \frac{\Sigma x^2}{1}}{N}} = \sqrt{\frac{\frac{(\Sigma xy)^2}{\Sigma x^2}}{N}}$$

but in the denominator of $r^2 = \frac{\sigma_{y'}^2}{\sigma_y^2}$, σ_y^2 equals $\frac{\Sigma y^2}{N}$.

Therefore,

$$\sqrt{\frac{\sigma_{y'}^2}{\sigma_y^2}} = \sqrt{\frac{\frac{(\Sigma xy)^2}{\Sigma x^2}}{N} \div \frac{\Sigma y^2}{N}} = \sqrt{\frac{(\Sigma xy)^2}{\Sigma x^2} \times \frac{1}{\Sigma y^2}} = \sqrt{\frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}}.$$

Extracting the root, $\frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} = \frac{\Sigma xy}{\sqrt{N\sigma_x}\sqrt{N\sigma_y}} = \frac{\Sigma xy}{N\sigma_x\sigma_y}$

The above worksheet may be used for small samples, but a shorter and preferable method is available. The only advantage of Worksheet No. 39 is that it shows in detail all the steps used in computing the Pearsonian coefficient. It is like placing the entire operation under a microscope.

In practical statistical computations, the following method should be used for individual items.

WORKSHEET NO. 34

X	Y	XY	X^2	Y^2
12	14	168	144	196
9	8	72	81	64
8	6	48	64	36
10	9	90	100	81
11	11	121	121	121
13	12	156	169	144
7	3	21	49	9
70	63	676	728	651

FORMULA NO. 51

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 728 - \frac{(70)^2}{7} = 728 - \frac{4900}{7} = 728 - 700 = 28$$

FORMULA NO. 52

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 651 - \frac{(63)^2}{7} = 651 - \frac{3969}{7} = 651 - 567 = 84$$

FORMULA NO. 53

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X \cdot \Sigma Y)}{N} = 676 - \frac{70 \times 63}{7} = 676 - \frac{4410}{7} = 676 - 630 = 46$$

FORMULA NO. 54

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{46}{\sqrt{2352}} = \frac{46}{48.5} = .95$$

FORMULA No. 55

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} = \frac{(46)^2}{28 \times 84} = \frac{2116}{2352} = .90$$

Worksheet No. 34 and the formulas associated with it are among the most useful methods in statistics. From it with a minimum of machine work and time may be computed the arithmetic means, standard deviation, the regression line, $Y = a + bX$, the standard error of estimate, $S_y = \sigma_y\sqrt{1 - r^2}$; and the coefficient of correlation,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}.$$

The student should become thoroughly familiar with it and use it whenever he works a sample on the basis of individual items.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

is a better formula than

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

for most cases because it saves one extraction of square root and gives the same result as the Pearsonian form.

FORMULA FOR UNGROUPED DATA

It is not necessary to reduce the data to deviations from the mean in order to compute correlation. The formula may be so altered in form without changing its value that the sums of the original data and their squares and products as given in Worksheet No. 34 may be used.

FORMULA No. 56

$$r = \frac{N\Sigma XY - (\Sigma X \cdot \Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{7 \cdot 676 - 70 \cdot 63}{\sqrt{[7 \cdot 728 - (70)^2][7 \cdot 651 - (63)^2]}}$$

$$= \frac{4732 - 4410}{\sqrt{[5096 - 4900][4557 - 3969]}}$$

$$= \frac{322}{\sqrt{196 \cdot 588}}$$

$$= \frac{322}{\sqrt{115{,}248}}$$

$$= \frac{322}{338} = .95$$

The ungrouped data formula maintains the same relationship throughout that are used in the Pearsonian coefficient formula of

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y},$$

but in the ungrouped data formula these relationships are expressed in the large sums of the original data while in the Pearsonian coefficient they are expressed as deviations from the mean.

1. The advantages of the ungrouped data method are that it (a) saves the time of computing the deviations from the mean and (b) eliminates all or most of the decimal fractions.

2. Its disadvantages are that it (a) runs into large numbers which are difficult to get on the calculating machine and (b) looks long and difficult to the beginners. It is an excellent method for a small sample when the individual items are used.

CORRELATION TABLE

Another device that is very useful in computing the coefficient of correlation and all related measures at the same time is the correlation table. It is of special value if one has a large sample which *must* be thrown into class intervals for economy of manipulation. One can manage a sample of ten thousand or even a hundred thousand items as easily, after the class intervals and tally sheet have been made, as he can thirty or a hundred items.

Several forms of this type of table are available in different texts and statistical journals. The form given below is as useful

and easy to manage as any. The first step is the creation of a cross-classification tally sheet as follows:

WORKSHEET NO. 35

CROSS-CLASSIFICATION TALLY SHEET FOR HEIGHT AND WEIGHT OF 75 GRADE SCHOOL CHILDREN HEIGHT X; WEIGHT Y

Weight Y	Height X						
	40–43	44–47	48–51	52–55	56–59	60–63	Fy
35– 54	1	~~1111~~ ~~1111~~	~~1111~~ ~~1111~~ ~~1111~~				26
55– 74			~~1111~~ 1	~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ 11	111		31
75– 94				11	~~1111~~ 111	11	12
95–114					1	11	3
115–143				1			1
135–154					1	1	2
Fx	1	10	21	25	13	5	75

After the tabulation and totals have been completed in the above cross-classification tally sheet, the class intervals with their appropriate frequencies and totals are set up as part of the Correlation Table shown in Worksheet No. 35, from which may be computed $\bar{X}$, $\bar{Y}$, $\sigma_{\bar{x}}$, $\sigma_{\bar{y}}$, r, r^2, σ_x, σ_y, σr, S_y, and $Y = a + bX$.

Explanation of Worksheet No. 36

The captions of the columns and the stubs of the lines are largely self-explanatory. This worksheet is a device for computing two standard deviations at the same time, one on X, the other on Y. Its elements are similar to those in Worksheet No. 24 for computing the single standard deviation, with the exception that two such worksheets are combined here. One of them is horizontal; the other is perpendicular. The totals in the bottom lines supply the X standard deviation. The totals on the right supply the Y standard deviation. Since the short method with class deviations from an assumed mean is used, it is necessary to

WORKSHEET NO. 36

CORRELATION TABLE

CORRELATION BETWEEN HEIGHT (X) AND WEIGHT (Y) OF 75 STILLWATER SCHOOL CHILDREN OF GRADES 1 TO 6

Y	X 40–43	44–47	48–51	52–55	56–59	60–63	F_y	d_y	fd_y	fd_y^2	$\Sigma d_x F_y$	$fd_x d_y$
35–54	1	10	15				26	− 1	− 26	26	− 38	38
55–74			6	22	3		31	0			− 3	
75–94				2	8	2	12	+ 1	+ 12	12	+ 12	12
95–114					1	2	3	+ 2	+ 6	12	+ 5	10
115–134				1			1	+ 3	+ 3	9	0	0
135–154					1	1	2	+ 4	+ 8	32	+ 3	12
F_x	1	10	21	25	13	5	75		+ 3	91	− 21	72
d_x	− 3	− 2	− 1	0	+ 1	+ 2						
fd_x	− 3	− 20	− 21		+ 13	+ 10	− 21					
fd_x^2	9	40	21		13	20	103					
$\Sigma d_y F_x$	− 1	− 10	− 15	+ 5	+ 14	+ 10	+ 3					
$fd_x d_y$	3	20	15		14	20	72					

	$\Sigma d_x f$	$\Sigma d_y f$	$\Sigma d_x^2 f$	$\Sigma d_y^2 f$	$\Sigma d_x d_y f$
Sums	− 21	+ 3.	103 0	91 0	72
Means of Sums	− .28	.04			
Corrections (subtracted)			5.88	0 12	− 84
Corrected Values			97.12	90 88	72.84

$$\frac{y \text{ class}}{x \text{ class}} = \frac{20}{4} = 5$$

b_{yx} in class intervals

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{72\ 84}{97\ 12} = .75$$

$b_{yx} = .75 \times 5 = 3\ 75$

in terms of original data

$$a = \overline{Y} - b\overline{X}$$
$$= 65\ 8 - 3\ 75 \times 52.88$$
$$= - 132\ 5$$

$$Y = a + bX$$
$$= - 132.5 + 3.75X$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{72\ 84}{\sqrt{97.12 \times 90.88}} = \frac{72\ 84}{93.9} = .776$$

$$\overline{X} = A + \frac{\Sigma f d_x}{N} i = 54 + \frac{-21}{75} 4 = 54 - 1.12 = 52.88$$

$$\overline{Y} = A + \frac{\Sigma f d_y}{N} i = 65 + \frac{3}{75} 20 = 65 + .8 = 65.8$$

$$\sigma_x = i\sqrt{\frac{\Sigma x^2}{N}} = 4\sqrt{\frac{97.12}{75}} = 4 \times 1.138 = 4.54$$

$$\sigma_y = i\sqrt{\frac{\Sigma y^2}{N}} = 20\sqrt{\frac{90.88}{75}} = 20 \times 1.10 = 22.0$$

correct for this error just as in Worksheets No. 30 and No. 31. The corrections are computed in the following form:

	$\Sigma d_x f$	$\Sigma d_y f$	$\Sigma d_x^2 f$	$\Sigma d_y^2 f$	$\Sigma d_x d_y f$
Sums	− 21	+ 3.	103	91	72.
Means	− .28	+ .04			
Corrections (subtracted)			5.88	0.12	− .84
Corrected Values			97.12	90.88	72.84

The sums used above are obtained from the appropriate columns and lines in the worksheet. It will be noted that the sums for $\Sigma d_x f$, $\Sigma d_y f$, $\Sigma d_x d_y f$ appear in both the X and Y totals. This device gives a check on the correctness of these totals. Any discrepancy in these paired sums indicates an error in the computations of the worksheet. These corrected values are used for all further computations.

In computing the correction above, the product of the − 21 times its mean, − .28, is the correction, 5.88, for $\Sigma d_x^2 f$. The product of the + 3 times its mean, + .04, is the correction, 0.12, for $\Sigma d_y^2 f$. The product of the − 21 times the mean, + .04, the sum of $\Sigma d_y f$ times the mean of $\Sigma d_x f$ is the correction of 0.84, for $\Sigma d_x d_y f$.

Since the value for b_{yx} is computed from the corrected totals in class intervals, it must be multiplied by the quotient of the class intervals, as follows:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{72.84}{97.12} = .75$$

$$\text{Class interval correction} = \frac{y \text{ class}}{x \text{ class}} = \frac{20}{4} = 5$$

b_{yx} corrected $= .75 \times 5 = 3.75$ in terms of pounds of weight for each inch of height. All the other computations have been explained in earlier chapters.

Regression Formula for Correlation

FORMULA No. 57 for determination $r^2 = b_{yx} \cdot b_{xy}$

FORMULA No. 58 for correlation

$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

Both of these b values can be computed from Worksheet No. 36. If one for other purposes has already computed the b's, this is the easiest method for computing the coefficient of correlation.

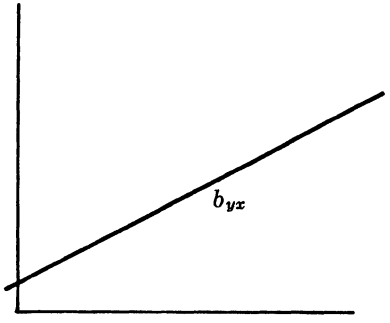


FIG. 59. Normal equations for computation of $Y = a + bX$

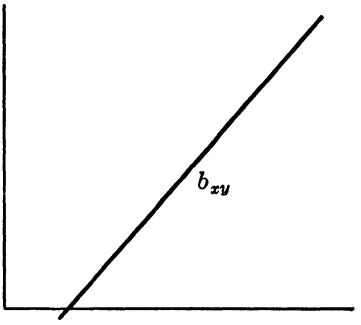


FIG. 60. Normal equations for computation of $X = a + bY$

$$\Sigma Y = Na + b\Sigma X \qquad \Sigma X = Na + b\Sigma Y$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Since $$r = \frac{\Sigma xy}{N\sigma x\sigma y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}},$$

$$r^2 = \frac{\Sigma xy \cdot \Sigma xy}{\Sigma x^2 \cdot \Sigma y^2} = \left(\frac{\Sigma xy}{\Sigma x^2}\right) \cdot \left(\frac{\Sigma xy}{\Sigma y^2}\right) = b_{xy} \cdot b_{xy}$$

Since

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \quad \text{and} \quad b_{xy} = \frac{\Sigma xy}{\Sigma y^2}, \; r^2 = b_{yx} \cdot b_{xy}$$

All of the above formulas for correlation are closely related and may be transformed algebraically one into the other. All of them give the same results. The statistician will ordinarily use the one which is most convenient for the type of calculations he finds desirable to make.

If one wishes only the coefficient of correlation or determination and the sample is not large, not over 50 or 100 items, Formula No. 50 or No. 54 is best.

If the sample is quite large and the regression lines are also desired, the correlation table and associated formulas are best. A thousand or even ten thousand or more items can be manipulated quite easily by this method. It is especially useful in economics, education, sociology, and other social sciences where large samples are often used.

If the standard error of estimate, S_y, has already been computed or is required, Formulas No. 47 and No. 48 are best. In most cases, Formulas No. 45 and No. 46 are less convenient to use, but sometimes may be useful.

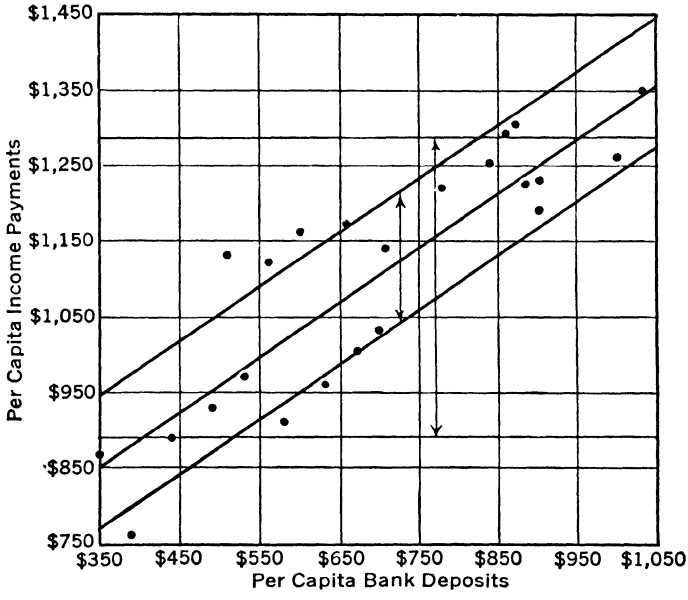


FIG. 61. Graphic method of computing correlation per capita bank deposits (X), per capita income payments (Y), 1948. (*Statistical Abstract of U.S.*)

This method may be useful when: (1) a short-cut is needed; (2) an approximate result is satisfactory, as in determining whether a given relationship justifies mathematical analysis; (3) the basic data are rough; (4) or a graphic teaching device is useful to make relationships clearer.

SHORT-CUT GRAPHIC METHOD [1] OF CORRELATION

By use of a scatter diagram and a freehand regression curve it is possible to obtain a quite accurate measure of the correlation of a sample.

Procedure

1. Plot a large size scatter diagram of the data.
2. Draw in a freehand regression curve.
3. Draw two lines *parallel* to the regression line, one on each side of it so that one-sixth of the dots fall above the top line and one-sixth fall below the bottom line, as in Fig. 61.
4. Measure the vertical width of the distance between these two lines on the y-axis scale. This value is roughly twice the standard error of estimate ($2S_y$).
5. Set the ruler on the scatter diagram and mark off two *horizontal* lines which separate off the top one-sixth and the bottom one-sixth of the items as in Fig. 61.
6. Measure the distance between these two lines on the Y-axis scale.
7. Divide the number obtained in step 4 by the number obtained in step 6. The result is the coefficient of alienation, $\left(\frac{S_y}{\sigma_y}\right)$.
8. Compute r by the formula $r = \sqrt{1 - \left(\frac{S_y}{\sigma_y}\right)^2}$.

TABLE 21

DATA ON PER CAPITA BANK DEPOSITS (X) AND PER CAPITA INCOME (Y) FOR 25 STATES. (SEE WORKSHEET NO. 31.)

X	Y	X	Y	X	Y	X	Y	X	Y
44	89	84	125	103	135	49	93	90	119
66	117	86	129	39	76	70	103	88	123
42	86	58	91	107	136	87	130	90	123
71	114	67	100	100	126	35	86	60	116
53	97	78	122	56	112	63	96	51	113

[1] From "A Short-Cut Measure of Correlation," William A. Spurr, Stanford University, from *Journal* of the *American Statistical Association*, Vol. 46, No. 253, pp. 89–94. Used with permission.

This method is not dependable for low coefficients of correlation. The least squares method which weights the items more accurately is superior for most data.

Computations of r

$$2S_y = \frac{17.3}{39.5} = .438 \qquad r^2 = 1 - \left(\frac{S_y}{\sigma_y}\right)^2 = 1 - .1918 = .8081$$

$$2\sigma_y = 39.5$$

$$80.81\% \qquad r = .90$$

RANK DIFFERENCE METHOD OF CORRELATION

Another type of correlation measure is the *rank difference* methods. They are somewhat cumbersome and are not widely used. The problem below illustrates one of them. The pairs of items are ranked for each variable. The largest item is given a rank of 1, the next largest, 2, the next, 3, and so on to the smallest

WORKSHEET NO. 37

COMPUTATION OF CORRELATION BY RANK DIFFERENCES METHOD FOR PER CAPITA BANK DEPOSITS AND PER CAPITA INCOME FOR STATES, 1948

	Per Capita Bank Deposits Divided by 10 1948	Per Capita Income Divided by 10 1948	Bank Deposits Rank	Income Rank	D (Difference)	D^2
Ala.	44	89	11	11	0	0
Ariz.	66	117	8	6	+ 2	4
Ark.	42	86	12	12	0	0
Fla.	71	114	6	7	− 1	1
Ga.	53	97	10	9	+ 1	1
Ida.	84	125	4	2	+ 2	4
Kan.	86	129	3	1	+ 2	4
Ky.	58	91	9	10	− 1	1
La.	67	100	7	8	− 1	1
Me.	78	122	5	4	+ 1	1
Texas	90	119	1	5	+ 4	16
Utah	88	123	2	3	+ 1	1
						34

item.[1] The difference of the paired ranks is then taken by subtracting the second variable rank from the first variable rank. These D's are squared and summed. If the ranks are all identical for the two variables the differences are all zero (0) and the correlation equals 1. This method is useful for small samples using individual items and when a high degree of accuracy is not required. It is the simplest and shortest of all mathematical methods presented in this text. It is not suited to large samples.

FORMULA NO. 59

$$r = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \times 34}{12(12^2 - 1)} = 1 - \frac{204}{1716} = 1 - .119$$

$$= .881$$

SUMMARY

1. The coefficient of determination is a percentage figure which measures the amount of variation in one variable called the dependent variable which is associated with, or varies with, variation in another variable called the independent variable. The coefficient of determination would measure the percentage of the variation in plant growth in an area that is associated with variation of rainfall in the identical area. Its symbol is r^2.

2. The coefficient of correlation is the square root of the coefficient of determination and is, therefore, not a percentage figure. Its symbol is r.

3. The base of both the coefficient of determination and the coefficient of correlation is unity, or 100%. When the coefficients of correlation and determination are unity (1), the variation in the one factor is identical with the variation in the other factor. When the coefficients are less than unity, part of the variation in the dependent variable may not be accounted for by variation in the independent variable.

4. The basic formula of the coefficient of determination rests on the ratio between the squared standard deviation of the estimated values of

[1] If two or more items have the same rank, both items are given the average of the ranks. If the third and fourth items are the same both are given the rank $3\frac{1}{2}$ and the next item takes it normal rank of 5. If the fifth, sixth and seventh items were equal all three would take the rank of 6, and the next item takes its proper place of 8.

the dependent variable computed from a regression line showing the average relationship between the two variables.

$$r^2 = \frac{\sigma_{y'}^2}{\sigma_y^2}$$

5. The formula of the Pearsonian coefficient of correlation, named for its originator, Karl Pearson, is $r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$.

6. The coefficient of determination is equal to the product of the reversed regression coefficients: $r^2 = b_{yx} \cdot b_{xy}$.

7. The correlation table is a device for computing the correlation between two variables which have been thrown into cross frequency distributions. It is especially useful for large samples.

8. The fact that there is correlation between two variables does not prove that there is a casual relationship between them, but only that their variations are associated.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics,* Chapter XXVII. Irwin, Inc., Chicago, 1941.

Chaddock, R. E., *Principles and Methods of Statistics,* Chapter XII. Houghton Mifflin Company, Boston, 1925.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics,* Chapters XXII, XXIII, XXIV. Prentice-Hall, Inc., New York, 1948.

Davies, G. R., and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences,* Chapter VIII. John Wiley & Sons, Inc., New York, 1933.

Ezekiel, M., *Methods of Correlation Analysis,* Chapters VII, XII, XIV. John Wiley & Sons, Inc., New York, 1941.

Garrett, H. E., *Statistics in Psychology and Education,* Chapters IV, V. Longmans, Green & Company, New York, 1926.

Mills, F. C., *Statistical Methods, Revised,* Chapters X, XII. Henry Holt and Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics,* Chapters IX, X, XII. John Wiley & Sons, Inc., New York, 1942.

Snedecor, G. W., *Statistical Methods,* Chapter 7, Collegiate Press, Inc., Ames, Iowa, 1938.

REVIEW QUESTIONS

1. What is the difference between determination and correlation?
2. In what respects is the coefficient of correlation misleading?
3. What is positive correlation? Negative? Give two illustrations of each.
4. What is the relation of correlation to regression? Why is it important to fit the best possible regression line to a sample?
5. Explain the formula $r^2 = \frac{\sigma_{y'}^2}{\sigma_y^2}$.
6. What relation does the formula in question 5 bear to $r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$?
7. What is the relation of the Pearsonian coefficient, $r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$, to $r^2 = \frac{\sigma_{y'}^2}{\sigma_y^2}$?
8. What are the advantages of Formula No. 55 over Formulas No. 50 and No. 56?
9. What are the advantages and disadvantages of a correlation table?
10. What is the relationship of the standard error of estimate to the coefficient of correlation? Explain fully.
11. Of what use is the coefficient of determination or correlation in scientific studies? Explain fully.
12. What is the difference in the units of measurement of correlation and determination?

EXERCISES

1. Compute the correlation between the heights and weights of children given in Worksheet No. 1.
2. Compute the determination between the age and weights of children given in Worksheet No. 1.
3. Compute the determination between the two variables in each of the five problems in the Exercises Nos. 1 and 2 at the end of Chapter 8.

CHAPTER 10

NON-LINEAR REGRESSION

In Chapter 8 the freehand and mathematical techniques of linear regression were presented. Such methods are adequate for only straight line relationships, which are relatively rare in the economic and social world as well as in the biological and physical universe. Most relationships among variables are *not* straight and can be described and measured accurately only by some curvilinear or non-linear function.

The long-time growth of population, whether for a village or city or a whole nation, is usually a non-linear relationship. The development of new industries, such as the production of automobiles, radios, or television sets, progresses through a non-linear trend. Percentage, logarithmic, exponential, hyperbolic, and logistic functions are widely used with physical, economic, and social data.

Supply and demand schedules often run in curved lines. Utility does not increase or diminish in successive uniform absolute amounts. It declines more rapidly at first and more slowly later. Overhead or fixed costs and even direct and selling costs vary according to curvilinear functions. To double output will reduce fixed costs per unit about 50%, but to treble production will cut overhead costs per unit only 16.7% more. Interest rates, wage scales, raw materials costs and most other business factors vary in curvilinear relations. Because trees and animals are three dimensional they tend to increase in volume and weight at a faster rate than their increase in height. Because of the burden of increased weight due to gravitation, which increases in animals faster than muscular strength, small animals are relatively stronger than large ones. A squirrel in proportion to its weight is stronger than

a horse. In proportion to its volume an ant can do more work than an elephant. The relationships among strength, weight, and volume are *not* straight line. The graphs and formulas with which every textbook on astronomy, chemistry, physics, and geology is filled indicate that the relationships in the physical sciences are curvilinear. It seems necessary, therefore, to include even in an elementary text in statistics a brief analysis of simple curvilinear regression. The methods presented here are simple and do not require a knowledge of mathematics beyond first semester algebra.

CLASS AVERAGES METHOD

The student is already familiar with the method of fitting a straight line trend through two sets of class averages. The only modification of this method necessary to fit a curved line is to set the data up in more than two class intervals, preferably five or more. The essential steps in the technique are as follows:

1. Plot the data on coordinate graph paper.
2. Arrange the data in an array for values of X, ranging from the smallest to the largest items, keeping each Y-value associated with its own X-value in a companion column regardless of the order or size of the Y-values, as indicated in Worksheet No. 38.
3. Obtain the sums of each column.
4. Compute the means for each column.
5. Plot these paired averages on the coordinate graph paper on which the original data were plotted.
6. Draw as smooth a curve as possible through or among these plotted averages. The reason for not always drawing the curve so that it passes exactly through the averages is that the averages may vary erratically because of the smallness of the sample. A larger sample would give more dependable averages. By drawing the curved line between or among the averages in the form of a smooth curve one avoids the influences of chance variations in a small sample.

Two class intervals limit one to a straight-line trend. In many cases such as the one above, the true regression is curvilinear.

WORKSHEET NO. 38

METHOD OF DETERMINING CURVILINEAR REGRESSION THROUGH CLASS AVERAGES FOR YIELD PER ACRE AND COST PER BUSHEL FOR WHEAT ON 25 NORTH DAKOTA FARMS

X 8–9		*X* 10–12		*X* 13–16		*X* 17–21		*X* 22–28		*X* 29–30	
X Yield	*Y* Cost	*X* Yield	*Y* Cost	*X* Yield	*X* Cost	*X* Yield	*Y* Cost	*X* Yield	*Y* Cost	*X* Yield	*Y* Cost
8	$2.00	10	$1.56	14	$ 93	17	$.86	22	$.54	29	$ 60
9	1.84	11	1 29	14	.89	18	.75	24	.50	30	.60
		12	1 08	15	.98	19	67	26	.48	31	.67
				16	.86	20	.67	27	.58	32	.62
						21	73	28	.49	33	.70
										34	.64
Sum 17	$3 84	33	$3.93	59	$3.66	95	$3 68	127	$2 59	189	$3 83
No. 2	2	3	3	4	4	5	5	5	5	6	6
Average 8.5	1 92	11 0	1 31	14.7	91	19 00	.733	25.4	.52	31 5	.638

To fit such a curve, more than two class intervals are necessary. The statistician must be guided by his best judgment in such cases, but may make as many classes as he thinks the form of the data requires. The class intervals need not be of exactly equal width. On the graph it will appear often that there are open places or definite breaks in the data. It is better to place the ends of the class intervals at these open spots, or points of definite break in the data. If such breaks do not occur, the class intervals may be made equal or unequal arbitrarily as seems most likely to result in the best regression line. The best line is the one that splits the plotted points approximately equally from one side of the graph to the other. Where the trend of the data changes direction rapidly, it is better to have more and narrower class intervals in order to indicate more definitely the correct variation in the line.

The six pairs of class averages from Worksheet No. 38 are located on Fig. 62 at the points of the small hollow circles. A smooth curve is then drawn through or among these averages. The curve is in the general form of a simple parabola.

It will be found that it is very easy and economical to fit curvilinear regression lines by this method. It not only requires very

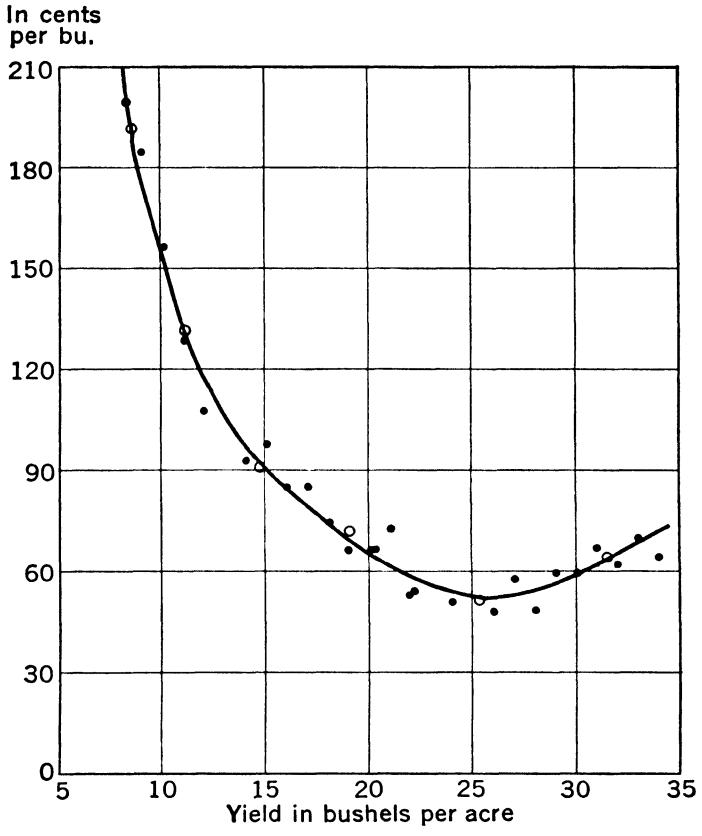


FIG. 62. Data and class averages for wheat yields and costs per bushel showing freehand curve plotted to class averages. See Worksheet No. 38.

little time and very few computations, but if the sample is of fair size, the line will fit remarkably well. It cannot be expressed in a formula and manipulated algebraically. With these limitations, it is still a very useful device. Such a trend line will usually give the statistician a much clearer idea of the type of mathematical line he should select for the data than he would have if he did not fit the trend line through class averages. Mathematical lines are so time consuming and expensive to fit that for some purposes this line is preferable.

STANDARD ERROR OF ESTIMATE AND CORRELATION COEFFICIENT FOR FREE-HAND REGRESSION LINES

Since the *free-hand* method has no regression equation, the usual mathematical computation of a standard error of estimate and coefficient of correlation is not possible. An approximation of these computations can be obtained by a graphic method of *reading off the deviations* from the plotted points, which represent the data, to the free-hand regression line. The *read-off* deviations are comparable in meaning to the mathematically computed deviations from a mathematically located regression line. The *read-off*

WORKSHEET NO. 39

COMPUTATION OF FREE-HAND CURVE STANDARD ERROR OF ESTIMATE AND COEFFICIENT OF CORRELATION FOR 25 NORTH DAKOTA FARMS

Independent Variable X	Read off Deviations Z	Z^2	Independent Variable X	Read off Deviations Z	Z^2
8	− 5	25	21	+ 10	100
9	+ 5	25	22	− 6	36
10	+ 3	9	24	− 5	25
11	− 3	9	26	− 5	25
12	− 9	81	27	+ 5	25
14	− 2	4	28	− 6	36
14	− 6	36	29	+ 2	4
15	+ 8	64	30	0	0
16	+ 2	4	31	+ 5	25
17	+ 7	49	32	− 3	9
18	0	0	33	+ 2	4
19	− 5	25	34	− 7	49
20	− 1	1			
					670

$$S_y^2 = \frac{\Sigma Z^2}{N} = \frac{670}{25} = 26.4 \qquad \sigma_y^2 = \frac{39{,}992.36^1}{25}$$

$$S_y = \sqrt{26.4} = 5.14 \qquad = 1{,}599.69$$

$$\rho^2 = 1 - \frac{S_y^2}{\sigma_y^2} = 1 - \frac{26.4}{1{,}599.69} \qquad \sigma_y = 40.0$$

$$= 1 - .0165 = .9835 \qquad \rho = .99$$

[1] Use Worksheet 34 and Formula 52 for computation of σ_y^2.

deviations on the graph bear the same relationships to the free-hand line that the computed deviations bear to the computed line. They measure in both cases the scatter of the data around the regression line. They both measure the extent to which the regression line fails to account fully for the exact size and location of the given data. They indicate the extent to which factors other than the independent, or X-variable are influencing the measurements of data for the dependent variable or on the Y-axis.[1]

One measure of S_y computed above when measured plus and minus from the free-hand regression line includes 17 of the 25 items in the sample, or 64 percent of them. In this it also conforms closely to the meaning of a mathematical standard error of estimate measured from a mathematical regression line. If the sample were larger the chances would be greater that one S_y plus and minus would include 68.26 percent of the sample items.

The free-hand method is especially well suited to the analysis of large samples which are long and tedious for complex mathematical analysis. The free-hand method enables the statistician to experiment quickly and cheaply with many large samples from a population as a means of deciding on a permanent mathematical function. It is an excellent descriptive device.

Curvilinear Correlation

Rho, ρ, is the symbol used to indicate the measure of the curvilinear coefficient of correlation. r, which was employed in earlier chapters, is used to indicate only straight line correlation. To distinguish between the curvilinear and the linear relationship ρ is employed. It is usually computed by subtracting the fraction, $\left(\frac{S_y^2}{\sigma_y^2}\right)$ from 1, as is indicated in

FORMULA NO. 60

$$\rho^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$ Curvilinear coefficient of determination

[1] The deviation can be measured by the device of a thin piece of cardboard or graph paper marked with 0 at a midpoint and the scale of the Y-axis of the graph marked on the cardboard from zero (0) up and down. If this measuring device is held vertically, parallel to the Y-axis, with 0 on the free-hand curve, the deviation of the points from the curve can be quickly read and recorded.

and FORMULA NO. 61

$$\rho = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$ Curvilinear coefficient of correlation

Since the fraction $\left(\frac{S_y^2}{\sigma_y^2}\right)$ is the measure of the amount of variation in Y which is *not* associated with variation in X and since unity, (1), is the measure of perfect or complete correlation, the subtraction of the amount of variation in Y which is *not* associated with variation in X gives the measure of the amount of variation in Y that *is* associated with X. These formulas may be used with either free-hand curves or mathematical regression lines.

Relation Between Regression and Coefficient of Correlation

The concepts of correlation and regression are closely related. The size of the coefficient of correlation for any given set of data depends to a considerable extent on the degree to which the underlying regression equation and line measure the true relation between the variables. If the relation between the variables is truly a straight line, a larger coefficient of correlation will be obtained if a straight regression line is used. If, however, the relation between the variables is actually curvilinear, a larger coefficient of correlation will be obtained if the correct curvilinear regression equation is chosen. It is impossible to get a coefficient of correlation which will measure the full amount of relationship in the data unless the regression equation which best measures that relationship is chosen.

Least Squares Method of Computing Simple Parabola

The equation for a simple parabola is $Y = a + bX + cX^2$. In the actual results obtained for curvilinear data the algebraic signs of b and of c are opposites. If the line rises to a maximum and then falls, there will be a $+ b$ and a $- c$. If on the other hand, the line falls to a minimum and then rises, there will be a $- b$ and a $+ c$. The a measures the Y-intercept. The b measures the

amount and direction of slope where $X = 0$. The c measures the change in slope per unit of X and its sign indicates the direction of the slope. Since b and c have opposite signs and since b is weighted by X while c is weighted by X^2, as the values of X increase the product of cX^2 tends to exceed the product bX, and gradually to turn the line in the opposite direction.

The necessity of computing three unknowns, a, b, and c, requires three normal equations which are obtained by multiplying successively the original equation $Y = a + bX + cX^2$ through for each pair of data by the coefficients of a, b, and c.

FORMULA NO. 62

The Normal Equation becomes

$$(1) \quad \Sigma Y = Na + b\Sigma X + c\Sigma X^2$$
$$(2) \quad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$
$$(3) \quad \Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

To compute a parabola from these equations the *sums* or totals of the data must be used. This has the disadvantage that for a large sample the numbers become long and difficult to get on the calculating machine. A great saving of time can be made by coding X^2 as U, or as a third variable, and reducing the equations to deviations from the means of the variables. With this transformation a is computed separately by the equation

$$a = \bar{Y} - b\bar{X} - c\bar{U} \text{ — Formula No. 63}$$

and b and c are computed from the equations

$$\left.\begin{array}{ll} (1) & (\Sigma x^2)b + (\Sigma xu)c = \Sigma xy \\ (2) & (\Sigma xu)b + (\Sigma u^2)c = \Sigma uy \end{array}\right\} \text{ Formula No. 64}$$

in which y, x, and u replace the Y, X, and U of the original normal equation. Each equation is shortened by one term and the number reduced from three equations to two. This is an economy well worth while. The use of the corrections and the deviations from the means computed at the bottom of Worksheet No. 40 further reduces the work required to fit a parabola. The short solution illustrated on page 249 minimizes the task.

WORKSHEET NO. 40

COMPUTATION OF SIMPLE PARABOLA FOR COST PER BUSHEL AND YIELD PER ACRE OF WHEAT FOR 25 NORTH DAKOTA FARMS

Yield of Wheat in Bushels X	Cost per Bushel in Cents Y	U X^2	XU X^3	U^2 X^4	XY	UY X^2Y
9	184	81	729	6,561	1,656	14,904
8	200	64	512	4,096	1,600	12,800
10	156	100	1,000	10,000	1,560	15,600
11	129	121	1,331	14,641	1,419	15,609
33	70	1,089	35,937	1,185,921	2,310	76,230
20	67	400	8,000	160,000	1,340	26,800
34	64	1,156	39,304	1,336,336	2,176	73,984
31	67	961	29,791	923,521	2,077	64,387
14	93	196	2,744	38,416	1,302	18,228
18	75	324	5,832	104,976	1,350	24,300
15	98	225	3,375	50,625	1,470	22,050
21	73	441	9,261	194,481	1,533	32,193
16	86	256	4,096	65,536	1,376	22,016
12	108	144	1,728	20,736	1,296	15,552
14	89	196	2,744	38,416	1,246	17,444
19	67	361	6,859	130,321	1,273	24,187
27	58	729	19,683	531,441	1,566	42,282
29	60	841	24,389	707,281	1,740	50,460
32	62	1,024	32,768	1,048,576	1,984	63,488
17	86	289	4,913	83,571	1,462	24,854
24	50	576	13,824	331,776	1,200	28,800
22	54	484	10,648	234,256	1,188	26,136
26	48	676	17,576	456,976	1,248	32,448
28	49	784	21,952	614,656	1,372	38,416
30	60	900	27,000	810,000	1,800	54,000
Sums 520	2,153	12,418	325,996	9,103,116	38,544	837,168
Means 20.8	86.12	496.72				
Corrections		10,816	258,294	6,168,269	44,782	1,069,438
Deviations		1,602	67,702	2,934,845	−6,238	−232,270

Detailed Computation of Corrections

$$\Sigma x^2 = \Sigma X^2 - N\bar{X}^2$$
$$= 12{,}418 - 25(20.8)^2$$
$$= 12{,}418 - 10{,}816$$
$$= 1{,}602$$

$$\begin{aligned}
\Sigma xu &= \Sigma XU - N\bar{X}\bar{U} \\
&= 325{,}996 - 25(20.8)(496.72) \\
&= 325{,}996 - 258{,}294 \\
&= 67{,}702 \\
\Sigma u^2 &= \Sigma U^2 - N\bar{U}^2 \\
&= 9{,}103{,}116 - 25(496.72)^2 \\
&= 9{,}103{,}116 - 6{,}168{,}269 \\
&= 2{,}934{,}845 \\
\Sigma xy &= \Sigma XY - N\bar{X}\bar{Y} \\
&= 38{,}544 - 25(20.8)(86.12) \\
&= 38{,}544 - 44{,}782 \\
&= -\ 6{,}238 \\
\Sigma uy &= \Sigma UY - N\bar{U}\bar{Y} \\
&= 837{,}168 - 25(496.72)(86.12) \\
&= 837{,}168 - 1{,}069{,}438 \\
&= -\ 232{,}270
\end{aligned}$$

The student should check these detailed computations back against the Sums, Means, Corrections, and Deviations at the bottom of Worksheet No. 40.

Short Method of Computing Deviations

$$\begin{aligned}
\Sigma x^2 &= \Sigma X^2 - (\Sigma X)(\bar{X}) = 12{,}418 - (520)(20.8) \\
&= 12{,}418 - 10{,}816 \ = 1{,}602
\end{aligned}$$

One multiplication and one subtraction are all the computations required by the short method as compared with two multiplications and a subtraction in the detailed method. The detailed method makes every step complete and clear. The short method saves time and should always be used after the student understands the relations involved. The short method is based on the fact that $\Sigma X = N\bar{X}$, $\Sigma Y = N\bar{Y}$, $\Sigma U = N\bar{U}$ and, therefore $(\Sigma X)\bar{X} = N\bar{X}^2$, $(\Sigma X)\bar{Y}$, or $(\Sigma Y)\bar{X} = N\bar{X}\bar{Y}$, $(\Sigma U)\bar{X} = N\bar{U}\bar{X}$, and $(\Sigma U)\bar{U} = N\bar{U}^2$.

$$\begin{aligned}
\Sigma xu &= \Sigma XU - (\Sigma X)(\bar{U}) = 325{,}996 - (520)(496.72) \\
&= 325{,}996 - 258{,}294 = 67{,}702
\end{aligned}$$

$$\Sigma u^2 = \Sigma U^2 - (\Sigma U)(\bar{U}) = 9{,}103{,}116 - (12{,}418)(496.72)$$
$$= 9{,}103{,}116 - 6\ 168{,}269 = 2{,}934{,}845$$
$$\Sigma xy = \Sigma XY - (\Sigma X)(\bar{Y}) = 38{,}544 - (520)(86.12)$$
$$= 38{,}544 - 44{,}782 = -\ 6{,}238$$
$$\Sigma uy = \Sigma UY - (\Sigma U)(\bar{Y}) = 837{,}168 - (12{,}418)(86.12)$$
$$= 837{,}168 - 1{,}069{,}438 = -\ 232{,}270$$

Five multiplications and five subtractions are sufficient to compute the deviations for the equations of a parabola after the sums of the original data are obtained and the three means computed. On the machines the student can soon perform the operations quickly, which takes the drudgery out of computing parabolas.

SOLUTION OF PARABOLA

Equations in terms of deviations from the means

$$(\Sigma x^2)b + (\Sigma xu)c = \Sigma xy$$
$$(\Sigma xu)b + (\Sigma u^2)c = \Sigma uy$$

Substitutions from Worksheet No. 40.

(1) $1{,}602\,b + 67{,}702\,c = -\ 6{,}238$
(2) $67{,}702\,b + 2{,}934{,}845\,c = -\ 232{,}270$

Divide each equation through by the coefficient of b in that equation

(1) $b + 42.2609c = -\ 3.89388$
(2) $b + 43.3494c = -\ 3.43077$

$0 \quad -\ 1.0885c = -\ .46311$

$c = .425457$

Subtract (2) from (1)

Substitute arithmetic value for c in equation (1)

$$b + 42.2609\ (.425457) = -\ 3.89388$$
$$b + 17.980196 = -\ 3.89388$$
$$b = (-\ 17.980196) + (-\ 3.89388)$$
$$b = -\ 21.874076$$

Compute $a = \overline{Y} - b\overline{X} - c\overline{U}$

$$a = 86.12 - (-21.87408)(20.8) - (.425457)(496.72)$$
$$a = 86.12 + 454.98 - 211.33$$
$$a = 330.87$$

Regression Line for Simple Parabola

$$Y = a + bX + cX^2$$
$$Y = 330.87 - 21.87408X + .42546X^2$$

Standard Error of Estimate for Simple Parabola

How closely does the simple parabola regression line fit these North Dakota wheat yield-cost data? With what degree of accuracy could the cost per bushel be computed from the simple parabola regression equation? The standard error of estimate is the answer.

The computation of the standard error of estimate by individual items requires the substitution of the X- and the X^2-values of the original data in the equation $Y = a + bX + cX^2$, for every pair of the original data. In this case each pair of X and X^2 is substituted in the equation $Y = 330.87 - 21.8741X + .4255X^2$ as is indicated in Worksheet No. 41.

The correlation coefficient based on the parabola is .97 which is much larger than that which is based on a straight line.

$$\rho^2 = 1 - \frac{S_y^2}{\sigma_y^2} = 1 - \frac{94.7444}{1{,}599.69} = 1 - .0592 = .9408$$
$$\rho = .97$$

In comparison with this result

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2} = 1 - \frac{625.}{1{,}599.69} = 1 - .39 = .61$$
$$r = .79$$

This small S_y of 9.74 for the parabola may be compared with the much larger S_y of 25.02 for the straight regression line computed for these same data in Chapter 8. This reduction of 61.0% in the size of the standard error of estimate indicates that the parabola fits the sample much better than does the straight

WORKSHEET NO. 41

COMPUTATION OF STANDARD ERROR OF ESTIMATE FOR COST OF WHEAT PER BUSHEL AND YIELD PER ACRE FOR 25 NORTH DAKOTA FARMS BASED ON SIMPLE PARABOLIC REGRESSION LINE

(*a*) plus	$-bX$		$+CX^2$		$=Y'$	Y	$(Y-Y')$	Z^2
	$(-b)$ times	X	$(+c)$ times	X^2			Z	
330.87	− 21.8741	9	+ .4255	81	168.5	184	+ 15.5	240.25
"	"	8	"	64	183.1	200	+ 16.9	285.61
"	"	10	"	100	154.7	156	+ 1.3	1.69
"	"	11	"	121	141.7	129	− 12.7	161.29
"	"	33	"	1,089	72.4	70	− 2.4	5.76
"	"	20	"	400	63.6	67	+ 3.4	11.56
"	"	34	"	1,156	79.0	64	− 15.0	225.00
"	"	31	"	961	61.7	67	+ 5.3	28.09
"	"	14	"	196	108.0	93	− 15.0	225.00
"	"	18	"	324	75.0	75	0.0	0.00
"	"	15	"	225	98.5	98	− 0.5	0.25
"	"	21	"	441	59.2	73	+ 13.8	190.44
"	"	16	"	256	89.8	86	− 3.8	14.44
"	"	12	"	144	129.7	108	− 21.7	470.89
"	"	14	"	176	108.0	89	− 19.0	361.00
"	"	19	"	361	68.9	67	− 1.9	3.61
"	"	27	"	729	50.5	58	+ 7.5	56.25
"	"	29	"	841	54.4	60	+ 5.6	31.36
"	"	32	"	1,024	66.6	62	− 4.6	21.16
"	"	17	"	289	82.0	86	+ 4.0	16.00
"	"	24	"	576	51.0	50	− 1.0	1.00
"	"	22	"	484	55.6	54	− 1.6	2.56
"	"	26	"	676	48.8	48	− 0.8	0.64
"	"	28	"	784	52.0	49	− 3.0	9.00
"	"	30	"	900	57.6	60	+ 2.4	5.76
								2,368.61

$$S_y^2 = \frac{2{,}368.61}{25} = 94.7444$$

$$S_y = \sqrt{94.7444} = 9.74$$

regression line. The accuracy with which cost of production per bushel can be estimated from the parabola is much greater than the estimate from the straight-line trend.

The free-hand curve through class averages as computed in Worksheet No. 38 and shown in Fig. 62 fits these data still better. The comparable standard error of estimate based on the *read-off* deviations from the free-hand line is only 5.14 which is little more than one-half of that based on the parabola. For these particular data the parabola is too extreme a curve. It falls too low at the minimum and rises too rapidly as it approaches the maximum values of the data. There seems to be no sound theoretical ground for thinking that the cost per bushel of producing wheat as related to wheat yields is a true parabolic function. The parabola represents the relationship fairly well but not perfectly. Since the function is not truly parabolic, the parabola becomes only a descriptive device less accurate in this case than the free-hand curve through class averages.

Cubic Parabola

Since the simple parabola does not measure accurately the relationship between cost per bushel of producing wheat as related to yields, it would be well to try the cubic parabola, which is less extreme at those points where the simple parabola missed the data farthest. The cubic parabola is a double reverse curve in the general form of an irregular letter "S," lying on its side "ꝏ."

The equation for a cubic parabola is $Y = a + bX + cX^2 + dX^3$.

The four normal equations necessary to compute the four constants a, b, c, and d are:

FORMULA NO. 65

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 \quad (1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 \quad (2)$$

$$\Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 \quad (3)$$

$$\Sigma X^3Y = a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 \quad (4)$$

When reduced to deviations from the means with X^3 coded as V in the formula, the equation becomes

$$a = \overline{Y} - b\overline{X} - c\overline{U} - d\overline{V} \text{ — Formula No. 66}$$

WORKSHEET NO. 42

ADDITIONAL COLUMNS REQUIRED TO COMPUTE CUBIC PARABOLA FOR COST PER BUSHEL OF WHEAT AND WHEAT YIELDS, 25 NORTH DAKOTA FARMS

U X^2	V X^3	UV X^5	V^2 X^6	VY X^3Y
		59,049	531,441	134,136
		32,768	262,144	102,400
		100,000	1,000,000	156,000
		161,051	1,771,561	171,699
		39,135,393	1,291,467,969	2,515,590
		3,200,000	64,000,000	536,000
		45,435,424	1,544,804,416	2,515,456
		28,629,151	887,503,681	1,995,997
		537,824	7,529,536	255,192
		1,889,568	34,012,224	437,400
		759,375	11,390,625	330,750
		4,084,101	85,766,121	676,053
		1,048,576	16,777,216	352,256
		248,832	2,985,984	186,624
		537,824	7,529,536	244,216
		2,476,099	47,045,881	459,553
		14,348,907	387,420,489	1,141,614
		20,511,149	594,823,321	1,463,340
		33,554,432	1,073,741,824	2,031,616
		1,419,857	24,137,569	422,518
		7,962,624	191,102,976	691,200
		5,153,632	113,379,904	574,992
		11,881,376	308,915,776	843,648
		17,210,368	481,890,304	1,075,648
		24,300,000	729,000,000	1,620,000
Sums 12,418	325,996	264,677,370	7,908,790,498	20,933,898
Means 496.72	13,039.84			
Corrections		161,928,733	4,250,935,680	28,074,775
Deviations		102,748,637	3,657,854,818	−7,140,877

and

$$\left.\begin{aligned}(\Sigma x^2)b + (\Sigma xu)c + (\Sigma xv)d &= \Sigma xy\\ (\Sigma xu)b + (\Sigma u^2)c + (\Sigma uv)d &= \Sigma uy\\ (\Sigma xv)b + (\Sigma uv)c + (\Sigma v^2)d &= \Sigma vy\end{aligned}\right\} \text{Formula No. 67}$$

In the actual solution of the problem for curvilinear data the signs of b, c, and d are alternately opposite. If b is plus, c is minus, and d is plus. If b is minus, c is plus, and d minus.

By coding X^2 as U and X^3 as V and computing the deviations from the means the labor of fitting a cubic parabola is reduced to a minimum. It is necessary to add only three columns of figures to Worksheet No. 40 in order to compute the cubic parabola. These columns are X^5, X^6, and X^3Y.

Arithmetic Quantities Required for Cubic Parabola

$$\begin{aligned}\Sigma x^2 &= 1{,}602\\ \Sigma xu &= 67{,}702\\ \Sigma u^2 &= 2{,}934{,}845\\ \Sigma xv &= 2{,}322{,}349\ ^1\\ \Sigma uv &= 102{,}748{,}637\\ \Sigma v^2 &= 3{,}657{,}854{,}818\\ \Sigma xy &= -\ 6{,}238\\ \Sigma uy &= -\ 232{,}270\\ \Sigma vy &= -\ 7{,}140{,}877\end{aligned}$$

NORMAL EQUATIONS AS DEVIATIONS FROM MEANS

(1) $(\Sigma x^2)b + (\Sigma xu)c + (\Sigma xv)d = \Sigma xy$

(2) $(\Sigma xu)b + (\Sigma u^2)c + (\Sigma uv)d = \Sigma uy$

(3) $(\Sigma xv)b + (\Sigma uv)c + (\Sigma v^2)d = \Sigma vy$

The student should note certain uniformities about these equations.

[1] Σxv is the only quantity which does not appear in Worksheets No. 40 and No. 42. It is computed as follows:

$$\begin{aligned}\Sigma xv &= \Sigma XV - (\Sigma X)(\overline{V})\\ &= 9{,}103{,}116 - (520)(13{,}039.84)\\ &= 9{,}103{,}116 - 6{,}780{,}716 = 2{,}322{,}349\end{aligned}$$

1. The coefficient, (Σxu), of the first term in the second equation is the same as the coefficient, (Σxu), of the second term in the first equation.

2. The coefficient, (Σxv), of the first term in the third equation is the same as the coefficient (Σxv) of the third term of the first equation.

3. The coefficient, (Σuv), of the second term in the third equation is the same as the coefficient (Σuv) of the third term of the second equation.

4. The sums of the squares, Σx^2, Σu^2, and Σv^2, of the three variables x, u, and v run diagonally through the three equations from the first term of the first equation to the third term of the third equation.

5. These relationships hold for these least squares equations because of the methods by which they were obtained — that of multiplying the original equation for the function desired successively by the coefficients of the unknowns desired, whether it be for a, b, c, d, e . . . or any number of them. An understanding of these relationships should aid the student in remembering these equations and in understanding their functions and meanings.

DOOLITTLE METHOD OF SOLVING EQUATIONS

When three or more simultaneous equations are to be solved, the ordinary arithmetic methods become long and cumbersome. The Doolittle Method systematizes the successive steps in a briefer process which is less likely to lead to errors. All students should learn it as an efficient economical statistical aid. It may seem difficult at first but, if many or long problems are to be solved, the time spent in learning it is a great economy.

For the solution of the equations in the Wheat Cost-Yield Problem the steps are:

1. Substitution of numerical values in terms of deviations from the means in the equation

(1) $1{,}602b + 67{,}702c + 2{,}322{,}349d = -6{,}238$

(2) $67{,}702b + 2{,}934{,}797c + 102{,}748{,}637d = -232{,}270$

(3) $2{,}322{,}349b + 102{,}748{,}637c + 3{,}657{,}854{,}818d = -7{,}140{,}877$

2. Solution of Normal Equation by Doolittle Method

WORKSHEET NO. 43

Forward Solution

(1)	1,602b +	67,702c +	2,322,349d =	− 6,238
(1′)	− b −	42.260924c −	1,449.656055d =	+ 3.89419
(2)	67,702b +	2,934,797c +	102,748,637d =	− 232,270
	− 67,702b −	2,861,149c −	98,144,614d =	+ 263,624
(Σ1, 2)		73,648c +	4,604,023d =	+ 31,354
(2′)		− c −	62.5138904d =	− .4296926
(3)	2,322,349b +	102,748,637c +	3,657,854,818d =	− 7,140,877
	− 2,322,349b −	98,144,614c −	3,366,607,290d =	+ 9,042,954
	−	4,604,023c	287,815,389d =	− 1,960,060
(Σ1.2.3.)			3,432,139d =	− 57,983
(3′)			− d =	+ .0169

Back Solution

b	c	d
(1) − 3.894190	+ .4296926	− .0169
(2) + 24.499187	+ 1.0564847	− .0169
(3) − 62.807226	1.4861773	
(4) − 42.202229		

Detailed Explanation of Doolittle Solution

Forward Solution

1. Bring down equation (1).

2. Divide equation (1) through by coefficient of b, which is 1,602 in this case, writing the quotients in appropriate columns in line below (1′) with signs changed. This is the first derived equation, (1′).

3. Bring down second equation (2).

4. Multiply equation (1) through by the coefficient of the second term, (c), in the first derived equation (1′), which is − 42.2609238 in this case, and write the products in the appropriate columns in the line immediately below equation (2).

5. Draw a line under this line and add. This result is the sum of equations (1) and (2) marked ($\Sigma 1, 2$), which eliminates the coefficient of b.

6. Divide the summary equation ($\Sigma 1, 2$) through by the coefficient of its first term, (c), which is 73,648 in this case, and write the quotients in the appropriate columns in the line below, with the signs changed.

7. Bring down equation (3).

8. Multiply equation (1) through by the coefficient of d in equation (1′), which is − 1,449.656055, in this case, and write the products in the appropriate columns in the first line below equation (3).

9. Multiply equation ($\Sigma 1, 2$) through by the coefficient of d in the second derived equation, (2′), which is 62.5138904, in this case, and write the products in the appropriate columns in the second line below equation (3).

10. Add equation (3) to the two lines below it. This result is summary equation ($\Sigma 1.2.3$).

11. Divide summary equation ($\Sigma 1.2.3$) by the coefficient of d, which is 3,432,139 in this case, and write the results in the line below, with the signs changed.

Backward Solution or Substitution of Known Values to Find Other Values

1. Write down in line (1) in the appropriate column the *last* value in each derived equation, with the sign changed. In this case these values are, under d, − .0169, under c, + .4296926, and under b, − 3.894190.

2. Multiply the next to the last value in the first and second derived equations by the known value of d. The computations in this case, are $(- 62.5138904) \times (- .0169) = + 1.056847$, and $(- 1{,}449.656055) \times (- .0169) = + 24.449187$. Write these two products in line (2) under "Back Solution."

3. Add the two values in the c column for the value c. In this case the computation is

$$(+ .4296926) + (+ 1.0564847) = + 1.4861773.$$

4. Multiply the third from the last value in the first derived equation (1′), which is − 42.4609238 in this case, by the value of c. The computation is

$$(-\ 42.2609238) \times (+\ 1.4861773) = -\ 62.807226.$$

Write this product in line (3) under "Back Solution."

5. Sum the values under b in lines (1), (2), and (3). The total is − 42.20229 in this case.

The back solution is merely a systematic concise method of substituting known values for unknowns until all the unknowns are computed. First, c is obtained by substituting the known value for d. Then b is obtained by substituting the known value of d and c.

Computation of *a*

$$\begin{aligned} a &= \bar{Y} - b\bar{X} - c\bar{U} - d\bar{V} \\ a &= 86.12 - (-\ 42.20223)(20.8) - (1.48618)(496.72) \\ &\quad - (-\ .0169)(13{,}039.84) \\ &= 86.12 + 874.4861 - 736.3377 + 220.3733 \\ &= 444.6417 \end{aligned}$$

Regression Equation

$$\begin{aligned} Y &= a + bX + cX^2 + dX^3 \\ Y &= 444.6417 - 42.2022X + 1.4862X^2 - .0169X^3 \end{aligned}$$

STANDARD ERROR OF ESTIMATE FOR CUBIC PARABOLA

By substituting successively all the quantities for X, X^2, and X^3 of the original data in the regression equation above the estimated values for Y may be computed. Here $(Y - Y') = z$.

$$S_y^2 = \frac{\Sigma z^2}{N}$$

For the cubic parabola fitted to the wheat cost-yield data

$$S_y = 7.15$$

TABLE 22

COMPARISON OF STANDARD ERRORS OF ESTIMATE FOR FOUR REGRESSION LINES

	Measure		S_y
1	Straight line, least squares	=	25.02
2	Simple parabola, least squares	=	9.74
3	Cubic parabola, least squares	=	7.15
4	Free-hand line through class average	=	5.14

The free-hand line through class averages "fits" these wheat cost-yield data more closely than any of the mathematically computed lines.[1] This does not mean that some mathematical line might not be fitted which would represent the data better than the free-hand curve. Such a line could be obtained but its form and function are not immediately evident. A considerable amount of study and experimentation would no doubt be necessary to find it.

SIGNIFICANCE OF REGRESSION LINE

If it were evident from logical, known relations and careful experimentation that some particular mathematical function accurately represented and measured the cost-yield relationship, it would be best to use such a line to express the relationship, even though the data in any particular sample failed to conform to such a line. If a mathematical function accurately and permanently measures the relationship between two or more variables, the relationship should be expressed in that function and equation. The law of falling bodies is an example. The distance a body falls, (d), is related to the time in which it falls, (t), as follows:

$$d = 16t^2$$

In 10 seconds it would fall 1,600 feet, or

$$d = 16 \times 10^2 = 16 \times 100 = 1{,}600$$

[1] See Ezekiel, M., *Methods of Correlation Analysis*, pp. 105–112 for further discussion of value of free-hand curves and the methods of determining their relative fit to the data as compared with mathematical curves.

In 30 seconds it would fall 14,400 feet, assuming no air resistance, or

$$d = 16 \times 30^2 = 16 \times 900 = 14{,}400 \text{ feet.}$$

The physical sciences are replete with such relationships, but the social sciences deal with such complex relationships involving so many variables that in most cases as yet no clearly recognized mathematical functions have been discovered which logically and permanently fit social and economic relationships. It is, perhaps,

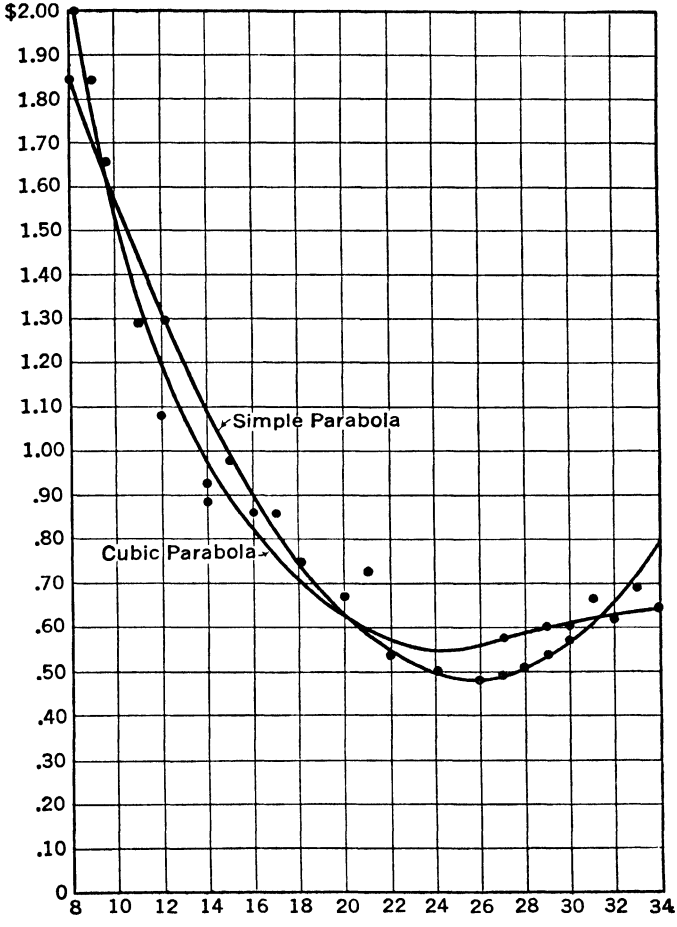


FIG. 63. Comparison of regression lines for wheat yields on 25 farms, simple parabola and cubic parabola

not too much to hope that in the future more study and experimentation may reduce some of these relationships to definite functions. Until such results are known a mathematical curve to such data is scarcely more than a descriptive device.

Logarithmic Regression Curves

Regression lines with three or four constants computed on the arithmetic scale become quite long and complex. It is frequently possible to obtain as good or better fit by the use of logarithms. Four types of such curves are presented in this section.

I. The Exponential Function

The first step in such a computation is to plot the data on semi-logarithmic paper with the arithmetic variable on the X-axis and the geometric variable on the Y-axis. If time is one of the factors it is always plotted on the X-axis. If the data on the semi-log scale fall in approximately in a straight line, as the population data of California from 1890 to 1950 do in Fig. 64, one may obtain a satisfactory regression line by the exponential curve, $Y = ab^x$. The logarithmic form of this curve is:

FORMULA NO. 68

$$\log Y = \log a + (\log b)X$$

For such a line when the point of origin for X is set in the center of the scale so that $\Sigma X = 0$, the normal equations are:

$$\log a = \frac{\Sigma \log Y}{N}; \quad \log b = \frac{\Sigma X (\log Y)}{\Sigma x^2}$$

If the values of X are not a time series the exponential curve cannot be solved by the short method used in Worksheet No. 44, but must be computed from the solution of the normal equations

$$\Sigma(\log Y) = Na + b\Sigma X$$
$$\Sigma(X \cdot \log Y) = a\Sigma X + b\Sigma X^2$$

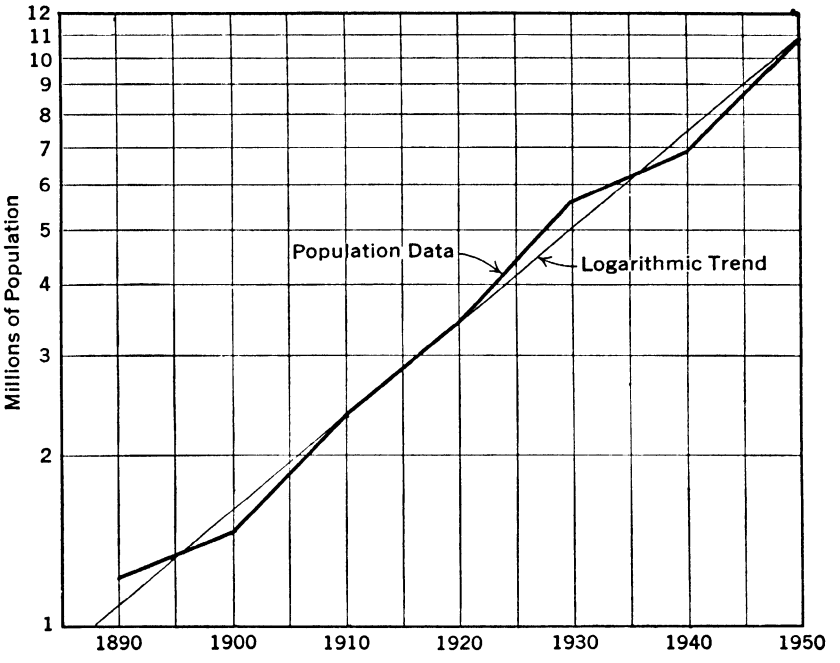


FIG. 64. Logarithmic trend of growth of population of California, 1890–1950. (*U.S. Decennial Census*, 1950)

WORKSHEET NO. 44

COMPUTATION OF LOGARITHMIC REGRESSION LINE FOR POPULATION GROWTH OF CALIFORNIA, 1890–1950

Year	Population (in 1000)					Trend in Logarithms	Trend in Natural Numbers (in 1000)
	X	Y	Log Y	X Log Y	X^2		
1890	− 3	1,213	3.08386	− 9.25158	9	3.05557	1,136.5
1900	− 2	1,485	3.17173	− 6.34346	4	3.21705	1,648.5
1910	− 1	2,378	3.37621	− 3.37621	1	3.37854	2,387.0
1920	0	3,427	3.53491			3.54002	3,467.4
1930	+ 1	5,677	3.75412	+ 3.75412	1	3.70150	5,029.2
1940	+ 2	6,907	3.83929	+ 7.67858	4	3.86299	7,294.3
1950	+ 3	10,472	4.02003	+ 12.06009	9	4.02447	10,579.5

$$\begin{array}{rr} 24.78015 + 23.49279 & 28 \\ -\ 18.97126 & \\ \hline +\ 4.52153 & \end{array}$$

$$\log a = \frac{\Sigma \log Y}{N} = \frac{24.78015}{7} = 3.54002$$

$$\log b = \frac{\Sigma X \log Y}{\Sigma x^2} = \frac{4.52153}{28} = .161483$$

$$\log Y = \log a + (\log b)X = 3.54002 + .161483X$$

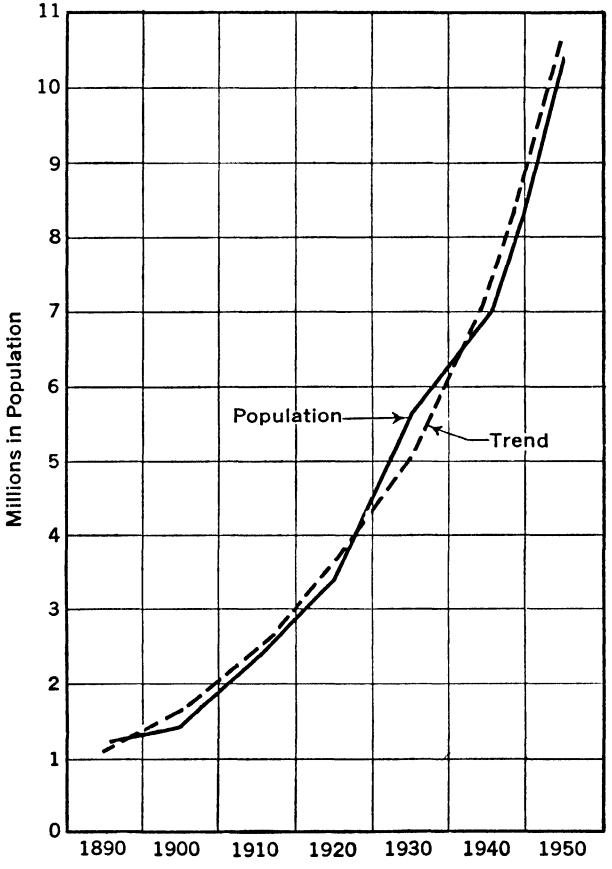


FIG. 65. Logarithmic trend of population growth of California plotted on absolute, or arithmetic scale. (*U.S. Decennial Census*)

II. The Hyperbolic Curve

If both X and Y increase in a geometric progression the relationship may be expressed by a hyperbolic function of the general type $y = ax^b$.

To fit this hyperbolic curve by the least squares method in logarithmic form the equation required is that stated as Formula No. 69.

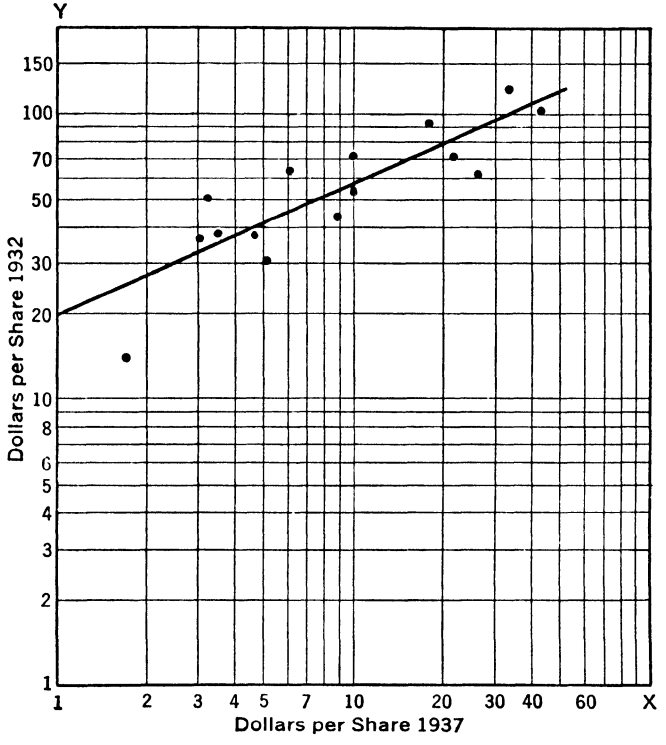


FIG. 66. Prices of 15 stocks in 1932 on the New York Stock Exchange (X), and prices of the same 15 stocks in 1937 on the N.Y. Stock Exchange (Y), plotted on double-logarithmic scales. A linear relationship on double-logarithmic scales indicates a Hyperbolic Curve on the Arithmetic scale. See Worksheet No. 43, page 257

FORMULA NO. 69

$$\log Y = \log a + b \ (\log X)$$

The normal equations required are:

I. $\Sigma(\log Y) = N \log a + b \ \Sigma(\log X)$

II. $\Sigma(\log X \cdot \log Y) = \log a \ \Sigma(\log X) + b \ \Sigma(\log X)^2$

To determine whether this type of curve is likely to fit the data, plot the original data on double logarithmic paper. If it falls in approximately a straight line as do the data in Fig. 66 the hyperbolic curve in logarithmic form will likely fit quite well.

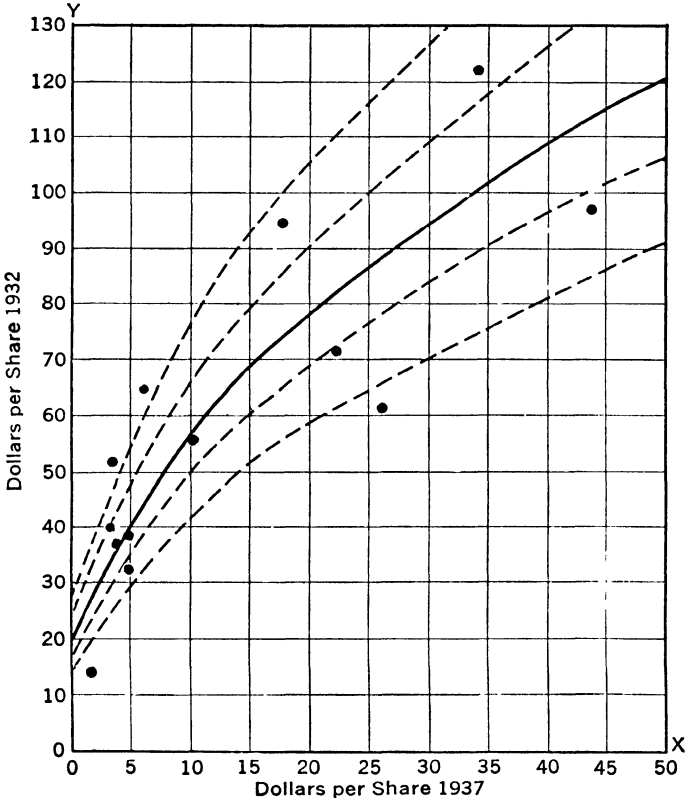


FIG. 67. Prices of 15 stocks in 1932 on the New York Stock Exchange (X), and prices of the same 15 stocks in 1937 (Y) plotted on the Arithmetic scale with Hyperbolic Curve. See Worksheet No. 43

WORKSHEET NO. 45

COMPUTATION OF HYPERBOLIC FUNCTION BY LOGARITHMS FOR 1932 PRICES OF 15 STOCKS (X) AND 1937 PRICE OF THE SAME 15 STOCKS (Y) NEW YORK EXCHANGES FROM NEW YORK TIMES — 1932 LOW (X) 1937 HIGH (Y)

1932 Prices X	1937 Prices Y	$\log X$	$\log Y$	$(\log X)^2$	$(\log Y)^2$	$\log X \cdot \log Y$
18.0	95.0	1.25527	1.97772	1.575703	3.911376	2.482573
10.0	56.0	1.00000	1.74819	1.000000	3.056168	1.748190
4.6	38.0	0.66276	1.57978	0.439251	2.337727	1.047015
6.0	65.0	0.77815	1.81291	0.605517	3.286642	1.410716
26.0	61.0	1.41497	1.78533	2.002140	3.187403	2.526188
34.0	122.0	1.53148	2.08636	2.345431	4.352898	3.195219
44.0	107.0	1.64345	2.02938	2.700928	4.118383	3.335186
10.0	72.0	1.00000	1.85733	1.000000	3.449675	1.857330
3.2	51.5	0.50515	1.71181	0.255177	2.930293	0.864721
9.0	45.0	0.95424	1.65321	0.910574	2.733103	1.577559
3.0	37.0	0.47712	1.56820	0.227643	2.459251	0.748220
5.0	32.0	0.69897	1.50515	0.488559	2.265477	1.052055
22.0	71.0	1.34242	1.85126	1.802091	3.427164	2.436841
1.7	14.0	0.23045	1.14613	0.053107	1.313614	0.264125
3.5	38.4	0.54407	1.58433	0.2996012	2.510102	0.861984
Sums		14.03850	25.89709	15.702133	45.309276	25.407922
Means		0.93590	1.72647			
Corrections				13.138632	44.710549	24.237086
Deviations				2.562501	.598727	1.170832

$$b = \frac{\Sigma(\log x) \cdot (\log y)}{\Sigma(\log x)^2} = \frac{1.170832}{2.562501} = .4569$$

$$\log a = \log Y - b(\log X) = 1.72647 - (.4569)(.935901)$$
$$= 1.72647 - .42761 = 1.29886.$$

$$\log Y = (\log a) + b(\log X)$$
$$= 1.29886 + .4569 \log X$$

The time and labor required to compute a hyperbolic curve can be reduced to a minimum by the use of logarithms in Worksheet No. 45, which employs the short method of reducing all values to deviations from the mean.

FORMULA NO. 70

$$S_{(\log y)}^2 = \frac{\Sigma(\log y)^2 - \dfrac{[\Sigma(\log x)\cdot(\log y)]^2}{\Sigma(\log x)^2}}{N}$$

$$= \frac{.59727 - \dfrac{(1.170832)^2}{2.562501}}{15}$$

$$= \frac{.59727 - \dfrac{1.370848}{2.562501}}{15}$$

$$= \frac{.59729 - .53496}{15} = \frac{.06233}{15} = 0.0041553$$

$$S_{(\log y)} = .0644$$

The computations for the same 15 stocks during the "bull-market" of 1941–46 are:

$N = 15$	Deviations
$\Sigma(\log X) = 17.01572$	$\Sigma(\log x)^2 = 1.490799$
$\Sigma(\log Y) = 26.87406$	$\Sigma(\log y)^2 = 0.337557$
$\Sigma(\log X)^2 = 20.793108$	$\Sigma(\log x \log y) = 0.480987$
$\Sigma(\log Y)^2 = 48.48523$	$b = .322667$
$\Sigma(\log X \log Y) = 30.96642$	$\log a = 1.425611$

$$\log Y = 1.425611 + .322637 (\log X)^*$$

$$S_{(\log y)}^2 = .012158 \qquad S_{(\log y)} = .1103$$

This method when Worksheet No. 45 is used with deviations from the mean is an easy and short device for computing regression for two variables both of which increase in geometric progressions.

Standard Error of Estimate for Logarithmic Regression Curves

The standard error of estimate is here expressed in terms of a ratio. In the example above it is

$$Sr = \text{antilog } S_{(\log y)} = \text{antilog of } .0644 = 1.16$$

* A comparison of these two samples will be made by the *t*-test in Chapter 20.

where Sr = represents the standard error of estimate in terms of ratios. It means that in 68 cases out of 100 the actual value will not exceed the estimated value by more than 16 percent. On the negative (−) side the logarithm is 9.9456 − 10, the antilog of which is .8822. In other words the actual values will not vary more than 16 percent above or 12 percent below the estimated regression values in 68 cases out of 100.

$$2S_{(\log y)} = 2 \times 0.0644 = 0.1288$$
$$3S_{(\log y)} = 3 \times 0.0644 = 0.1932$$

for which the antilogs are obtained by the same method as for $S_{(\log y)}$ above. See Fig. 67.

Index of Correlation Based on Logarithmic Values

$$P^2(\log x)(\log y) = \frac{[\Sigma(\log x)(\log y)]^2}{\Sigma(\log x)^2 \cdot \Sigma(\log y)^2} = \frac{(1.170832)^2}{2.56250 \times .598727}$$
$$= \frac{1.370847}{1.534268} = .89348$$

$P(\log x)(\log y) = .945$ which is a ratio and states in ratio or logarithmic terms the relationship, or degree of similarity, in the variation of log X and log Y.

III. Compound Interest Curve

FORMULA NO. 71

$$y = P(1 + r)^x$$

r = rate of interest
x = years (interest compounded annually)
P = principal
y = the sum to which the principal amounts at the end of x-years.

In logarithmic form the equation is $\log y = \log P + \log (1 + r)x$

IV. The Gompertz Curve

The Gompertz curve is one type of "growth curve." Since growth generally passes through the three phases of (a) slow be-

ginning when the organism or business is getting rooted and started, (2) rapid expansion, and (3) maturity or the saturation period; this curve fits the growth function quite well.

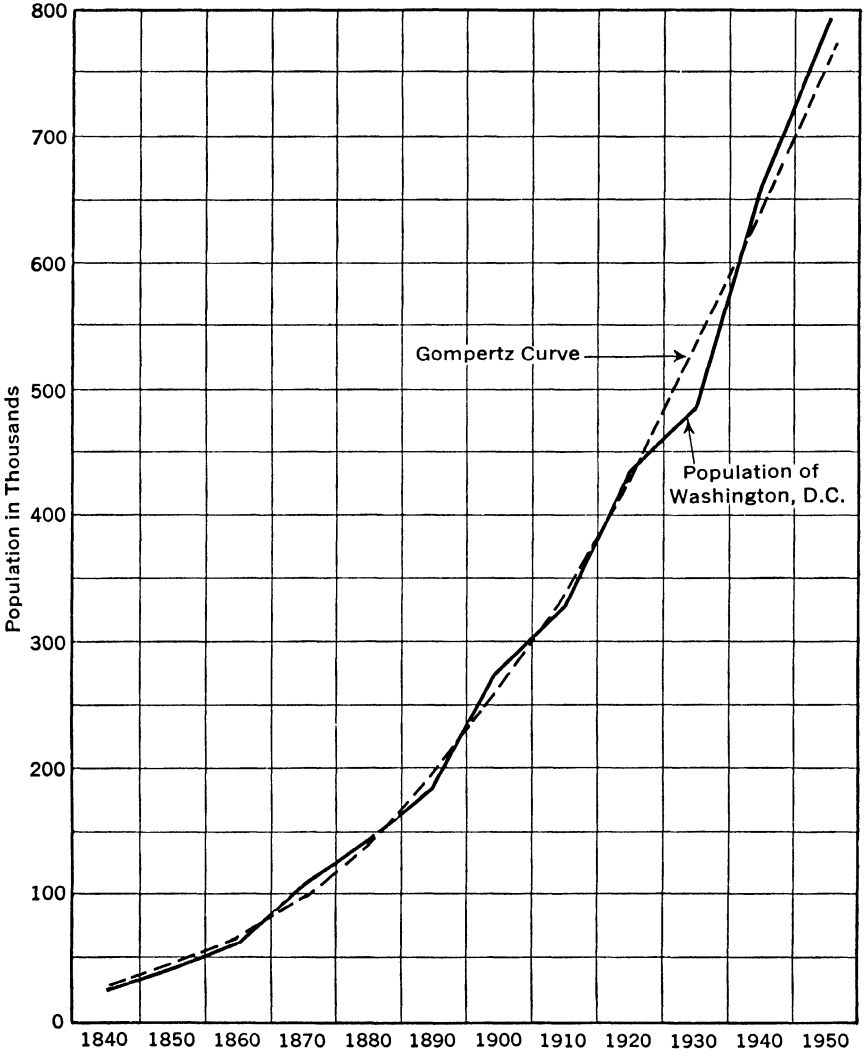


FIG. 68. Gompertz curve fitted to population of Washington, D.C., 1840–1950. (*U.S. Decennial Census*, 1950. See Worksheet No. 44)

FORMULA NO. 72

$$y = ab^{c^x}$$

This natural form of the curve is transformed to the logarithmic form

FORMULA NO. 73

$$\log_y = \log a + (\log b)c^x$$

In this curve (log a) represents the maximum growth or ceiling which the curve approaches as the growth function moves toward maturity. (Log $b)c^x$, in which x represents time or successive time periods measures the amount the trend falls short of the maximum or ceiling at a given time or for a specific x value. C will have a value between 0 and 1. And as X increases the curve continues to increase at a decreasing rate.

WORKSHEET NO. 46, PART I

COMPUTATION OF QUANTITIES REQUIRED TO FIT A GOMPERTZ CURVE TO POPULATION OF WASHINGTON, D.C., 1840 TO 1950

1	2	3	4	5
Year	Actual Population Y	Log Y	Sub-totals	First Differences
1840	23,364	4.36856		
1850	40,001	4.60206	$S_1 = 18.79532$	
1860	61,161	4.78648		
1870	109,199	5.03822		$d_1 = S_2 - S_1$
				$= 2.61392$
1880	147,203	5.16791		
1890	188,932	5.27626	$S_2 = 21.40924$	
1900	278,718	5.44516		
1910	331,069	5.51992		
1920	437,571	5.64105		$d_2 = S_3 - S_2$
1930	486,869	5.68741	$S_3 = 23.04889$	$= 1.63965$
1940	663,091	5.82158		
1950	792,234	5.89885		

FORMULA No. 72. $y = ab^{c^x}$

Logarithmic form $\log y = \log a + (\log b)c^x$

FORMULA No. 74
$$\begin{cases} c^n = \dfrac{d_2}{d_1} \\ \log b = \dfrac{d_1(c-1)}{(c^n-1)^2} \\ \log a = \dfrac{1}{n}\left\{ S_1 - \dfrac{d_1}{(c^n-1)} \right\} \end{cases}$$

$$C^n = \frac{d_2}{d_1} = \frac{1.63965}{2.61392} = .627276$$

$$C = \sqrt[4]{.627276}$$

$$\log \text{ of fourth root of } .627276 = \frac{9.797463 - 10}{4} = 9.949366 - 10$$

$$C = .88995$$

$$\log b = \frac{2.61392 \times (-.11005)}{(.627276 - 1)^2}$$

$$= \frac{-.2876619}{.1389272} = -2.0705945$$

$$\log a = \tfrac{1}{4}\left\{18.79532 - \frac{2.61392}{-.372724}\right\}$$

$$= \tfrac{1}{4}\{18.795320 - (-7.013018)\}$$

$$= \tfrac{1}{4}(25.808338) = 6.452084$$

The required equation, therefore, is

$$\log y = 6.452084 - 2.0705945\ (.88995^x)$$

(Formula No. 72 $y = y = ab^{c^x}$. Computations of a, b, and c).

The student will note that the Gompertz curve provides a better fit to the data on the population of Washington, D.C. than does the logistic curve in Worksheet No. 47. Compare Figures 68 and 69. In the logistic curve the product of b and c is used, but in the Gompertz c is the index of the power of b.

WORKSHEET NO. 46, PART II

COMPUTATION OF TREND OF A GOMPERTZ CURVE FITTED TO DATA OF POPULATION OF WASHINGTON, D.C., 1840 TO 1950

(1) Year	(2) x	(3) c^x	(4) $(\log b)c^x$	(5) $\log y$ (4) + $\log a$	y antilog of (5)
1840	0	1.000000	− 2.070594	4.381490	24,071
1850	1	.889950	− 1.842725	4.609359	40,678
1860	2	.792011	− 1.639933	4.812151	64,885
1870	3	.704851	− 1.459460	4.992624	98,312
1880	4	.627276	− 1.298834	5.153250	142,420
1890	5	.558244	− 1.155897	5.296187	197,775
1900	6	.496815	− 1.028702	5.423382	265,116
1910	7	.442140	− 0.915492	5.536592	344,125
1920	8	.393475	− 0.814727	5.637357	433,970
1930	9	.350173	− 0.725066	5.727018	533,325
1940	10	.311636	− 0.645272	5.806812	641,916
1950	11	.277340	− 0.574258	5.877826	754,792

Logistic Curve

Raymond Pearl and L. J. Reed developed the logistic curve which is frequently called the Pearl-Reed growth curve which measures or describes the usual expansion of a large population. The larger the population, usually, the more accurately this curve describes its growth. It conforms quite closely to the population expansion of the United States over the past 110 years as shown in Worksheet No. 47 and Fig. 69. It is not the fault of the equation or curve, perhaps, that it does not conform to the population line more closely after 1940. The rapid growth of Washington, D.C. during and after World War II, an expansion of 20 percent in ten years, is an unusual increase for so large a city. This high rate is, no doubt, temporary and in the future this city, as most others will return to a more gradual rate of expansion.

This curve, however, also fails to fit the growth of Boston, Chicago or Detroit. The reason for the discrepancy in these cases

is that unlike Washington which is growing fast, these industrial centers are now expanding slowly. This discrepancy is an illusion since the recent growth of most large cities is in the suburban areas which are not listed in the census as part of the city.

This curve, as well as the Gompertz Curve, is quite useful in measuring the growth of industries such as railways, automobile production, rayon, coal, steel, electricity, and the like over a long period of time. It is primarily a time series trend device, because the X-variable must fall in units of uniform length or size.

The logistic curve does not employ logarithms. Its technique is simple and the elementary student need have no difficulty with it. For a long-time period it gives a much better fit for the growth of most populations and industries than a straight line or a parabola. If, however, the period covered is sufficiently long to cover the decline or decay of the population or industry after it has passed its peak the parabola may give a better fit, as is indicated in Worksheet No. 48 on Post Office data.

WORKSHEET NO. 47, PART I

COMPUTATION OF QUANTITIES REQUIRED IN FITTING A LOGISTIC CURVE TO DATA OF POPULATION OF WASHINGTON, D.C., 1840–1950

Year	Actual Population	$\frac{10{,}000{,}000}{Y}$	Sub-totals	First Differences
1840	23,364	428		
1850	40,001	250		
1860	61,162	164	$S_1 = 934$	$d_1 = S_2 - S_1$
1870	109,199	92		$= 187 - 934$
				$= -747$
1880	147,203	68		
1890	188,932	53		
1900	278,718	36	$S_2 = 187$	
1910	331,069	30		
				$d_2 = S_3 - S_2$
1920	437,571	23		$= 71.2 - 187$
1930	486,869	20.5		$= 115.8$
1940	663,091	15.1	$S_3 = 71.2$	
1950	792,234	12.6		

FORMULA NO. 75

$$\frac{1}{Y} = a + bc^x$$

FORMULA NO. 76

$$\begin{cases} c^n = \dfrac{d_2}{d_1} \\ b = \dfrac{d_1(c-1)}{(c^n-1)^2} \\ a = \dfrac{1}{N}\left(S_1 - \dfrac{d_1}{c^n-1}\right) \end{cases}$$

$$C^n = C^4 = \frac{d_2}{d_1} = \frac{-115.8}{-747} = +155$$

$$C = \sqrt[4]{.155} = .6275$$

Computation of C

The 4th root of .155 is the antilogarithm of the logarithm of .155 divided by 4

$$\frac{39.19033 - 40}{4} = 9.79758 - 10$$

$$b = \frac{d_1(c-1)}{(c^n-1)} = \frac{-747(.6275-1)}{(.155-1)^2} = \frac{-747-.3725}{(-.845)^2}$$

$$= \frac{278.26}{714025} = 389.7$$

$$a = \tfrac{1}{4}\left(934 - \frac{-747}{-.845}\right)$$

$$= \tfrac{1}{4}\,(934 - 884)$$

$$= \tfrac{1}{4}\,(50) = 12.5$$

In the computation of both Gompertz and logistic curves, decades or centuries of data are not necessary. Although time periods must fall in three equal sections, the individual time units may be short, such as years, or two-year or three-year periods. Even shorter periods may be used if seasonal factors can be eliminated.

WORKSHEET NO. 47, PART II

COMPUTATION OF ORDINATES OF TREND OF A LOGISTIC CURVE TO DATA OF POPULATION OF WASHINGTON, D.C., 1840 TO 1950

Year	x	c^x	bc^x	$a + bc^x$	$\frac{Y'}{10{,}000{,}000}$ $\frac{}{a + bc^x}$
1840	0	1.0000	389.7	402.2	24.863
1850	1	.6275	244.5	257.0	38.910
1860	2	.3938	153.5	166.0	60.241
1870	3	.2471	96.3	108.8	91.712
1880	4	.1551	60.4	72.9	137,174
1890	5	.0973	37.9	50.4	198,413
1900	6	.0611	23.8	36.3	275,482
1910	7	.0383	14.9	27.4	364,964
1920	8	.0240	9.4	21.9	456,621
1930	9	.0151	5.9	18.4	543,478
1940	10	.0095	3.7	16.2	617,284
1950	11	.0060	2.3	14.8	675,676

Measures of Growth and Decay

An excellent example of the increase and successive decrease of a series is the number of Post Offices in the United States from 1850 to 1950. The number reached a peak in 1900 from which it has declined rather steadily.

Since growth, maturity, and decay are universal phenomena not only in the physical and biological worlds but also in business and social activities their measurement is of great practical importance in all fields of human interest. A new business, such as radio was in 1920, makes a rapid growth for a few years, reaches maturity as it did in the later 1940's and, if television continues to expand, radio may possibly decline. Such measurements are of great practical value in business management.

The reason the logistic curve does not fit the population growth

of Washington, D.C. any better is that Washington's normal growth was highly disturbed by the two World Wars and the recent expansion of all Federal activities.

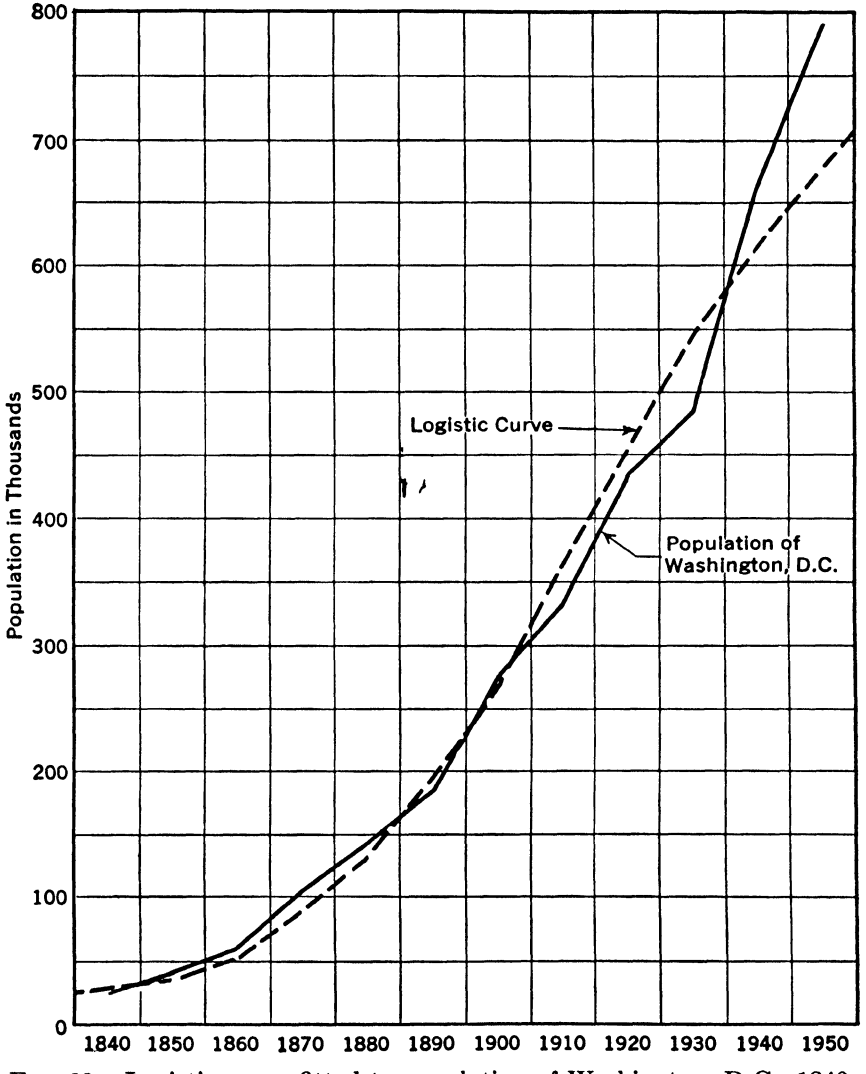


FIG. 69. Logistic curve fitted to population of Washington, D.C., 1840–1950. (*U.S. Decennial Census*, 1950. See Worksheet No. 45)

WORKSHEET NO. 48

EXPANSION AND DECLINE OF POST OFFICES IN U.S., 1850 TO 1950
Source: *Statistical Abstract of U.S.*, 1950

Year	X	(P.O. in 1000) Y	XY	X^2	X^3	X^4	X^2Y
1850	− 5	18.4	− 92.0	25	− 125	625	46.00
60	− 4	28.5	− 114.0	16	− 64	256	456.0
70	− 3	28.5	− 85.5	9	− 27	81	256.5
80	− 2	43.0	− 86.0	4	− 8	16	172.0
90	− 1	62.4	− 62.4	1	− 1	1	62.4
1900	0	76.7					
10	+ 1	59.6	+ 59.6	1	+ 1	1	59.6
20	+ 2	52.6	+ 105.2	4	+ 8	16	210.4
30	+ 3	49.0	+ 147.0	9	+ 27	81	441.0
40	+ 4	44.0	+ 176.0	16	+ 64	256	704.0
50	+ 5	41.0	+ 205.0	25	+ 125	625	1,025.0
	0	503.7	+ 692.8	110	0	1,958	3,846.9
			− 439.8				
			+ 253.0				

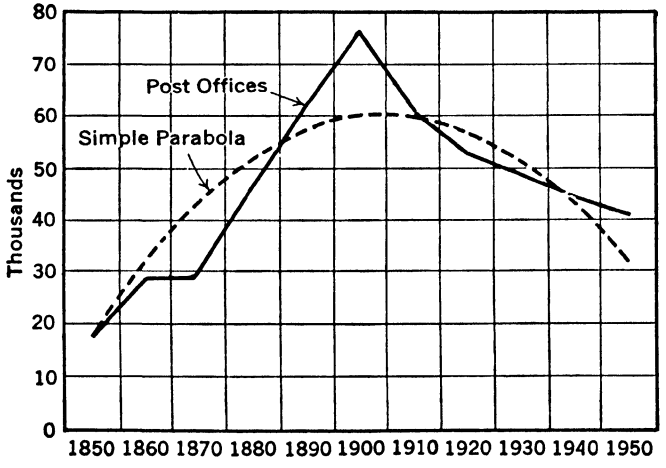


FIG. 70. Simple parabola fitted to expansion and decline of number of post offices in U.S., 1850–1950. See Worksheet No. 48. (*Statistical Abstract of U.S.*, 1950)

FORMULA NO. 77

$$Y = a + bx + cx^2$$

FORMULA NO. 78

$$b = \frac{xy}{x^2} = \frac{+153.3}{110.0}$$
$$= 1.3946$$

FORMULA NO. 79

$$\begin{aligned} Na + \Sigma x^2 c &= \Sigma Y \\ \Sigma x^2 a + \Sigma x^4 c &= \Sigma x^2 Y \\ 11a + 110c &= 503.7 \\ 110a + 1958c &= 3{,}846.9 \\ 1a + 10c &= 45.7909 \\ 1a + 17.8c &= 34.9718 \\ \hline -7.8c &= 10.8191 \\ -c &= 1.3871 \\ c &= -1.3871 \end{aligned}$$

$$\begin{aligned} 11a + 110c &= 503.7 \\ 11a + 110(-1.3871) &= 503.7 \\ 11a - 152.581 &= 503.7 \\ 11a &= 152.581 + 503.7 \\ 11a &= 656.281 \\ a &\frac{656.281}{11} = 59.6619 \end{aligned}$$

$$Y = a + bx + cx^2$$
$$Y = 59.6619 + .39461\,X - 1.3871X^2$$

SUMMARY

1. Most relationships between variables in the physical as well as in the social sciences are not straight-line, but conform to some type of non-linear function.

2. Non-linear lines may be computed from class averages and drawn into coordinate graphs of the plotted data. Such curves are often ex-

cellent descriptive devices and by reading off the estimated values of the data from the class averages curved line, values may be obtained comparable to the standard error of estimate and the coefficients of determination and correlation.

3. Mathematical non-linear lines may be computed from any desired formula. The simplest and most commonly used functions are (1) the simple parabola, (2) the cubic parabola, (3) the exponential function or compound interest curve, and (4) the logistic and other growth curves.

4. In selecting a mathematical curve for data, a prejudgment must usually be made as to the nature of the function which expresses the real relationship in the data.

5. Unless the nature of the function existing in the population is known several curves may have to be computed before the suitable one is discovered, or extensive experimentation may have to be done.

6. If the nature of the function is known, a mathematical line is much superior to any freehand or class average line because it makes more precise and easily available the information.

7. Unless a mathematical line does describe the true function, it becomes merely a descriptive device and may be no better than a freehand line.

SELECTED REFERENCES

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XVI. Prentice-Hall, Inc., New York, 1948.

Davies, G. R., and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences*, Chapter VI. John Wiley & Sons, New York, 1933.

Ezekiel, M., *Methods of Correlation Analysis*, Chapters VI, VII, VIII, XIV, XVI, XVIII, XXII. John Wiley & Sons, New York, 1941. (Contains a complete and excellent reference on regression.)

Fisher, R. A., *Statistical Methods for Research Workers*. Oliver and Boyd, Ltd., Edinburgh, 1925.

Holzinger, K. J., *Statistical Methods for Students in Education*, Chapter XVI. Ginn & Company, Boston, 1928.

Mills, F. C., *Statistical Methods, Revised*, Chapters VII, XVII. Henry Holt and Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapters VIII, XII, XIII. John Wiley & Sons, New York, 1942.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, Chapter XV. McGraw-Hill Book Company, New York, 1940.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chapters VI and VIII. Harcourt, Brace & Company, New York, 1943.

Waugh, A. E., *Elements of Statistical Method*, Chapters IX, X. McGraw-Hill Book Company, New York, 1938.

REVIEW QUESTIONS

1. Why do straight-line trends not fit many series of data well? Give examples that are not straight-line.
2. State the procedure in fitting a non-linear freehand regression line through class averages.
3. Of what use is a freehand non-linear trend line and what are its deficiencies?
4. How may a standard error of estimate be computed for a freehand line?
5. What are the normal equations required to compute a simple parabola? Write them.
6. What is their form when changed to deviations from the means? What economies are effected by this change? Explain.
7. Explain the required form of worksheet to compute a simple parabola from deviations from the means?
8. What are the advantages of the short cut method over the detailed method in the computation of the deviations from the mean? Explain in detail.
9. Does a simple parabola fit well many series of data in the social sciences? Why? In the physical sciences? Why?
10. Describe the method of computing the standard error of estimate data fitted with a parabola.
11. What are the differences between a simple parabola and a cubic parabola? Explain in detail.
12. What additions are required to the least squares worksheet of a simple parabola in order to compute a cubic parabola?
13. Write the normal equations for a cubic parabola.
14. Explain the principles of the Doolittle Method of solving equations.
15. What is the Back Solution?
16. What is the difference between the equation, $Y = a + b_{yx}X$ and the equation, $X = a + b_{xy}Y$? Explain fully.
17. Explain why no regression line is complete without its standard error of estimate.

EXERCISES

1. Population of San Antonio, Texas

1860	8,235
1870	12,256
1880	20,550
1890	37,673
1900	53,321
1910	96,614
1920	161,379
1930	231,542
1940	253,854
1950	406,811

Source: *U.S. 1950 Census*

Compute $\frac{1}{y} = a + bc^x$

2. Population of Baltimore, Maryland

1810	46,555
1820	62,738
1830	80,020
1840	102,313
1850	169,054
1860	212,418
1870	267,354
1880	332,313
1890	434,429
1900	508,957
1910	558,485
1920	733,826
1930	804,874
1940	859,100
1950	940,205

Source: *U.S. 1950 Census*

Compute $\frac{1}{y} = a + bc^x$

3. Production of Crude Petroleum in U.S. (1,000,000 bbls.)

1910	210
1915	281
1920	443
1925	764
1930	898
1935	997
1940	1,353
1945	1,714
1950	2,260

Source: *Statistical Abstract of U.S.*

4. Electrical Energy Production in U.S. (1,000,000 KW)

1920	39,405
1925	61,451
1930	91,112
1935	95,287
1940	141,837
1945	222,486

Source: *Statistical Abstract of U.S.*

$$\text{Compute } \frac{1}{y} = a + bc^x$$

5. Number of Divorces Granted in U.S.

1905	67,976
1910	83,045
1915	104,298
1920	170,505
1925	175,449
1930	195,961
1935	218,000
1940	264,000
1945	485,000

Source: *Statistical Abstract of U.S.*

6. Currency Outside of Banks (in millions of dollars)

1938	5,417
1939	6,005
1940	6,699
1941	8,204
1942	10,936
1943	15,814
1944	20,881
1945	25,096
1946	26,516

Source: *Federal Reserve Bulletin*

Compute $y = ab^{cx}$

7. Trend of Change of prices of 15 stocks on N.Y. Exchange, 1941 prices (X) and 1946 prices of same stocks (Y).

	1941 X	1946 Y
Atchison	25.0	12.1
Atl. Coast Lines	15.0	8.3
Ill. Central	4.2	45.4
So. Pacific	9.1	70.0
Reynolds Tobacco	20.0	46.7
Ligget Myers	50.4	101.6
American Tobacco	34.0	100.0
Pullman Company	16.7	69.2
Johns Manville	14.5	55.6
Otis Elevator	9.6	39.2
Pitt-Plate Glass	13.0	48.6
Kress S.S.	17.0	45.4
Woolworth Co.	21.4	62.4
Sharp & Dohne	3.0	39.6
General Tire	6.4	60.0

Source: *N.Y. Times quotations of N.Y. Exchange*

Compute: $\log y = \log a + b \log X$.

CHAPTER 11

MULTIPLE AND PARTIAL CORRELATION

All reference to multiple correlation is usually omitted from introductory texts as a calculation that is unnecessary or too difficult for a first semester's work in statistics. It is, however, often quite useful for a student who does not specialize in statistics and may be presented in a form that is as easy to compute as simple correlation.

In many statistical problems the relationships among the variables are so complex that simple correlation gives a very incomplete and inadequate answer to the question. Only multiple correlation is sufficient for satisfactory results. In the case of the relationship of corn yield and rainfall, the very important factors of soil fertility, amount of fertilizer, methods of cultivation, time of planting, variety of corn, and amount of sunshine are all omitted. In any case corn yield is a resultant of many factors. Successive simple correlations between corn yield and each one of these independent variables give answers that are both incomplete and incorrect. In the method of simple correlation all the independent factors except one are eliminated and ignored. In multiple correlation all these independent variables may be included, measured, and their influence on the dependent factor duly apportioned and related. Another reason for presenting an introduction to multiple correlation in an elementary text is that it aids the student to grasp the total concept of correlation much more clearly.

Although correlation does not prove a causal relationship among associated factors, it does measure the degree to which a

factor or group of factors is associated with another factor. If, for example, the rate of farm income is associated with (1) size of farm, (2) number of cows on farms, and (3) wheat yield, it is possible to measure the degree of association between rate of income and the combined other factors. This is *Multiple Correlation.*

Since the several independent factors do not have equal degrees of association with the dependent factor, to obtain a correct measure of their total association with it, each one of them must be weighted according to its contribution to the total of the independent factors. If the association of wheat yield is twice as close as that of size of farm it must have twice as much weight in the total degree of association. So it is with all independent factors. Each one must be weighted properly in the total result. The multiple net regression coefficients are the correct weights, because they measure the relationship between the dependent factor and each independent factor after the influence of all the other factors has been taken into consideration and computed out or eliminated.

To reduce these logical relationships to a mathematical formula three steps are necessary:

1. Compute the deviations from the mean for each variable as a measure of the variation in that variable which is to be compared with the variation of the dependent factor.
2. Express these summed deviations in terms of their own standard deviations in order to reduce them to a common standard of comparison, and,
3. Compute the multiple net regression coefficients as relative weights for each variable.

If the various factors are represented by X_1, X_2, X_3, X_4, $\cdots$ etc., the deviations from the means will be x_1, x_2, x_3, x_4, $\cdots$ etc. If these deviations are summed and divided by their own standard deviations, the results are $\frac{\Sigma x_1}{\sigma_1}$, $\frac{\Sigma x_2}{\sigma_2}$, $\frac{\Sigma x_3}{\sigma_3}$, $\frac{\Sigma x_4}{\sigma_4}$, $\cdots$ etc. If the multiple net regression coefficients are now used as weights of the relative importance of each factor they become

$$b_2 \frac{\Sigma x_2}{\sigma_2},\ b_3 \frac{\Sigma x_3}{\sigma_3},\ b_4 \frac{\Sigma x_4}{\sigma_4},\ \cdots \text{ etc.}$$

When the coefficients for the factors are solved by the least squares multiple net regression method the correlation equation becomes

$$R^2_{1.234} = \frac{(b_{12.34}\Sigma x_1 x_2)}{\Sigma x_1^2} + \frac{(b_{13.24}\Sigma x_1 x_3)}{\Sigma x_1^2} + \frac{(b_{14.23}\Sigma x_1 x_4)}{\Sigma x_1^2}, \text{ etc.}$$

$$= \frac{b_{12.34}\Sigma x_1 x_2 + b_{13.24}\Sigma x_1 x_3 + b_{14.23}\Sigma x_1 x_4 +}{\Sigma x_1^2} \text{ etc.,}$$

in which the correct weight of each independent variable is included in a total, which is expressed as a ratio to the total variation in the dependent variable. This ratio is $R^2_{1.234\ldots n}$, or the coefficient of Multiple Determination. It expresses as a percentage the amount of the total variation in the dependent variable which is associated with the combined variation of the independent variables. If, for instance, $R^2_{1\,234\ldots n} = .75$, one is to understand that 75% of the variation of X_1 is associated with the weighted combined variation of $X_2 + X_3 + X_4 + \cdots X_n$, and that 25% of the variation of X_1 is *not* associated with the independent variables included in this problem, but must be accounted for by other variables.

Meaning of symbols:

X_1 = the dependent variable, usually designated by Y in simple correlation

X_2, X_3, X_4, etc., = the several independent variables
R = multiple straight-line coefficient of correlation
R^2 = multiple straight-line coefficient of determination
$b_{12.34\ldots n}$ = net regression coefficient between X_1 and X_2
$b_{13.24\ldots n}$ = net regression coefficient between X_1 and X_3
$b_{14.23\ldots n}$ = net regression coefficient between X_1 and X_4

The multiple linear regression equation is

$$X_1 = a_{1.234\ldots n} + b_{12.34\ldots n}X_2 + b_{13.24\ldots n}X_3 + b_{14.23\ldots} + \cdots + b_{1n\ldots n-1}X_n$$

For four variables the equations are:

FORMULA NO. 80

$$a_{1.234} = \overline{X}_1 - b_{12.34}\overline{X}_2 - b_{13.24}\overline{X}_3 - b_{14.23}\overline{X}_4$$

FORMULA NO. 81

$$\Sigma x_2^2 b_{12.34} + \Sigma x_2 x_3 b_{13.24} + \Sigma x_2 x_4 b_{14.23} = \Sigma x_1 x_2$$
$$\Sigma x_2 x_3 b_{12.34} + \Sigma x_3^2 b_{13.24} + \Sigma x_3 x_4 b_{14.23} = \Sigma x_1 x_3$$
$$\Sigma x_2 x_4 b_{12.34} + \Sigma x_3 x_4 b_{13.24} + \Sigma x_4^2 b_{14.23} = \Sigma x_1 x_4$$

All the above formulas may be expanded as desired to include more variables.

FORMULA NO. 82

$$S_1^2 = \frac{\Sigma z^2}{N - M}, \quad \text{or} \quad S_1^2 = \frac{\Sigma x_1^2 - \Sigma (x_1')^2}{N - M}$$

S = check sum of paired variables.

Multiple net regression equations and coefficients form the basis for multiple determination. If the multiple net regression equations express straight-line relationships the coefficient of multiple determination is based on a straight line relationship and is indicated by the symbol $R_{1.234\ \ldots\ n}$.

FORMULA NO. 83

$$R^2_{1.234\ \ldots\ n} = 1 - \frac{S_1^2}{\sigma_1^2}$$

If the multiple net regression equations and coefficients are curvilinear, the coefficient of multiple determination is based on curvilinear relationships and is indicated by the symbol P (Capital Rho)

FORMULA NO. 84

$$\mathrm{P}^2_{1.234\ \ldots\ n} = 1 - \frac{S_1^2}{\sigma_1^2}$$

The distinction between R and P is quite important. If any type of curvilinear regression is used in the solution, the correlation should be designated with P.

PRACTICAL EXAMPLE OF MULTIPLE CORRELATION APPLIED TO ECONOMIC DATA

It is well known by practical businessmen, as by professional economists, that the profits of any particular business venture or unit are affected by and associated with many diverse factors. The profits of a drug store may rest on (1) location, (2) delivery service, (3) quality of goods sold, (4) dependability of pharmacist, (5) advertising, (6) efficiency of management, (7) personality of clerks, and many other factors. These factors may not all be equal in importance.

The profits of a factory may be associated with (1) loyalty of labor force, (2) nearness to raw materials, (3) cheapness of transportation, (4) nearness to consumer market, (5) credit standing, (6) good will, and many other factors of varying strength and weight.

The relationships treated in the sciences of sociology, psychology, education, biology, business management, as well as economics are all measurable by multiple correlation insofar as they may be stated in quantitative or measurable terms. Divorces may be associated with (1) economic conditions of the home, (2) different ages of couple, (3) different religious views, (4) sex inhibitions, and many other factors. The achievement of students in school may be due to (1) I.Q., (2) home environment, (3) health, (4) quality of teachers, (5) interest in subject, and scores of other factors of various weights. Educators have made extensive use of correlation in educational research, but these problems are so complex that only multiple correlation is adequate to treat many of them. The student will readily recall many other fields of study and particular cases in which the powerful statistical analytical device of multiple correlation may be used.

The following problem on the rate of return on nineteen mixed wheat and dairy farms of various sizes is a good illustration of the detailed procedures and methods for the computation of extensions for variables.

WORKSHEET NO. 49

COMPUTATION OF EXTENSIONS FOR FOUR VARIABLES USING CHECK SUM [1]

Variables					Extensions with X_2				
X_2	X_3	X_4	X_1	S	X_2^2	X_2X_3	X_2X_4	X_2X_1	X_2S
220	9	16	3.5	248.5	48,400	1,980	3,520	770.0	54,670.0
325	16	20	7.2	368.2	105,625	5,200	6,500	2,340.0	119,665.0
520	8	17	9.0	554.0	270,400	4,160	8,840	4,680.0	288,080.0
400	13	16	4.2	433.2	160,000	5,200	6,400	1,680.0	173,280.0
285	18	27	16.9	346.9	81,225	5,130	7,695	4,816.5	98,866.5
297	14	22	9.7	442.7	88,309	4,158	6,534	2,880.9	101,781.9
307	11	20	5.2	343.2	94,249	3,377	6,140	1,596.4	105,362.4
215	8	20	6 5	249.5	46,225	1,720	4,300	1,397.5	53,642.5
220	1	20	8.3	249.3	48,400	220	4,400	1,826.0	54,846 0
255	9	22	10.0	296.0	65,025	2,295	5,610	2,550.0	75,480.0
225	3	20	5.2	253.2	50,625	675	4,500	1,170.0	56,970 0
400	8	25	14.6	447.6	160,000	3,200	10,000	5,840.0	179,040.0
135	3	20	5.1	163.1	18,225	405	2,700	688 5	22,018.5
148	7	15	4.0	174.0	21,904	1,036	2,220	592.0	25,752.0
360	4	22	6.2	392.2	129,600	1,440	7,920	2,232.0	141,192.0
280	6	20	10 2	316.2	78,400	1,680	5,600	2,856.0	88,536.0
480	6	25	7.4	518.4	230,400	2,880	12,000	3,552.0	248,832.0
305	7	18	5 6	335 6	93,025	2,135	5,490	1,708 0	102,358.0
400	9	23	10.6	442.6	160,000	3,600	9,200	4,240.0	177,040.0
5,777	160	388	149.4	6,474.4	1,949,937	50,491	119,569	47,415.8	2,167,412.8

Extensions with X_3				Extensions with X_4			Extensions with X_1	
X_3^2	X_3X_4	X_3X_1	X_3S	X_4^2	X_4X_1	X_4S	X_1^2	X_1S
81	144	31.5	2,236.5	256	56 0	3,976 0	12.25	869.75
256	320	115.2	5,891.2	400	144.0	7,364 0	51 84	2,651.04
64	136	72.0	4,432.0	289	153.0	9,418.0	81 00	4,986.00
169	208	54.6	5,631.6	256	67.2	6,931 2	17 64	1,819.44
324	486	304.2	6,244.2	729	456.3	9,366.3	285 61	5,862.61
196	308	135.8	4,797.8	484	213.4	7,539.4	94.09	3,324.19
121	220	57.2	3,775 2	400	104 0	6,864.0	27 04	1,784 64
64	160	52.0	1,996.0	400	130.0	4,990.0	42.25	1,621.75
1	20	8 3	249.3	400	166.0	4,986.0	68 89	2,069.19
81	198	90.0	2,664.0	484	220.0	6,512.0	100.00	2,960.00
9	60	15.6	759.6	400	104.0	5.064.0	27.04	1,316 64
64	200	116.8	3,580.8	625	365.0	11,190.0	213.16	6,534.96
9	60	15.3	489.3	400	102.0	3,262.0	26.01	831 81
49	105	28.0	1,218.0	225	60 0	2,610 0	16.00	696.00
16	88	24.8	1,568.8	484	136.4	8,628.4	38.44	2,431.64
36	120	61.2	1,897 2	400	204.0	6,324.0	104.04	3,225.24
36	150	44.4	3,110.4	625	185.0	12,960.0	54.76	3,836.16
49	126	39.2	2,349.2	324	100 8	6,040.8	31.36	1,879.36
81	207	95.4	3,983.4	529	243.8	10,179.8	112.36	4,691.56
1,706	3,316	1,361.5	56,874.5	8,110	3,210.9	134,205.9	1,403.78	53,391.98

[1] This table is adapted from Mordecai Ezekiel, *Methods of Correlation Analysis*, p. 360, and used by permission of John Wiley and Sons, Inc., New York.

SHORT METHOD OF COMPUTING EXTENSIONS IN WORKSHEET NO. 49

1. Set the data down in the first columns under the heading of "variables."

2. Add the data in each line crosswise for the check sum. See that the total of the check sum column exactly equals the totals of all the variables.

3. For Extensions with X_2, multiply all the items in each line by the X_2 value in that line.

4. For Extensions with X_3, multiply all the items in the line beginning with column X_3 by the value of the X_3 column.

5. In each succeeding set of extensions begin with the column of data for which extensions are being computed. That is, the Extensions with X_4 begin with column X_4, Extensions with X_5 begin with column X_5, and so on to X_1, which is always the last.

The check sum column in each set of extensions checks all the totals for that variable. Check sum X_2S should equal the sum of columns containing an X_2. The check sum X_3S should equal the sums of all columns containing an X_3, and so on for all variables.

WORKSHEET NO. 50

COMPUTATIONS OF DEVIATIONS FROM MEANS WITH CHECK SUM

	X_2	X_3	X_4	X_1	S	Line
Sums	5,777.	160	388	149.4	6,474 4	1
Means	304.05263	8.42105	20.42105	7.86316	340.75789	2
Extensions X_2	1,949,937.	50,491.	119,569.	47,415.8	2,167,412.8	3
Corrections	1,756,512.04	48,648 42	117,972.42	45,425.46	1,968,558 35	4
Deviations of X_2	193,424.96	1,842.58	1,596.58	1,990.34	198,854.45	5
Extensions X_3		1,706.	3,316.	1,361.50	56,874.50	6
Corrections		1,347.37	3,267.37	1,258.10	54,521.25	7
Deviations of X_3		358.63	48.63	103.40	2,353.25	8
Extensions X_4			8,110.	3,210.90	134,205.90	9
Corrections			7,923.37	3,050.90	132,214.05	10
Deviations of X_4			186.63	160.00	1,991.85	11
Extensions X_1				1,403.78	53,391.98	12
Corrections				1,174.76	50,909.24	13
Deviations of X_1				229.02	2,482.74	14

The corrections in Worksheet No. 50 are computed by multiplying all the line *Sums* through successively by the *Means*. For the corrections for X_2 *all* the sums are multiplied by the mean of X_2. These multiplications are (5,777)(304.05263) and (160)(304.05263) and (388)(304.05263) and so on to include the check sum total. For the corrections for X_3 the sums beginning with X_3 (160 in this case), are multiplied through by the mean of X_3. These computations are (160)(8.42105) and (388)(8.42105) and so on to include the check sum. The extensions with X_2, X_3, and all the little x's, or deviations from the mean, are obtained by subtracting the *corrections* from the Extensions in the line above.

WORKSHEET NO. 51

SOLUTION OF NORMAL EQUATIONS BY THE DOOLITTLE METHOD WITH CHECK SUM *

Line	X_2	X_3	X_4	X_1	S
I	193,424.96	1,842.58	1,596.58	1,990.34	198,854.45
I′	− 1.00000	− .00953	− .00825	− .01029	− 1.02807
II	(1,842.58)	358.63	48.63	103.40	2,353.25
	(− 1,842.58)	− 17.56	− 15.22	− 18.97	− 1,895.08
2		341.07	33.41	84.43	458.17
II′		− 1.00000	− .09796	− .24754	− 1.34333
III	(1,596.58)	(48.63)	186.63	160.0	1,991.85
	(− 1,596.58)	(− 15.20)	− 13.17	− 16.42	− 1,640.55
		(− 33.41)	− 3.27	− 8.27	− 44.88
3			170.19	135.31	306.42
III′			− 1.00000	− .79505	− 1.80045

Back Solution

$b_{12.34}$	$b_{13.24}$	$b_{14.23}$
.01029	.24754	.79505
− .00656	− .07788	.79505
− .00162	.16966	
.00211		

* See explanation of Doolittle Method in Chapter 10.

MEANING OF NET REGRESSION COEFFICIENTS

A regression coefficient computed for two variables only leaves entirely out of the picture the effects of all other independent variables. If a regression coefficient is computed for wheat yields, (X_2), and income, (X_1), alone, no account is taken of the influence on income of such evidently present factors as (X_3) size of farm, (X_4) rainfall, (X_5) fertilizer, (X_6) man power, and many other factors. These other variables are simply excluded from the problem. This elimination of vital factors usually results in over-magnifying the coefficient of the variable which is included. When these other factors are included in the problem and multiple net regression equations are computed for all of them at once, each net regression coefficient is corrected for the influence of the others by the mathematical process of holding all the others constant while the value of this particular one is computed. In the first case, the other independent variables are simply eliminated. In the second case, they are all included in the problem and adjustments are made for their separate effects. The latter method, the use of multiple net regression equations and coefficients, gives a much more accurate analysis.

For the variables and equations developed in Worksheets No. 49 and No. 50, for income (X_1), size of farms (X_2), number of cows on farms (X_3), and yield of wheat per acre (X_4), the differences between the two methods are as follows:

1. b_{12} (the regression coefficient between income and size of farm alone) = .01029.

$b_{12.34}$ (the net regression coefficient between income and size of farms after the influence of the number of cows, (X_3), and wheat yields, (X_4) has been removed) = .00211.

The net regression coefficient is only about one-fifth as large as the simple coefficient. The simple coefficient, (b_{12}), was magnified beyond its true size by the influence of X_3 and X_4, which was not computed out.

2. b_{13} (the regression coefficient between income and number of cows on farms) = .2883.

$b_{13.24}$ (the net regression coefficient between income and number of cows on farms after the influence of the size of farms, (X_2), and wheat yields, (X_4), has been removed) = .16966. The net coefficient is less than two-thirds as large as the simple coefficient. In b_{13} the true relationship is obscured by the influence of X_2 and X_4 which had not been computed out.

3. b_{14} (the regression coefficient between income and wheat yield per acre) = .8573.

$b_{14.23}$ (the net regression coefficient between income and wheat yields after the influence of size of farms, (X_2), and number of cows on farms, (X_3), has been removed) = .79505. The net regression coefficient is only eleven-twelfths as large as the simple coefficient. The multiple net regression coefficients appear at the end of the back solution in Worksheet No. 51. This example is a fair illustration of the principle stated in the preceding paragraph, *a simple regression coefficient tends to obscure* the true value *of the relationship between two variables.* The multiple net regression coefficients approach more nearly the true relationship.

The coefficient b_{12} expresses an obscured and usually magnified relationship between X_1 and X_2. The net regression coefficient $b_{12.3}$ measures the relationship between X_1 and X_2 after the influence of X_3 is taken into consideration and its weight

WORKSHEET NO. 52

FINAL STEPS IN SOLUTION OF MULTIPLE CORRELATION PROBLEM, CHECK, COMPUTATION OF R^2 AND OF A *

Variables	Regression Coefficients	Equation III	Check	Equation X_1	Computation of R^2	Means	Computation of A
	1	2	(1 × 2) 3	4	5	6	7
X_2	.00211	1,596.58	3.36878	1,990.34	4.19962	304.05	.6415
X_3	.16966	48.63	8.25057	103.40	17.54284	8.42	1.4285
X_4	.79505	186.63	148.38018	160.00	127.20800	20.42	16.2349
Sums		160.00	159.99953	229.02	148.95046	7.86	18.3049

* Adapted from Mordecai Ezekiel, *Methods of Correlation Analysis*, and used by permission of John Wiley and Sons, Inc., New York.

removed from the effect of X_2 on X_1. In the still more complex net regression coefficient, $b_{1.2345}$, the net relationship between X_2 and X_1 is computed after the effects of X_3, X_4, and X_5 have been brought into the problem and their effects on the relationship between X_2 and X_1 have been computed out. The fact that the influence of the other independent variables has been removed in each case is the reason such coefficients are called net regression coefficients. They are much more exact measures of the relationship between variables than the simple two variable coefficients.

$$\begin{aligned} a_{1.234} &= \bar{X}_1 - b_{12.34}\bar{X}_2 - b_{13.24}\bar{X}_3 - b_{14.23}\bar{X}_4 \\ &= 7.86 - 18.3 = -10.44 \\ X_1 &= a_{1.234} + b_{12\,34}X_2 + b_{13.24}X_3 + b_{14\,23}X_4 \\ &= -10.44 + .00211X_2 + .16966X_3 + .79505X_4 \end{aligned}$$

$$R^2 = \frac{\Sigma x_1x_2b_{12\,34} + \Sigma x_1x_3b_{13\,24} + \Sigma x_1x_4b_{14\,23}}{\Sigma x_1^2}$$

$$= \frac{4.19962 + 17.54284 + 127.20800}{229.02}$$

$$= \frac{148.95046}{229.02} = .6504 \quad \text{and}$$

$$R = .8065$$

STANDARD ERROR OF ESTIMATE FOR MULTIPLE REGRESSION

FORMULA NO. 85

$$S_1^2 = \frac{\Sigma x_1^2 - (b_{12\,34}\Sigma x_1x_2 + b_{13.24}\Sigma x_1x_3 + b_{14.23}\Sigma x_1x_4)}{N - M}$$

$$= \frac{\Sigma x_1^2 - \Sigma x_1'^2}{N - M}$$

$$= \frac{229.2 - 148.95}{19 - 4}$$

$$= \frac{80.25}{15} = 5.36$$

See Chapter 13, Formula No. 100.

$$S_1 = 2.32$$

Since $R^2_{1.234}$ = the percentage of the total variation in X_1 that is associated with all the independent variables combined,

$1 - R^2$ = the percentage of variation in X_1 that is *not* explained by $X_2 + X_3 + X_4$, etc.

Σx_1^2 = total amount of variation in X_1.

$(b_{12.34}\Sigma x_1x_2 + b_{13.24}\Sigma x_1x_3 + b_{.4.23}\Sigma x_1x_4)$ = total amount of variation in X_1 that is explained by or associated with variation in $X_2 + X_3 + X_4$, etc. Therefore,

$$\Sigma x_1 - (b_{12.34}\Sigma x_1x_2 + b_{13.24}\Sigma x_1x_3 + b_{14.23}\Sigma x_1x_4 + \text{etc.})$$

is the portion of variation in X_1 that is not associated with the included independent variables. In this problem these values are:

$\Sigma x_1^2 = 229.02$ total variation to be explained

$(b_{12.34}\Sigma x_1x_2 + \text{etc.},) = 148.95$ total amount of variation explained by X_2, X_3, and X_4

$NS_1^2 = 80.25$ the amount of variation not explained

$$\bar{S}_1^2 = \frac{80.25}{N - M} = \frac{80.25}{19 - 4} = \frac{80.25}{15} = 5.36$$

$$\bar{S}_1 = 2.32$$

CORRECTION OF COMPUTATIONS FOR SIZE OF SAMPLE

Ezekiel develops the following set of corrections for coefficients computed from small samples. The standard error for multiple correlation increases (1) with the increase in the number of variables employed and also (2) with the *decrease* in size of samples.

N = the number of paired items of data used
M = the number of variables included

Following the rule for degrees of freedom the correlation becomes: $\frac{N - 1}{N - M}$ which in this case is $\frac{19 - 1}{19 - 4} = \frac{18}{15}$.

Applied to $R^2 = .6504$ the coefficient of multiple correlation corrected for size of sample and number of variables becomes:

$$\bar{R}^2 = 1 - (1 - .6504)\frac{18}{15}$$
$$= (1 - .4195)$$
$$= .5805 \quad \text{and} \quad \bar{R} = .762$$

FORMULA NO. 86

$$\bar{R}^2 = I - (I - R^2)\left(\frac{N - I}{N - M}\right)$$

PARTIAL CORRELATION

Multiple correlation measures the total combined association of the independent variables with the dependent variable but gives no measure of the relative importance of the several independent factors. In the problem above $R^2_{1.234} = .6504$, but this does not explain how much of this total is due to (1) wheat yield, (2) number of cows, (3) size of farm. This additional information is usually desirable and sometimes necessary, but it must be obtained from additional computations.

In this text three methods are presented for measuring the association of each independent variable with the dependent variable after the influence of the other factors is removed.

1. *Separate determination* is given in Formula Nos. 87, 88 and 89. For the present problem the values for the three variables are:

$$\text{No. 87} \quad d_{12.34} = \frac{b_{12.34}\Sigma x_1 x_2}{\Sigma x_1^2} = \frac{4.1996}{229.02} = .0183$$

$$\text{No. 88} \quad d_{13.24} = \frac{b_{13.24}\Sigma x_1 x_3}{\Sigma x_1^2} = \frac{17.5428}{229.02} = .0766$$

$$\text{No. 89} \quad d_{14.23} = \frac{b_{14.23}\Sigma x_1 x_4}{\Sigma x_1^2} = \frac{127.208}{229.02} = .5554$$

$$R^2_{1\,234} = .6504 = \text{Total}$$

Separate determination has the advantage that the separate determinations exactly equal the total determination, but it has the great disadvantage that it is erratic and undependable whenever there is a high degree of inter-correlation between two or more of the independent variables. Also, its standard error is not determinable.[1]

2. *Partial correlation* is the standard and most widely used method of measuring the separate values of several variables. Ezekiel gives, perhaps, the best definition: "The coefficient of partial correlation may be defined as a measure of the extent to which that part of the variation in the dependent variable which was *not* explained by the other independent factors can be ex-

[1] Mordecai Ezekiel, *Methods of Correlation Analysis*, pp. 498–99.

plained by the addition of the new factor." [1] If (1) size of farm and (2) wheat yields explain .5943 of the variation in farm income, how much more of the variation would be explained by including (3) the number of cows? .4057 is yet unexplained. Including cows raises the total determination to .6504, which is (.6504) − (.5947) = .0561. Adding the variable on cows explained .0561 of the yet unexplained variation in X_1 of .4507; $\frac{.0561}{.4057} = .1383$ the amount of additional explanation of variation supplied by including cows.

The formulas for partial correlation for three independent variables are [2]:

FORMULA NO. 90

$$r^2_{12.34} = 1 - \frac{(1 - R^2_{1.234})}{(1 - R^2_{1.34})}$$

FORMULA NO. 91

$$r^2_{13.24} = 1 - \frac{(1 - R^2_{1.234})}{(1 - R^2_{1.24})} \quad \text{and}$$

FORMULA NO. 92

$$r^2_{14.23} = 1 - \frac{(1 - R^2_{1.234})}{(1 - R^2_{1.23})}$$

in which the numerator of the fraction, $(1 - R^2_{1.234})$, remains the same for all equations, but the denominator varies for each computation. The denominator in the equation for each variable is 1 minus the multiple determination coefficient for all the dependent factors except the one being measured. For $r^2_{12.34}$ the multiple determination is $R^2_{1.34}$ which omits X_2. For $r^2_{13.24}$ this multiple

[1] *Ibid.*, p. 214. Quoted by permission of the author and John Wiley & Sons, Inc.

[2] There is another method for computing partial correlation based on building up the partial coefficients from successive simple correlation coefficients which has been in wide use for many years. It is not presented here because it reaches the same results obtained by the shorter method given below. The student may find it in many of the more advanced texts or older books as Mills, F. C., *Statistical Methods, Revised*, pp. 554–60, or Pearson and Bennett, *Statistical Methods Applied to Agriculture*, p. 191.

determination is $R^2_{1.24}$ which omits X_3. All the formulas follow the same principle.

To compute the partial correlation for each factor it is necessary first to compute the several multiple determinations successively eliminating the desired factor. This result requires the simultaneous solution of the required equations for the new regression coefficients. This computation is neither long nor difficult, because the same equations developed in Worksheet No. 50, for the original total multiple determination may be used by omitting successively one term, as

for $R^2_{1.34}$

$$\Sigma x_3^2 b_{13.4} + \Sigma x_3 x_4 b_{14.3} = \Sigma x_1 x_3$$
$$\Sigma x_3 x_4 b_{13.4} + \Sigma x_4^2 b_{14.3} = \Sigma x_1 x_4$$

for $R^2_{1.24}$

$$\Sigma x_2^2 b_{12.4} + \Sigma x_2 x_4 b_{14.2} = \Sigma x_1 x_2$$
$$\Sigma x_2 x_4 b_{12.4} + \Sigma x_4^2 b_{14.2} = \Sigma x_1 x_4$$

for $R^2_{1.23}$

$$\Sigma x_2^2 b_{12.3} + \Sigma x_2 x_3 b_{13.2} = \Sigma x_1 x_2$$
$$\Sigma x_2 x_3 b_{12.3} + \Sigma x_3^2 b_{13.2} = \Sigma x_1 x_3$$

The equations to be taken from Worksheet No. 50 for solutions are:

$R_{1.34}$ { (1) 358.63 + 48.63 = 103.40
(2) 48.63 + 186.63 = 160.00

$R_{1.24}$ { (1) 193,424.96 + 1,596.58 = 1,990.34
(2) 1,596.58 + 186.63 = 160.00

$R_{1.23}$ { (1) 193,424.96 + 1,842.58 = 1,990.34
(2) 1,842.58 + 358.63 = 103.40

WORKSHEET NO. 53

SOLUTION OF SIMULTANEOUS EQUATIONS

$R^2_{1.34}$, $R^2_{1.24}$, AND $R^2_{1.23}$

$R^2_{1.34}$				
I	358.63	48.63	=	103.40
I′	− 1.0000	− .135599	=	− .288319
II	48.63	186.63	=	160.00
	− 48.63	− 6.59418	=	− 14.02094
1,2		180.03582	=	145.7906
II′		− 1.000	=	− .810833

Back Solution

$b_{13.4}$	$b_{14.3}$
.288319	.810833
− .109948	.810833
.178371	

Regression Coefficients		Equation I	Computation of $R^2_{1\ 34}$	
X_3	.178371	103.40	18.444	$R^2_{1.34} = \frac{148.177}{229.02}$
X_4	.810833	160.00	129.733	
		229.02	148.177	$= .6470$

$$r^2_{12.34} = 1 - \frac{(1 - .6504)}{(1 - .6470)} = 1 - \frac{.3496}{.3530} = 1 - .9904 = .0096$$

This coefficient of partial correlation is 0.96% of the unexplained determination before size of farms was included in the problem. In this sample, size of farm, X_2, has little association with rate of income.

WORKSHEET NO. 54

$R^2_{1.24}$

Front Solution

I	193,424.96	+	1,596.58	=	1,990.34
I′	− 1.0000	−	.0082543	=	− .01029
II	1,596.58		186.63	=	160.00
	− 1,596.58	−	13.17865	=	− 16.42886
1,2			173.45135	=	143.57114
II′				=	− .806132

Back Solution

$b_{12.4}$	$b_{14.2}$
.010290	.806132
− .006554	.806132
.003736	

Regression Coefficients		Equation I	Computation of $R^2_{1.24}$	
X_2	.003636	1,990.34	7.2369	$R^2_{1.24} = \frac{136.2180}{229.02}$
X_4	.806132	160.00	128.9811	
		229.02	136.2180	$= .5943$

$$r^2_{13.24} = 1 - \frac{(1 - .6504)}{(1 - .5943)} = 1 - \frac{.3496}{.4057} = 1 - .8617 = .1383$$

This coefficient of partial correlation is 13.83% of the unexplained determination before number of cows was included in the problem. In this sample number of cows on farms does not have a large degree of association with rate of income, but it is much more important than size of farms.

WORKSHEET NO. 55

$R^2_{1.23}$					
I	193,424.96	+	1,842.58	=	1,990.34
I′	− 1.0000	−	.0095426	=	− .01029
II	1,842.58	+	358.63	=	103.40
	− 1,842.58	−	17.5526	=	− 18.9062
1,2			341.0774	=	+ 84.4398
II′		−	1.0000	=	− .247567

Back Solution

$b_{12.3}$	$b_{13\,2}$
.01029	.247567
− .002358	
.007932	.247567

Regression Coefficients		Equation I	Computation of $R^2_{1\,23}$	
X_2	.007932	1,990.34	15.78738	$R^2_{1.23} = \frac{41.38581}{229.02}$
X_3	.247567	103.40	25.59843	
		229.02	41.38581	$= .1806$

$$r^2_{14.23} = 1 - \frac{(1 - .6504)}{(1 - .1806)} = 1 - \frac{.3496}{.8194} = 1 - .4266 = .5734$$

This coefficient of partial correlation is 57.34% of the unexplained determination before wheat yields were included in the problem. In this sample wheat yield is by far the most important factor associated with rate of income.

BETA COEFFICIENTS

As was stated earlier in this chapter it is possible to measure the degree of association between the dependent and each of the independent variables by weighting each independent factor with its multiple regression coefficient. This measure requires, however, that the value of each factor be stated in "standard measures," or in terms of its own standard deviations. The computation reduces the variance in all variables to a common measure or a kind of common denominator. When the variation in each of the X_1, X_2, $X_3, \cdots X_n$ is reduced to σ_1, σ_2, $\sigma_3, \cdots \sigma_n$, the ratios between the standard deviations weighted by the regression coefficients are called the "*beta*" coefficients and are expressed as follows:

$$\frac{X_1}{\sigma_1} = \beta_{12.34}\frac{X_2}{\sigma_2} + \beta_{13.24}\frac{X_3}{\sigma_3} + \beta_{14.23}\frac{X_4}{\sigma_4} + a_3',$$

and the partial betas as

$$\beta_{12.34}\frac{\sigma_2}{\sigma_1}, \text{ etc.}$$

The computations for this problem are, by Formula Nos. 93, 94, and 95.

$$\begin{cases} \beta_{12.34} = b_{12.34}\dfrac{\sigma_2}{\sigma_1} = .00211\dfrac{100.9}{3.47} = \dfrac{.222989}{3.47} = .0642 \\ \beta_{13.24} = b_{13.24}\dfrac{\sigma_3}{\sigma_1} = .16966\dfrac{4.84}{3.47} = \dfrac{.7353244}{3.47} = .2119 \\ \beta_{14.23} = b_{14.23}\dfrac{\sigma_4}{\sigma_1} = .79505\dfrac{3.13}{3.47} = \dfrac{2.488506}{3.47} = .7171 \end{cases}$$

The squared betas are:

	Beta	Squared Beta
$\beta_{12.34}$	.0642	.0041
$\beta_{13.24}$	.2119	.0449
$\beta_{14.23}$	.7171	.5142

TABLE 22A

COMPARISON OF SEPARATE DETERMINATION, PARTIAL CORRELATION AND BETA COEFFICIENTS

	Variables	*d*	Partial	Betas	Betas Squared
X_2	Size of farms	.0183	.0096	.0642	.0041
X_3	Number of cows	.0766	.1383	.2119	.0449
X_4	Wheat yields	.5554	.5734	.7171	.5142

Although these several measures for each variable are not identical they rank in the same order in each case, and, in fact, tell the same story. They check and corroborate each other.

SECOND MULTIPLE DETERMINATION PROBLEM

The purpose of the first multiple and partial correlation problem was to explain and illustrate the theory and techniques of these statistical methods for the beginning student and to indicate some of their practical applications to current business problems. Multiple and partial correlation methods are useful in agriculture, manufacturing, marketing, management and other industrial and business activities. Business is never a matter of only two variables. Its essential nature is multiple and simple two-variable analyses usually exaggerate the weight of the single independent variable included at the expense of the excluded variables. Only a multiple and partial analysis can properly weight and balance their individual and total contributions to the result.

Many qualities are desired in an employee, but time and expense preclude the checking of all of them. If, however, a careful multiple and partial determination study indicates that 90% of the success in the job is associated with some four or five major factors, the many minor factors can be largely disregarded at a considerable saving of cost. The analysis of markets and the fitting of a good to its market as well as the creation of index numbers can be greatly improved by the multiple-partial analysis of the problem.

Perhaps one of the very greatest and most devastating faults of human beings is their almost universal tendency to oversimplify their thinking about all the problems of life. We are so often wrong in our judgments because we are so incomplete in our analysis. The wider and more complete view gained from multiple and partial analysis in undergraduate education is a step toward a more adequate understanding of business, social, and political problems.

The purpose of this second problem is (1) to expand the students' understanding of the theory and techniques to larger and more complex problems, (2) to apply the multiple and partial analysis to primary data obtained by experimental research, (3) to show how the data of the physical and biological sciences may be organized in complex problems which have a most tremendous impact on large populations and society in general.

DATA [1]

The data in this problem are taken from the *Third Interim Report on The Smog Problem in Los Angeles County* by the Stanford Research Institute, 1950, for the Committee on Smoke and Fumes of the Western Oil and Gas Association of which William L. Stewart, Jr. is chairman.

Seventeen series of data are included in the Report. The five series on Temperature in the original Report are represented in this problem by the one series, X_5, Temperature, Degrees Above Normal. The other series included in this problem, the S.R.I. Smog Index, X_1; the S.R.I. Inversion Intensity Index, X_2; the 24-Hour Wind Movement, X_3; and the Visibility in Los Angeles, X_4, are all taken as given in the Report, page 22.

WORKING HYPOTHESIS

The hypothesis underlying the smog study is that combustion wastes which include chemically active hydrocarbons and other compounds gradually accumulate in the atmosphere unless carried

[1] Data are used with the permission of The Committee on Smoke and Fumes of the Western Oil and Gas Association and are taken from page 22 of *The Third Interim Report.*

away by fairly strong and regular wind currents. In a large city or industrial area the exhaust from automobiles, smoke and fumes from factories, homes, trash burners and other sources gradually saturate the air if there is little wind movement. It is believed that the intense sunshine accelerates the smog condition by increasing the chemical action among the nitrogen oxides and hydrocarbons resulting in an excess of ozone and other active chemical compounds which are painful to the population and injurious to many substances. Smog results from (1) excessive smoke and fumes, (2) intense sunshine, and (3) slow air movement.

Certain mining and manufacturing towns in Pennsylvania and other areas located in deep, closed valleys have suffered results similar to those manifest in Los Angeles. In no place is the smog condition uniform or continuous, as these data indicate. If the relative weights of the major factors were known, steps could be devised in many cases to remove them. The purpose of this multiple and partial determination problem is to measure the relative, total and partial weights of the several factors involved.

Worksheet No. 56 includes the data, check sums and the column sums. The Worksheet on Extensions is omitted since it is quite long and is similar to Worksheet No. 49. The other Worksheets, Nos. 57, 58 and 59 are all included in small print as notes for use as references by teachers and students to clarify points in the complex techniques of such problems.

The use of multiple and partial regression and correlation analysis is quite extensive. February business problems can be solved adequately by simple correlation and regression. The factors associated with sales are numerous. Location of store, price of goods, type of sales force, advertising, delivery of goods, guarantee of products, credit terms, competition, and many other factors affect volume of sales and prices.

Problems of production, finance, and personnel are equally complex. Rarely is it possible for a personnel manager to employ laborers on only one qualification. Age, education, experience, health, character, and even family and home relationships affect the value of employees. Credit risks rest on many factors and investment opportunities are varied.

WORKSHEET NO. 56

MULTIPLE CORRELATION STUDY OF LOS ANGELES, 1949–1950

Smog, 1949, 1950. For 34 Days

S.R.I. Inversion Intensity	Wind Movement 24 hours	Visibility Los Angeles Noon Miles	Temperature Degrees Above Normal	S.R.I. Smog Index	Check Sum
X_2	X_3	X_4	X_5	X_1	S
11.1	107	3.0	13	12.7	146.8
18.4	110	2.0	11	16.9	158.3
10.0	147	5.0	13	9.7	184.7
16.5	161	5.0	21	15.9	219.4
25.2	101	3.0	15	14.3	158.5
23.5	110	2.5	22	29.0	187.0
12.8	119	2.0	12	15.0	160.8
16.6	117	3.0	12	11.0	159.6
13.7	156	2.0	13	14.4	199.1
10.3	179	15.0	18	8.2	230.5
11.7	99	1.0	7	10.6	129.3
14.3	86	.7	1	4.9	106.9
18.2	117	.5	7	16.3	159.0
13.8	96	.7	10	12.4	132.9
8.7	107	1.5	10	14.4	141.6
18.1	87	.2	7	19.6	131.9
9.6	93	1.0	7	12.5	123.1
13.8	139	.6	15	12.1	180.5
14.2	122	.7	11	13.0	160.9
27.8	109	1.0	9	12.9	159.7
28.3	105	1.5	11	12.2	158.0
28.2	117	2.0	14	16.5	177.7
9.0	121	1.0	3	2.3	136.3
9.0	106	1.5	6	4.8	127.3
8.9	130	5.0	11	6.5	161.4
23.0	88	1.0	5	12.1	129.1
12.3	181	30.0	27	10.3	260.6
11.5	109	2.0	8	10.3	140.8
12.8	130	5.0	10	6.0	163.8
14.1	110	2.0	11	10.7	147.8
14.6	97	3.0	10	11.3	135.9
19.5	86	.6	7	15.0	128.1
14.4	80	2.0	10	19.3	125.7
15.0	88	.5	2	7.1	112.6
528.9	3910	107.5	369	420.2	5,335.6

WORKSHEET NO. 57

Computation of Deviations From the Means for Normal Equations on Los Angeles Smog Data

	X_2	X_3	X_4	X_5	X_1	S	Line	
SUMS	528.9	3910.0	107.5	369	420.2	5,335.6	1	
Means	15.55588	115.00000	3.161765	10.852941	12.358823	156.9294117	2	
Extensions X_2	9,316.23	59,591.10	1,471.86	5,854.10	7,006.91	83,240.20	3	
Corrections	8,227.50	60,823.49	1,672.26	5,740.12	6,536.58	82,999.95	4	
Deviations X_2	1,088.73	− 1,232.39	− 200.40	113.98	470.33	240.25	5	Equation I
Extensions X_3		471,288.00	15,552.70	45,685.00	47,682.30	639,799.10	6	
Corrections		449,650.00	12,362.50	42,435.00	48,323.00	613,594.00	7	
Deviations X_3		21,638.00	3,190.20	3,250.00	− 640.70	26,205.10	8	Equation II
Extensions X_4			1,309.73	1,824.00	1,199.69	21,357.98	9	
Corrections			339.89	1,166.69	1,328.57	16,869.91	10	
Deviations X_4			969.84	657.31	− 128.88	4,488.07	11	Equation III
Extensions X_5				4,989.00	4,933.10	63,285.20	12	
Corrections				4,004.74	4,560.41	57,906.95	13	
Deviations X_5				984.26	372.69	5,378.25	14	Equation IV
Extensions X_1					6,021.64	66,843.64	15	
Corrections					5,193.18	65,941.74	16	
Deviations X_1					828.46	901.90	17	

WORKSHEET NO. 58

Solution of Normal Equation in Worksheet No. 57 by Doolittle Method for Multiple Regression Coefficients of Los Angeles Smog Data

Forward Solution

Line	X_2	X_3	X_4	X_5	X_1	S
I	1,088.73	− 1,232.39	− 200.40	113.98	470.33	240.25
I′	− 1.000000000	+ 1.131951900	+ .184067675	− .104690786	− .431998751	− .220669955
II		21,638.00	3,190.20	3,250.00	− 640.70	26,205.10
		− 1 395.01	− 226.84	+ 129.02	+ 532.39	+ 271.95
1,2		20,242.99	2.963.36	3,379.02	− 108.31	26,477.05
II′		− 1.000000000	− .146389441	− .166922969	+ .005350494	− 1.307961422
III			969.84	657.31	− 128.88	4,488.07
			− 36.89	20.98	86.57	44.22
			− 433.80	− 494.65	+15.86	− 3,875.96
1,2,3			499.15	183.64	− 26.45	656.33
III′			− 1.000000000	− .367905439	+ .052990083	− 1.31489532
IV				984.26	372.69	5,378.25
				− 11.93	− 49.24	− 25.15
				− 564.04	+ 18.08	4,419.63
				− 67.56	+ 9.73	− 241.47
				340.73	351.26	692.00
				− 1.000000000	− 1.030904235	− 2.030933583

Back Solution

b 12.345	b 13.245	b 14.235	b 15.234
.431998751	− .005350494	− .052990083	1.030904235
− .107926175	− .172081596	− .379275275	1.030904235
− .079566079	+ 063279084	− .432265358	
− .129215712	− .114153006		
.115290785			

WORKSHEET NO. 59

FINAL STEPS IN SOLUTION OF MULTIPLE CORRELATION PROBLEM CHECK, R^2 AND a.

	Regression Coefficients 1	Equation IV 2	Check (1 × 2) 3	Equation X_1 4	Determination $R_{1.2345}$ (1 × 4) 5	Means 6	(1 × 6) 7
X_2	.1152907	113.98	13.14083	470.33	54.22467	15.556	1.79346
X_3	− .1141530	3,250.00	− 370.99725	− 640.70	73.13783	115.000	− 13.12759
X_4	− .4322654	657.31	− 284.13237	− 128.88	55.71036	3.162	− 1.36682
X_5	1.0309042	984.26	1.014.67776	372.69	384.20769	10.853	11.18840
Sums		372.69	372.68897	828.46	567.28055	12.359	− 1.51255

$$R^2_{1.2345} = \frac{567.28}{828.46} = .6847 \qquad R_{1.2345} = .827$$

TABLE 23

COMPARISON OF SEPARATE DETERMINATION, PARTIAL CORRELATION AND BETA COEFFICIENTS

	Variables	d	Partial	Beta	Beta Squared
X_2	Inversion Intensity	.0655	.0554	.1325	.0175
X_3	Wind Velocity	.0882	.2758	− .5608	.3145
X_4	Visibility at Noon	.0672	.4992	− .4682	.2920
X_5	Deviation of Temperature above Daily Mean	.4638	.5826	1.1250	1.2650

$R^2_{1\ 2345}$ = Total Determination = .6847, or 68.47%

$a_{1.2345} = \bar{X}_1 - b_{12.345}\bar{X}_2 - b_{13.245}X_3 - b_{14.235}X_4 - b_{15.231}X.$

$a_{1.2345} = 12.359 - (- 1.513) = 13.872.$

$X_1 = a_{1.2345} + b_{12\ 345}X_2 + b_{13.245}X_3 + b_{14\ 235}X_4 + b_{15.234}X_5$

$X_1 = 13.872 + .1152907X - .1141530X - .4322654X + 1.030942X$

BRIEF SUMMARY OF THE SMOG PROBLEM

1. Inversion intensity (X_2) measures the rate or lack of up-flow or rise of the air near the earth. Often the temperature of the air 1000 feet or so above the surface in the Los Angeles area is warmer than the ground temperature, preventing an upward movement of air. The higher the inversion intensity is, the greater the smog.

2. Wind velocity (X_3) is often quite slow in the Los Angeles

Basin. If the wind is strong, it carries away more of the smog. Its relationship to smog is negative. The slower the wind movement is, the more smog.

3. Visibility at noon (X_4) is a fair index of the accumulation of smoke and fumes in the air. This variable is related to both X_2 and X_3. The lower the visibility is, the greater is the smog.

4. Deviation of temperature above daily mean (X_5) is the strongest active or positive factor in the problem. It ranks above the other factors in all partial measures. The hotter the sun is, the more active are the various poisonous chemical reactions in the accumulated smoke and fumes, and the more violent the smog condition is.

MULTIPLE CURVILINEAR CORRELATION

The formulas and methods presented in this chapter up to this point are all straight-line equations and methods. As has already been shown many relationships are curvilinear. Straight-line methods do not give the best results when the relationships are truly curvilinear.

Mathematical Curvilinear Methods

Several formulas for curvilinear regression were presented in the previous chapter. Any combination of them may be used in curvilinear correlation problems. The straight-line multiple regression line

FORMULA NO. 96

$$X_1 = a + b_{12.34\ldots n}X_2 + b_{13.24\ldots n}X_3 + \cdots b_{1n\ldots N-1}X_n$$

may be expanded to fit any set curvilinear factors by combinations such as the following examples:

EXAMPLE 1

1. Relationship between X_1 and X_2 simple parabola
2. Relationship between X_1 and X_3 straight line
3. Relationship between X_1 and X_4 cubic parabola

Such a regression line would be:

$$X_1 = a + b_{12.34\ldots n}X_2 + b'_{12.34\ldots n}X_2^2 + b_{13.24\ldots n}X_3 + b_{14.23\ldots n}X_4 + b'_{14.23\ldots n}X_4^2 + b''_{14.23\ldots n}X_4^3.$$

Such an equation contains seven unknowns and is very long and tedious to compute.

EXAMPLE 2

1. Relationship between X_1 and X_2 parabolic log line
2. Relationship between X_1 and X_3 cubic log parabola
3. Relationship between X_1 and X_4 simple log line

Such relationships would be expressed by the equation:

$$\text{Log } X_1 = \text{Log } a + b_{12}\text{Log } X_2 + b'_{12}\text{Log } X_2^2 + b_{13}\text{Log } X_3 + b'_{13}\text{Log } X_3^2 + b''_{13}\text{Log } X_3^3 + b_{14}\text{Log } X_4$$

Usually it is a waste of time to compute such expensive mathematical lines in practical problems unless the statistician knows before the work is done that such lines fit the data well. Such lines are definitely beyond the limits of an elementary text. The references at the end of this chapter are adequate to guide students who may wish to study such regression lines.

FREEHAND MULTIPLE CORRELATION

Quite dependable freehand methods for estimating multiple and practical correlation have been developed during the past fifteen years. Although they are much briefer and less expensive than mathematical curvilinear methods they are too complex and advanced to be included at length in an elementary text.[1]

Multiple and Partial Correlation combined, form one of the most powerful and useful methods of advanced statistical research available. If the elementary student does not master them, he would do well to become acquainted with their basic

[1] Ezekiel, M., *Methods of Correlation Analysis*, Chapters 14, 15, 16, and 17, gives an adequate presentation of these methods. Also, Pearson, F. A., and Bennett, K. R., in *Statistical Methods Applied to Agricultural Economics*, Chapter 13, gives a sufficient presentation of the methods.

principles and more simple methods as a means of appreciating and interpreting in a more intelligent way much of the material with which he will be confronted in many textbooks and corporation and governmental reports. Such a knowledge is useful from the standpoint of consuming as well as of producing statistics.

LIMITATIONS AND USES OF MULTIPLE AND PARTIAL CORRELATION

Straight-line multiple correlation assumes that (1) the relationships among the variables are linear, and (2) are additive. Often neither assumption is entirely corrrect. As has been indicated earlier many, perhaps most, relationships in the social and biological as well as in the fields of the physical sciences are *not* straight line. In some cases the deviations from linear equations are not sufficient to invalidate the results; in others they are. The researcher should be well enough informed in his field of study to know whether linear relations are sufficient, or whether he must use curvilinear equations. Multiple curvilinear equations are long and laborious to use and will often be avoided for this reason unless they are necessary. The methods of computing them are not given in this text but may be found in several more advanced texts.[1]

The second assumption, that the relationships are additive, often does not hold. The assumption of additive relationships means that each variable increases or decreases without being affected by the other variable, just as the total weight of a loaded truck is equal to the weight of the truck plus the weight of the load. Combining such factors does not increase their total. In the fields of the social and biological sciences this assumption often does not hold. In many cases when two or more independent factors are brought together they react on each other. Fertilizer and rainfall are an example. Since a given amount of

[1] Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agriculture*, John Wiley & Sons, Inc., New York, 1942; Ezekiel, M., *Multiple Correlation Analysis*, John Wiley & Sons, Inc., New York, 1941; and Mills, F. C., *Statistical Methods, Revised*, Henry Holt and Co., New York, 1938.

moisture is required to dissolve the fertilizer and bring its elements to the roots of the plant, a small amount of rain may make little or none of the fertilizer available, but a little more rain may make it all available. The two factors support and multiply each other. Fire and gunpowder are perhaps a violent illustration of joint action. It is a problem of which the statistician must always take cognizance.

Multiple correlation can be effectively used on samples of smaller size than is possible with tabular analysis; but as the number of variables included in the problem increases, the size of the sample must increase in order to hold the standard error to a safe minimum. For four variables a sample of at least 50 items is usually necessary for dependable results and a still larger sample would be preferable. Much of the unfavorable criticism to which the results of multiple correlation have been subjected is due to the use of samples which were too small and to the additional fact that the studies were not carefully and logically planned. The more powerful a research tool is, the greater must be the care with which it is used. Multiple and partial correlation, if carefully and logically applied to samples of adequate quality and size in problems which have been properly planned, is one of the most powerful and adequate engines of analysis available to the statistician. It is not a device for novices. To obtain good results the researcher must be thoroughly informed in the entire field of science in which his problem lies.

The standard error of the coefficient of multiple correlation is computed from

FORMULA No. 97

$$\sigma_R = \frac{1 - R^2}{\sqrt{N - M}}$$

See Formula Nos. 100 and No. 102, Chap. 13.

For partial correlation coefficients the formula is similar.

FORMULA No. 98

$$\sigma_{r12.34\ldots n} = \frac{1 - r^2_{12.34\ldots n}}{\sqrt{N - M}}$$

In both formulas the symbol M equals the number of independent variables included in the problem.

Multiple and partial correlation have found wide use in the various fields of agricultural science. They are also widely used in the testing of variables to be included in indexes. Sociology, psychology, economics, education, and market analysis and price forecasting are among the larger fields in which they have proved useful.

SUMMARY

1. The coefficient of multiple correlation expresses the total combined relationship between a group of independent variables and one dependent factor.

2. Partial correlation expresses the degree of net relationship between a dependent factor and each one of a group of independent variables after the influence of the other independent factors has been taken into consideration and computed out or removed.

3. Since most effects are the result of a number of factors, multiple and partial correlation analysis is essential to obtain an accurate measure of the influence of each separate factor in the total complex result.

4. The symbol of straight-line multiple correlation is R, and of curvilinear multiple correlation is capital Rho, P.

5. To prevent errors in the computation of the long net regression and correlation equations a *check sum* device is important.

6. The Doolittle method of solving simultaneous equations is especially helpful in the solution of such long multiple equations.

7. The beta coefficients, based on the standard deviations of the variables and the net regression coefficients are an alternative device for computing the net effect of each independent factor.

8. The beta coefficients will not be identical with the partial coefficients, but will rank in the same order in size.

9. The several coefficients of separate determination will equal the coefficient of total determination but are not dependable when the intercorrelation between two or more of the independent variables is high.

10. Multiple and partial determination or correlation are highly effective and powerful devices for statistical analysis, but like any other complicated mechanism require a maximum of exact knowledge and skill in their effective use. Any problem to which they are applied should be carefully planned and should rest on sufficient dependable data.

11. As the number of variables included in a multiple correlation problem increases, the number of items in the sample must increase if dependable results are to be obtained.

SELECTED REFERENCES

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XXIV. Prentice-Hall, Inc., New York, 1948.

Ezekiel, M., *Methods of Correlation Analysis*, Chapters 12–17. John Wiley & Sons, New York, 1941.

Mills, F. C., *Statistical Methods, Revised*, Chapter XVI. Henry Holt and Co., New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapters 10, 11, 12, 13. John Wiley & Sons, New York, 1942.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, McGraw-Hill Book Co., New York, 1940.

REVIEW QUESTIONS

1. What is the difference between simple and multiple correlation?
2. Why should multiple correlation be studied by beginners in statistics?
3. What is meant by "weighting" the independent variables?
4. Why must the variables be stated in "standard measures" or in terms of their standard deviations in multiple correlation?
5. What are the advantages of the Doolittle Method of solving equations?
6. What is separate determination? What are its advantages and its weaknesses?
7. What is partial correlation? Exactly what does it measure? Explain.
8. What is the "beta" coefficient? What does it measure? Explain in detail.

EXERCISE

Problem in Multiple and Partial Determination and Regression [1]

State	Av. Ann. Temp. through 1947 X_2	% of Pop. Urban 1950 X_3	Per Pupil School Expenditure 1947 X_4	Per Capita Investment in Plant and Equipment, 1947 X_5	Per Capita Income 1947 X_1
Conn.	50	78	$169	$74	$1,671
N.Y.	50	85	209	38	1,781
W.Va.	55	35	100	46	1,031
Ala.	66	43	68	25	837
Vt.	45	36	122	29	1,183
Del.	55	63	152	70	1,646
Ark.	62	33	65	18	710
N.C.	60	34	81	36	890
Okla.	60	51	112	16	930
Calif.	63	81	169	43	1,643
Wyo.	45	49	171	42	1,472
Mont.	44	44	215	14	1,641
Iowa	50	48	143	27	1,144
Miss.	66	28	46	11	659
Ga.	64	45	64	26	885
Minn.	45	54	155	30	1,195
Mich.	48	71	144	73	1,424
Ky.	57	37	86	29	850
Kans.	57	52	141	22	1,315
Ind.	53	60	144	82	1,287
Idaho	51	43	116	21	1,290
N.D.	41	27	154	3	1,678
Ore.	47	54	155	48	1,253
Texas	66	63	118	47	1,128
Mass.	50	84	180	46	1,449

Source: *Statistical Abstract of U.S.*, 1949.

[1] The instructor may vary the assignment of this problem in the following ways:

1. Length. Assign 12, 15, 18 or 20 states, or all.
2. Width. Assign (1) X_2, X_3, X_4 and X_1, or (2) X_3, X_4, X_5 and X_1 or (3) X_2, X_4, X_5 and X_1, or (4) X_2, X_3, X_5 and X_1. Any one of the assignments suggested above is of sufficient length for class practice and by varying the assignments in this way these data may be used for several semesters.

CHAPTER 12

TABULAR ANALYSIS

A cross-classification table is the simplest and easiest device for discovering the more simple and apparent relationships which exist among variables. Tabular analysis is not the making of tables. It is the *use of tables as analytical devices* to discover and measure statistical relationship. If one variable changes in any constant or relative proportion to changes in another variable, this quantity can be measured by dividing one variable into class intervals and computing the averages of the associated values of the other variable for each class interval. The mathematics required for this type of analysis is quite simple, consisting almost entirely of totals and simple averages. The principles upon which cross-classification tables are constructed were explained in detail in Chapter 3. A brief review of those materials at this point will aid the student in the present analysis. The easiest way for the beginning student to comprehend this method is to follow through the successive stages of a complete problem.

DATA TABLES

Tables 24–32 inclusive reveal the steps required. Three types of tables are used. The first is the *data table*, which includes the material to be analyzed. Table 24 illustrates this method of assembling data. Figures for 85 farms in the hard winter wheat area were collected for (1) size of farm in acres, (2) number of milk cows, (3) wheat yields in bushels, and (4) rate of income. In Table 24 these data are recorded by farms as they were received from the schedules without any thought or plan for logical ar-

TABLE 24. DATA TABLE

FARM DATA ON SIZE OF FARM, NUMBER OF COWS, YIELD OF WHEAT, AND RATE OF INCOME USED IN ILLUSTRATING THE THEORY AND METHODS OF TABULAR ANALYSIS

No.	Size Acre	Milk Cows	Wheat Yield	Rate of Income	No.	Size Acre	Milk Cows	Wheat Yield	Rate of Income
1	160	7	12	1.6	44	180	2	20	2.5
2	160	7	14	.1	45	402	4	17	10.5
3	160	4	13	2.8	46	480	6	20	7.7
4	160	4	18	1.0	47	450	1	14	3.7
5	160	5	16	3.0	48	420	1	26	11.0
6	160	4	20	2.3	49	640	2	12	3.0
7	160	4	15	1.9	50	260	4	12	4.8
8	160	4	11	1.4	51	395	42	26	14.7
9	160	5	14	2.9	52	220	9	16	3.5
10	160	2	11	.2	53	325	16	23	7.2
11	160	5	13	6.5	54	520	8	17	9.0
12	160	5	31	10.5	55	400	13	16	4.2
13	160	6	15	5.7	56	285	18	27	16.9
14	160	3	17	9.2	57	297	14	22	9.7
15	160	2	16	8.0	58	307	11	20	5.2
16	190	2	9	.6	59	215	8	20	6.5
17	230	6	11	1.8	60	220	1	21	8.3
18	400	2	14	5.6	61	255	9	22	10.0
19	260	1	13	2.3	62	225	3	20	5.2
20	220	13	24	17.8	63	400	8	25	14.6
21	250	8	15	3.8	64	135	3	20	5.1
22	271	2	20	9.1	65	148	7	14	4.0
23	303	7	15	1.4	66	360	4	22	6.2
24	165	5	16	8.9	67	280	6	23	10.2
25	380	8	19	13.7	68	480	6	25	7.4
26	150	6	20	6.6	69	305	7	18	5.6
27	250	6	18	7.9	70	400	9	23	10.6
28	390	9	18	8.3	71	165	6	12	1.8
29	310	5	14	2.0	72	210	9	10	.8
30	260	4	15	6.5	73	304	6	10	1.1
31	480	12	11	2.8	74	250	7	24	5.0
32	300	6	27	8.0	75	200	5	14	4.5
33	155	6	18	8.9	76	230	6	25	13.2
34	180	9	22	8.3	77	160	4	22	5.7
35	450	8	14	6.4	78	320	4	26	8.4
36	320	5	18	5.4	79	160	16	11	5.4
37	320	10	25	7.9	80	160	4	10	1.0
38	320	10	14	2.6	81	160	6	12	3.5
39	240	14	18	2.5	82	160	6	12	2.0
40	400	19	13	5.0	83	160	2	17	10.2
41	262	12	22	6.2	84	160	6	11	4.3
42	180	5	20	6.6	85	160	9	14	3.0
43	232	6	12	3.1					

rangement. For tabular analysis this method and order of recording the data for the first time is quite sufficient. Regardless of the order in which the data first appear they must be completely rearranged during the process of analysis.

ORGANIZATION TABLES

The heart of the whole process of tabular analysis is the *organization table.* In complex or long problems there may be several such tables. In this example three are sufficient. As is emphasized in Chapter 23 on "Organizing a Statistical Problem," a large amount of thought should be given to the form and contents of each organization table before the schedules are printed or the data collected. The statistician should decide in planning his study what data he wishes to collect and the exact form of organization table in which he expects to organize them. An organization table is an analytical device. It may be as simple or as complex as its purpose demands. For two variables a simple one-way table such as Table 25 is sufficient.

Class Intervals of Independent Variable

The first point in designing an organization table after the data are at hand is to decide on the number and width of class intervals for the independent variable or variables. Table 25 is set up in the simplest form with only two variables. Wheat yields are the independent variable, and the class intervals are chosen for it. These may be of uniform or of varying width. The range in this case is from 9 to 31 bushels inclusive, or 23 bushels. This range in this case is divided into three class intervals as follows: (1) 14 bushels and under, (2) 15–21 bushels, and (3) 22 bushels and over. This gives a sufficiently equal distribution of the data among the three classes. There would be no objection to having four or even five or more classes. Three were chosen in this case to show a brief simple analysis. A larger number would ordinarily be preferred because they would reveal the relationship in greater detail. However, too many classes make a table confusing and unwieldy.

TABLE 25. ONE-WAY ORGANIZATION TABLE

ORGANIZATION OF THE DATA ON THE TWO VARIABLES, WHEAT YIELD AND RATE OF INCOME TO REVEAL THE CROSS RELATIONSHIP BETWEEN THEM

14 bu. and Under		15 bu. to 21 bu.		22 bu. and Over	
Wheat Yield	Rate of Farm Income	Wheat Yield	Rate of Farm Income	Wheat Yield	Rate of Farm Income
12	1.6	18	1.0	31	10.5
14	0.1	16	3.0	24	17.8
13	2.8	20	2.3	27	8.0
11	1.4	15	1.9	22	8.5
14	2.9	15	5.7	23	7.9
11	0.2	17	9.2	22	6.8
13	6.5	16	8.0	26	11.0
9	0.6	15	3.8	26	14.7
11	1.8	20	9.1	23	7.2
14	5.6	15	1.4	27	16.9
13	2.3	16	8.9	22	9.7
14	2.0	19	13.7	22	10.0
11	2.8	20	6.6	25	14.6
14	6.4	18	7.9	22	6.2
14	2.6	18	8.3	23	10.2
13	5.0	15	6.5	25	7.4
12	3.1	18	8.9	23	10.6
14	3.7	18	5.4	24	5.0
12	3.0	18	2.5	25	13.2
12	4.8	20	6.6	22	5.7
14	4.0	20	2.5	26	8.4
12	1.8	17	10.5		
10	0.8	20	7.7		
10	1.1	16	3.5		
14	4.5	17	9.0		
11	5.4	16	4.2		
10	1.0	20	5.2		
12	3.5	20	6.5		
12	2.0	21	8.3		
11	4.3	20	5.2		
14	3.0	20	5.1		
		18	5.6		
		17	10.2		
Total 381	90.6	589	204.2	510	210.3
Means 12.3	2.9	17.8	6.2	24.3	10.0

Paired Values Under Each Class Interval

In Table 25, under each class interval of wheat yields is a double column of figures, the left-hand column of which contains the wheat yield of the farms falling in that class and the right-hand column of which contains the rate of income for the identical farms. Each column is totaled and simple averages computed. These paired averages are the basis for comparing and measuring the relationship between the two variables of wheat yields and rate of income for these 85 farms.

TABLE 26

SUMMARY TABLE

PARTIAL SUMMARY OF RELATIONS BETWEEN WHEAT YIELDS AND RATE OF INCOME AS REVEALED IN CONSTRUCTION TABLE NO. 18

Class Intervals for Wheat Yields	Frequencies No. of Farms	Totals for Dependent Factor, Farm Incomes	Average Rate of Farm Income
14 bu. and under	31	90.6	2.9
15 bu. to 21 bu.	33	204.2	6.2
22 bu. and over	21	210.3	10.0
Totals	85	505.1	5.9

SUMMARY TABLES

The third type of table used in tabular analysis is the *summary table*. It is composed of materials, class intervals, totals, averages and frequencies drawn from the organization table. Table 26 consists of four columns: (1) class intervals of wheat yields, (2) frequencies, or the number of farms in each class, (3) the total of the dependent factor for each class, and (4) average of dependent factor for each class.

This summary table reveals considerable information about the relationship between wheat yields and rate of farm income. Farms with yields of 14 or fewer bushels per acre have an unweighted average rate of income of only 2.9%, which is not a very

profitable enterprise. Farms with yields of 15 to 21 bushels per acre have on the average a return of 6.2%, which is more than twice as large as the low yield farms. Farms with yields of 22 or more bushels per acre have on the average an income of 10.0% or more than three times that of the small production farms. This simple table reveals relationships which a farm economist would easily comprehend. Since the costs of preparing the soil, sowing the crop and harvesting are approximately the same per acre regardless of yield, it follows that large yields are much more profitable than small ones. From such a table it is possible to predict on the average the rate of income for farms within any given class.

TABLE 27

SUMMARY TABLE

FINAL SUMMARY OF AVERAGE RELATIONSHIP BETWEEN WHEAT YIELDS IN BUSHELS AND RATE OF INCOME ON 85 OKLAHOMA FARMS, 1939–1940

Wheat Yields Average per Class	Farm Income in Per Cent
12.3	2.9
17.8	6.2
24.3	10.0

Summary Table 27 is a still more brief but no less meaningful presentation of significant relationships. In this case the comparison is between the average wheat yield of each class and the average rate of income for the same class. It is a positive relationship in which the rate of income increases in approximately the same ratio as the increase of wheat yields. Figure 71 indicates graphically the significance of this relationship.

If it is desired to bring two independent variables into the analysis at once, a *two-way organization table* is necessary. In this case in addition to wheat yields the influence of size of farms on rate of income is taken into the analysis. The inclusion of two variables increases the complexity of the organization table.

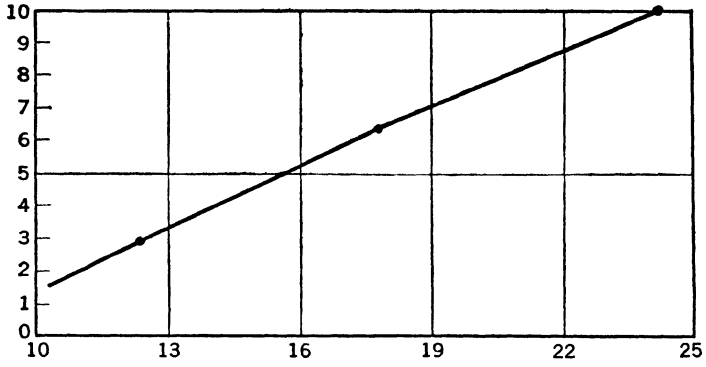


FIG. 71. Relationship between yield of wheat and rate of income for 85 farms as shown in Summary Table 27

Class intervals must be chosen for the additional variable. In this case farm sizes are divided into three class intervals as follows: (1) Farms under 200 acres, (2) Farms of 200 acres to 320 acres inclusive, and (3) Farms of more than 320 acres. These class intervals are laid out across the top of the organization table. Under each class of the top variable a complete set of class intervals for the second variable is placed. This second set of class intervals must be repeated over and over in exactly the same form under each class of the top variable, if a complete cross classification of the data is to be organized. This is the point that usually gives beginning students the greatest difficulty. *The exact necessary relationships should be noted carefully.*

At this point in a two-way organization table three alternatives are presented.

First Form. The values of the dependent variable only may be recorded in the appropriate class-interval columns. That means in this case that *rates of income* only would be recorded in the table. There would be only one column of figures under each class interval and they would be the rate of income figures. The first three figures of the first column would be 1.6, 0.1, 2.8. The first three figures of the second column would be 1.0, 3.0, 2.3. Only the % column would appear. The *bu.* column would be absent. For many purposes this type of table would be sufficient

TABLE 28. TWO-WAY ORGANIZATION TABLE

ORGANIZATION OF THE DATA ON SIZE OF FARMS, WHEAT YIELDS, AND RATE OF INCREASE TO REVEAL THE CROSS RELATIONSHIP AMONG THE THREE VARIABLES

	Farms Under 200 Acres						Farms 200 Acres to 320 Acres Inclusive						Farms Over 320 Acres					
	Wheat Yields 14 Bu. and Under		Wheat Yields 15–21 Bu.		Wheat Yields 22 Bu. and Over		Wheat Yields 14 Bu. and Under		Wheat Yields 15–21 Bu.		Wheat Yields 22 Bu. and Over		Wheat Yields 14 Bu. and Under		Wheat Yields 15–21 Bu.		Wheat Yields 22 Bu. and Over	
	Bu.	%	Bu.	%	Bu.	%	Bu.	%	Bu.	%	Bu.	%	Bu.	%	Bu.	%	Bu.	%
	12	1.6	18	1.0	31	10.5	11	1.8	15	3.8	24	17.8	14	5.6	19	13.7	26	11.0
	14	0.1	16	3.0	22	8.3	13	2.3	20	9.1	27	8.0	11	2.8	18	8.3	26	14.7
	13	2.8	20	2.3			14	2.0	15	1.4	25	7.9	14	6.4	17	10.5	23	7.2
	11	1.4	15	1.9	22	5.7	14	2.6	15	6.5	22	6.2	13	5.0	20	7.7	25	14.6
	14	2.9	15	5.7			12	3.1	18	5.4	27	16.9	14	3.7	17	9.0	22	6.2
	11	0.2	17	9.2			12	4.8	18	2.5	22	9.7	12	3.0	16	4.2	25	7.4
	13	6.5	16	8.0			10	0.8	16	3.5	22	10.0					23	10.6
	9	0.6	16	8.9			10	1.1	20	5.2	23	10.2						
	14	4.0	20	6.8			14	4.5	20	6.5	24	5.0						
	12	1.8	18	7.9					21	8.5	25	13.2						
	11	5.4	18	8.9					20	5.2	26	8.4						
	10	1.0	20	6.6					18	5.6								
	12	3.5	20	2.5														
	12	2.0	20	5.1														
	11	4.3	17	10.2														
	14	3.0																
Totals	193	41.1	266	88.0	75	24.5	110	23.0	216	63.2	267	113.3	78	26.5	107	53.4	170	71.7
Number	16	16	15	15	3	3	9	9	12	12	11	11	6	6	6	6	7	7
Means	12.1	2.6	17.7	5.3	25	8.2	12.2	2.6	18.0	5.3	24.3	10.3	13.0	4.4	17.8	8.9	24.3	10.3

Second Form. If it is desired to show in the summary table a more exact measure of the relationship between one of the independent variables and the dependent variable, a *double column* including the figures for the *two* variables desired must be included under each class interval. In Table 28 these are (bu.) and (%). This form of organization makes it possible to obtain totals and averages for both variables and to compute the ratios between them. Some of these results are shown in Summary Tables 29 and 30.

Third Form. If it is desired to compute such summary ratios and detailed relationships between the dependent and both of the independent variables, the figures of the original data for both independent variables must be included under each class interval. For such a computation for the above data there would appear under each class of wheat yields, three sub-columns as follows: (acres), (bu.), (%). With such an organization table three cross summaries could be shown, (1) Size of Farm to Wheat Yields, (2) Size of Farm to Rate of Income, and (3) Wheat Yields to Rate of Income. Such an organization table is *not* included here because it would be too large to place on one page of the book, but it is of little greater difficulty to make than Table 28. It would have to be at least 50% wider than Table 28.

Summary Table 29 shows the average rate of income as in-

TABLE 29

SUMMARY TABLE

SUMMARY OF THE RELATIONSHIPS BETWEEN RATE OF INCOME AND WHEAT YIELD, AND BETWEEN RATE OF INCOME AND SIZE OF FARM, SHOWING COMBINED EFFECT OF TWO INDEPENDENT VARIABLES ON INCOME

Wheat Yields Class Intervals	Size of Farm Units			Averages
	Under 200 Acres	200 to 320 Acres	Over 320 Acres	
14 bu. and under	2.6	2.6	4.4	2.9
15 bu. to 21 bu.	5.3	5.3	8.9	6.2
22 bu. and over	8.2	10.3	10.3	10.0
Averages	4.5	6.2	8.0	5.9

fluenced by yields of wheat and size of farms. Wheat yields are the most important factor in income for these farms. For low yields the rate is only 2.9%, while for high yields it is 10.0%, or a range of 7.2%. On the other hand trebling the size of farms increases the income only from 4.5% to 8.0% or a range of 3.6%. Yield, therefore, has approximately twice as much effect on income as does size of farm.

Computing Totals and Averages for Summary Tables

The averages for a summary table are computed from the totals in the sub-classes of the organization table as follows: (1) Add the totals of all sub-classes for each class interval. Example:

CLASS INTERVAL

WHEAT YIELDS 14 BU. AND UNDER

Farms	Sub-class Totals (bu.)	Sub-class Totals (%)
16	193	41.1
9	110	23.0
6	78	26.5
31	381	90.6

$$\frac{381}{31} = 12.3 \qquad \frac{90.6}{31} = 2.9$$

All the other averages of the summary tables were obtained from the organization table by the same method. This method gives a weighted average for the summary table averages.

TABLE 30. SUMMARY TABLE

VARIATIONS IN WHEAT YIELDS FOR FARMS OF VARIOUS SIZES

Wheat Yields Class Intervals	Size of Farm Units			Averages
	Under 200 Acres	200 to 320 Acres	Over 320 Acres	
14 bu. and under	12.1	12.2	13.0	12.3
15 bu. to 21 bu.	17.7	18.0	17.8	17.8
22 bu. and over	25.0	24.3	24.3	24.4

Summary Table 30 shows average variations in wheat yields for size of farms and classes of yields. It is merely a further indication of the various types of relationships which may be drawn out of a well-designed organization table.

TABLE 31. THREE-WAY ORGANIZATION TABLE

ORGANIZATION OF THE DATA FOR THREE INDEPENDENT VARIABLES, WHEAT YIELD, NUMBER OF COWS AND SIZE OF FARMS, AS RELATED TO THE RATE OF INCOME OF 85 OKLAHOMA FARMS

	Farms with 6 Cows or Less								
	Under 200 Acres			200 to 320 Acres			Over 320 Acres		
	14 Bu and Under	15 to 21 Bu.	22 Bu. and Over	14 Bu. and Under	15 to 21 Bu.	22 Bu. and Over	14 Bu. and Under	15 to 21 Bu.	22 Bu. and Over
	2 8	1 0	10 5	1 8	9.1	8.0	5.6	10.5	11.0
	1.4	3 0	5 7	2 3	6.5	8.4	3.7	7.7	7.4
	2 9	2 3		2 0	5 4	10 2	3.0		6.2
	0 2	1 9		3 1	5.2	13.2			
	6 5	5.7		1.1	8.3				
	0 6	9 2		4 5					
	1 8	8 0		4.8					
	1 0	8 9							
	3 5	6 6							
	2.0	7.9							
	4.3	8 9							
		6 6							
		2 5							
		5 1							
		10 2							
Total	27 0	87 8	16 2	19 6	34 5	39.8	12 3	18 2	24.6
Means	2 5	5 9	8 1	2 8	6.9	10.0	4.1	9.1	8 2

	Farms with More Than 6 Cows								
	Under 200 Acres			200 to 320 Acres			Over 320 Acres		
	14 Bu. and Under	15 to 21 Bu.	22 Bu and Over	14 Bu and Under	15 to 21 Bu.	22 Bu. and Over	14 Bu. and Under	15 to 21 Bu.	22 Bu. and Over
	1.6		8 5	2 6	3.8	17.8	2 8	13.7	14.7
	0.1			0.8	1.4	7 9	6.4	8.3	7.2
	4.0				2.5	6.2	5.0	9.0	14.6
	5.4				5.8	16.9		4.2	10.6
	3.0				6.5	9.7			
					5.6	10.0			
					3.5	5.0			
Total	14.1		8.5	3.4	29.1	73 5	14.2	35.2	47 1
Means	2.8		8.5	1.7	4.4	10 5	4.7	8.8	11.8

If one wished to exhaust the possibilities of analyzing the data on farms in Data Table 24 it would be necessary to use a *three-way organization table*, such as Table 31.

The fundamental principle is laid down at this point, that *the complexity of organization tables increases in proportion to the product of the number of class intervals for each variable.* This principle may be illustrated by Organization Tables 25, 28, and 31. In Table 25 there are 6 columns, 3 class intervals and 2 sub-classes, or $3 \times 2 = 6$ columns. In Table 28 there are 3 class intervals for farm sizes, 3 class intervals for wheat yields, and 2 sub-classes.

$$3 \times 3 \times 2 = 18 \text{ columns}$$

In Table 31, there are 2 classes for number of cows, 3 classes for size of farms, 3 classes for wheat yields, and 1 sub-class.

$$2 \times 3 \times 3 \times 1 = 18 \text{ columns}$$

If there had been 3 classes for cows, and four sub-classes for the items of data, the requirement for such a complete organization table would have been,

$$3 \times 3 \times 3 \times 4 = 108 \text{ columns}$$

It is, of course, possible to make any number of class intervals desired. If there had been only 4 classes for each variable, the result would have been

$$4 \times 4 \times 4 \times 4 = 256 \text{ columns}$$

for the organization table, a table so large that it would cover the entire top of a large desk. For this reason, there is a practical necessity for keeping both the number of class intervals and the number of variables used in tabular analysis to a workable minimum.

Another cogent reason for not going beyond four variables in tabular analysis is that the relationships among the variables become so complex that it is increasingly difficult for the statistician to discover and measure and explain them all adequately in any manageable space or time. This method is easy and effective within narrower limits but becomes unmanageable when expanded too far.

A three-way organization table is usually best set up in the form of Table 31 in which two variables are placed one above the other across the table, or on the X-axis, and the third variable is placed on the side, or on the Y-axis. In Table 31 size of farms is placed in three classes across the entire table. Wheat yields are placed in three classes, under each of the farm size classes. This arrangement takes care of two variables. The third variable, number of cows, is introduced by duplicating the above table so that the upper section is marked "6 cows or less" while the lower section is marked "over 6 cows." If we wished to divide cows into three classes, it would be necessary to make still another section of the table below with the third class of cows as a heading. By this process a table can be expanded indefinitely on both axes.

Many types and models of summary tables may be developed for such a complex organization table as Table 31. Table 32 is one of the simpler general forms. It shows the relationships between (1) wheat yields and rate of income, (2) size of farms and rate of income, and (3) number of cows and rate of income for class intervals and totals. If two or more sub-classes for data had been included under the wheat yield class intervals a summary table could have been set up for relationship between (1) cows and wheat yields, (2) cows and size of farms and (3) size of farms and wheat yields.

From Table 32 it is evident that the number of cows on a wheat farm does increase its income, but to a smaller degree than size of farm or wheat yields. The measurement of the influence of number of cows on income is clouded in this table by the fact that it does not appear to be uniform. For some sizes of farms and for some class yields it is smaller. For others in which one would expect a higher rate, the income is smaller. As an illustration of this variation one may note that the larger farms with "over 6 cows" have a higher rate than smaller farms with "6 cows or less." The rates are 2.5 and 2.8. For medium size farms, the rates are reversed. They are 2.8 and 1.7. Do more cows increase the income on some farms and decrease it on others? If so, what are the reasons and does this table reveal them? It does not.

TABLE 32

SUMMARY TABLE

SUMMARY OF THE INTERRELATIONSHIPS BETWEEN THE THREE INDEPENDENT VARIABLES OF WHEAT YIELD, NUMBER OF COWS AND SIZE OF FARMS AND THE DEPENDENT VARIABLE OF RATE OF INCOME FOR 85 OKLAHOMA FARMS, 1939-1940

	Wheat Yields			Totals
	14 Bu. and Under	15 to 21 Bushels	22 Bu. and Over	
6 cows or less	2.8	6.4	9.0	5.4
Under 200 acres	2.5	5.9	8.1	4.7
200 to 320 acres	2.8	6.9	10.0	5.9
Over 320 acres	4.1	9.1	8.2	6.9
Over 6 cows	3.2	5.9	10.8	6.8
Under 200 acres	2.8		8.5	3.8
200 to 320 acres	1.7	4.4	10.5	6.6
Over 320 acres	4.7	8.8	11.8	8.8
Average of Totals	2.9	6.2	10.0	5.9

LARGE SAMPLES FOR LARGE TABLES

It will be noted that in Table 32 there is one cell, or location, which is vacant. There were no farms in this sample with less than 200 acres and a yield of 15 to 21 bushels that had over 6 cows. Why does this vacancy occur? *This sample of 85 farms is too small to be spread over so large and complex a table. Tabular analysis is definitely a large sample analysis.* It is not effective with small groups of data. In order to work out tables such as Table 31 and Table 32 with dependable results there should be 200 or more items in the sample, 400 or 500 would be still better. In one cell of Table 31 there is a clear vacancy. Another cell has only one item. Three more cells have only two items. Only two cells, the first and second, have sufficiently large samples to be dependable. A much better sample would have been 500 farms.

TABLE 33

DATA TABLE

DATA ON ELEVATION, ANNUAL MEAN TEMPERATURE, MEAN RELATIVE HUMIDITY, AND MEAN ANNUAL PRECIPITATION OF 61 AMERICAN CITIES, THROUGH 1947 *

Elevation above Sea Level (ft.)	Mean Annual Temperature	Mean Annual Relative Humidity	Mean Annual Precipitation (in.)	Elevation above Sea Level (ft.)	Mean Annual Temperature	Mean Annual Relative Humidity	Mean Annual Precipitation (in.)
10	67.5	72	62	4,289	48.8	52	9
201	65.6	67	52	8	52.8	75	41
1,083	70.3	38	8	5,314	55.2	47	8
257	62.2	66	48	277	48.1	72	37
277	63.3	52	9	10	52.8	65	43
312	63.0	62	15	513	47.7	68	32
52	56.5	73	22	2,192	55.5	68	38
5,221	50.4	48	14	400	60.3	72	46
72	55.5	67	42	1,670	41.1	66	16
18	69.3	73	51	787	48.7	61	33
11	75.2	73	59	1,254	60.1	66	32
1,054	62.3	69	49	30	53.1	70	42
2,858	51.1	52	13	351	52.6	64	38
611	49.5	69	33	1,286	52 6	67	36
793	53.0	66	40	43	66.2	72	49
800	49.9	71	32	1,282	44.5	66	20
1,380	56.6	62	30	485	59.6	69	46
459	57 0	65	42	3,590	54 6	55	21
8	69.6	72	60	3,920	63 8	38	9
32	41 8	77	39	688	65.7	63	32
15	49.6	68	41	51	69.2	73	46
619	48.7	71	31	4,222	51.0	57	14
721	38.1	77	30	331	44.7	77	32
830	45.0	64	27	11	59.8	72	45
234	65.8	73	52	162	58.2	68	42
908	55.3	66	36	14	52.0	73	33
465	56.3	64	39	1,954	48.6	60	16
3,893	43.8	56	13	615	54.6	72	39
2,629	45.8	61	13	938	46.1	71	31
2,805	49.5	62	18	6,139	44.6	53	15
978	51.1	67	28				

* Source: *Statistical Abstract of U.S.* 1949, pp. 145–155.

Second Illustrative Problem

It is difficult to bring out all phases and possibilities of the method of tabular analysis in one problem. This second illustration falls in the physical rather than the economic realm and includes data on (1) elevation above sea level, (2) mean annual temperature, (3) mean annual relative humidity, and (4) mean annual precipitation for 61 cities of the United States including

all the states except Connecticut, Delaware, Rhode Island, Maryland, and New Hampshire. Space limits the number of analyses which may be obtained from these data, but those employed indicate how broad and versatile this method is and how quickly, simply and accurately the inner relationships of the data may be revealed.

An important part of this method of tabular analysis is the choice and manipulation of class intervals of the independent variables. The fewer the classes the easier is the problem, but the less adequate are the results. In the previous problem no more than three classes were used for any variable. In this study we shall use four or more in some of the tables.

TABLE 34

SUMMARY TABLE

SUMMARY OF RELATIONSHIPS BETWEEN ELEVATION AND PRECIPITATION OF 61 AMERICAN CITIES WITH CLASS INTERVALS OF EQUAL WIDTH, THROUGH 1947

Elevation Class Intervals	*f*	Total Precipitation Per Class Interval	Mean Precipitation Per Class Interval
1 — 1000	41	1597	38.95
1001 — 2000	8	207	25.88
2001 — 3000	4	82	20.50
3001 — 4000	3	43	14.33
4001 — 5000	2	23	11.50
5001 — 6000	2	22	11.00
6001 — 7000	1	15	15.00
	61	1989	32.61

The frequency distribution used in Table 34 based on class intervals of equal width is highly skewed declining from 41 cities in the first class to only one in the seventh. In spite of this defect this table reveals important information about the relation of elevation to precipitation. As elevation increases precipitation declines, but at a decreasing rate for the higher elevations. The two methods used to avoid the difficulty arising from a skewed distri-

bution are (1) To use a wide *open-end* class to include all the small frequencies at the end, such as 3001 and over, (2) To set the classes up disregarding uniformity of width but approximating uniformity of frequencies. This is illustrated by Table 35 in which there are four classes as follows:

TABLE 35

Class	*Frequency*
0 — 100	15
101 — 650	17
651 — 1300	14
1301 — 6200	15

With a larger number of items the class frequencies can be made more nearly equal.

UNIFORM CLASS FREQUENCIES

By dividing the number of cases into terciles, quartiles, quintiles, or other equal distributions, this uneven frequency distribution may be avoided. The possible divisions up to 10 are:

Name	*Number of Classes*
Halves	2
Tertiles	3
Quartiles	4
Quintiles	5
Sextiles	6
Septiles	7
Octiles	8
Noniles	9
Deciles	10

Of course, even a larger number of classes could be used, but as was indicated above, such a division tends to make the problem too detailed and complex and requires too large a sample. For these reasons, halves, tertiles, and quartiles are used most frequently with occasionally a quintile. The more numerous divisions are rare.

In addition to the difficulty of mere size involved in a five or six

class-interval table there is the problem of analyzing the relationships revealed and of computing the measures of error required to indicate the significance of the results.

TABLE 36

ONE-WAY ORGANIZATION TABLE

ORGANIZATION OF THE DATA WITH ONE INDEPENDENT VARIABLE (ELEVATION) AND ONE DEPENDENT VARIABLE (PRECIPITATION) FOR 61 AMERICAN CITIES THROUGH 1947

Elevations 0 — 300		Elevations 301 — 1,100		Elevations 1,101 — 6,200	
Elev. (ft.)	Prec. (in.)	Elev. (ft.)	Prec. (in.)	Elev. (ft.)	Prec. (in.)
8	60	312	15	1,254	32
8	41	331	32	1,282	20
10	62	351	38	1,286	36
10	43	400	46	1,380	30
11	59	459	42	1,670	16
11	45	465	39	1,954	16
14	33	485	46	2,192	38
15	41	543	32	2,629	13
18	51	611	33	2,805	18
30	42	615	39	2,858	13
32	39	619	31	3,590	21
43	49	688	32	3,893	13
51	46	721	30	3,920	9
52	22	787	33	4,222	14
72	42	793	40	4,289	9
162	42	800	32	5,221	14
201	52	830	27	5,314	8
234	52	908	36	6,139	15
257	48	938	31		
277	9	978	28		
277	37	1,054	49		
		1,083	8		
Totals 1,793	915	14,762	739	55,898	335
Items 21	21	22	22	18	18
Means 85.4	43.6	671.0	33.6	3,105.4	18.6

In Table 36 only three classes are used. They vary greatly in width but are almost equal in frequencies. This unequal width of classes in no way violates the principle stated in Chapter 5 that "the classes of frequency distributions must be of equal width" because here we base our comparisons on the *means of classes* instead of *totals of classes.*

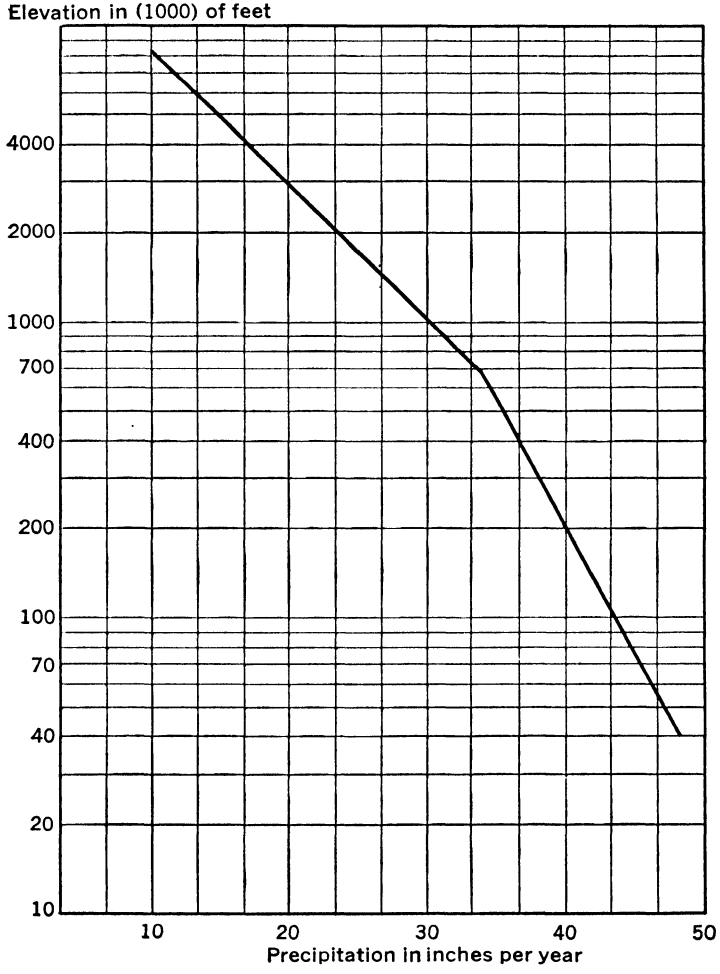


FIG. 72. Relation of mean elevation to mean precipitation, 61 American cities. (See Table 38)

TABLE 37

SUMMARY TABLE FOR ORGANIZATION TABLE PARTIAL SUMMARY OF RELATIONSHIP BETWEEN ELEVATION AND PRECIPITATION IN 61 AMERICAN CITIES THROUGH 1947

Class Intervals Elevations in Feet	Frequencies Number of Cities	Totals for Dependent Factor Precipitation	Mean Precipitation per Class
0– 300	21	915	43.6
301–1,100	22	739	33.6
1,101–6,200	18	335	18.6
	61	1989	32.6

TABLE 38

FINAL SUMMARY BETWEEN ELEVATION AND PRECIPITATION IN 61 AMERICAN CITIES SHOWING TREND BETWEEN ELEVATION AND PRECIPITATION THROUGH CLASS MEANS

Mean Elevation	Mean Precipitation
85.4	43.6
671.0	33.6
3,105.4	18.6

If one wishes to expand the analysis to include the effects of humidity along with elevation a *two-way* organization table is necessary in which one independent is placed above and the other one is repeated over and over as many times as there are classes in the outside variable.

TABLE 39

SUMMARY TABLE

PARTIAL SUMMARY BETWEEN RELATIVE HUMIDITY AND PRECIPITATION IN 61 AMERICAN CITIES

Class Intervals Relative Humidity	Frequencies No. of Cities	Total Dependent Factor Precipitation	Average of Precipitation
38–60	12	149	12.4
61–68	25	833	33.3
69–77	24	1007	42.0
	61	1989	32.61

TABLE 40

TWO-WAY ORGANIZATION TABLE

ORGANIZATION OF TWO INDEPENDENT VARIABLES (ELEVATION AND RELATIVE HUMIDITY) WITH ONE DEPENDENT VARIABLE (PRECIPITATION) FOR DATA ON 61 AMERICAN CITIES THROUGH 1947

Humidity 38–60						Humidity 61–68						Humidity 69–77					
El. 0–300		El. 301–1100		El. 1101–6200		El. 0–300		El. 301–1100		El. 1101–6200		El. 0–300		El. 301–1100		El. 1101–6200	
Elev.	Prec.	Elev.	Prec.	Elev.	Prec.	Elev.	Prec.	Elev.	Prec.	Elev.	Prec.	Elev.	Prec.	Elev.	Prec.	Elev.	Prec.
270	9	1,083	8	5221	14	201	52	312	15	1380	30	10	62	1054	49		
				2858	13	257	48	793	40	2629	13	52	22	611	33		
				3893	13	72	42	459	42	2805	18	18	51	800	32		
				4289	9	15	41	830	27	2192	38	11	59	619	31		
				5314	8	10	43	908	36	1670	16	8	60	721	30		
				3590	21	162	42	465	39	1254	32	32	39	400	46		
				3920	9			978	28	1286	36	234	52	485	46		
				4222	14			543	32	1282	20	8	41	331	32		
				1954	16			787	33			277	37	615	39		
				6139	15			351	38			30	42	938	31		
								688	32			43	49				
												51	46				
												11	45				
												14	33				
Sum 277	9	1,083	8	41,406	132	717	268	7,114	362	14,498	203	799	638	6,574	369		
Items 1	1	1	1	10	10	6	6	11	11	8	8	14	14	10	10		
Means 277	9	1,083	8	4.1406	13.2	119.5	44.7	646.7	32.9	1.812	25.4	57	45.6	657.4	36.9		

TABLE 41

SHOWING METHODS OF COMPUTING MEANS FOR CLASS 38–60 OF PRECIPITATION IN TABLE 40

Cities	Sub-class Totals Elevation	Sub-class Totals Precipitation
1	277	9
1	1,083	8
10	41,406	132
12	42.766	149

$$\frac{\text{Total Elevation}}{N} \quad \frac{42{,}776}{12} = 3564$$

$$\frac{\text{Total Precipitation}}{N} \quad \frac{149}{12} = 12.4$$

TABLE 42

SUMMARY TABLE

SUMMARY OF RELATIONSHIPS BETWEEN ELEVATION AND PRECIPITATION BETWEEN RELATIVE HUMIDITY AND PRECIPITATION SHOWING COMBINED EFFECT OF TWO INDEPENDENT VARIABLES ON PRECIPITATION

Elevation in feet	Relative Humidity			Means
	38–60 %	61–68 %	69–77 %	
0– 300	9	44.7	45.6	43.6
301–1,100	8	32.9	36.9	33.6
1,101–6,200	13.2	25.4		18.6
	12.4	33.3	42.0	32.6

Table 42 indicates that a high humidity is associated with a low elevation while a low humidity follows a high elevation. These two independent factors have an inverse relationship. One increases as the other declines. Their combined effect on precipitation for the various classes is here summarized.

ANALYSIS BASED ON FOUR CLASSES

TABLE 43

TWO-WAY ORGANIZATION TABLE

ORGANIZATION TABLE BASED ON FOUR CLASSES OF ELEVATION AND FOUR CLASSES OF HUMIDITY SHOWING MEAN PRECIPITATION AS RELATED TO TWO INDEPENDENT VARIABLES FOR 61 AMERICAN CITIES THROUGH 1947

Elevation 0–100 ft.							
Humidity 38–59		Humidity 60–65		Humidity 66–71		Humidity 72–77	
Humid.	Prec.	Humid.	Prec.	Humid.	Prec.	Humid.	Prec.
		65	43	67	42	72	62
				68	41	73	22
				70	42	73	51
						73	59
						72	60
						77	39
						75	41
						72	49
						73	46
						72	45
						73	33
		65	43	205	125	805	507
			43		41.7		46.1
		Elevation 101–650 ft.					
52	9	62	15	67	52	73	52
		65	42	66	48	72	37
		64	39	69	33	72	46
		64	38	71	31	72	39
				68	32	77	32
				69	46		
				68	42		
52	9	255	134	478	284	366	206
			33.5		40.6		41.2

Elevation 651–1300 ft.							
Humidity 38–59		Humidity 60–65		Humidity 66–71		Humidity 72–77	
Humid.	Prec.	Humid.	Prec.	Humid.	Prec.	Humid.	Prec.
38	8	64	27	69	49	77	30
		61	33	66	40		
		63	32	71	32		
				66	36		
				67	28		
				66	32		
				67	36		
				66	20		
				71	31		
38	8	188	92	609	304	77	30
	8		30.7		33.8		30
		Elevation 1301–6200 ft.					
48	14	62	30	68	38		
52	13	61	13	66	16		
56	13	62	18				
52	9	60	16				
47	8						
55	21						
38	9						
57	14						
53	15						
458	116	245	77	134	54		
	12.9		18.5		27		

Table 43 is a good example of the difficulties of attempting to make a long and complex organization table with a small sample. With the three variables of (1) elevation (2) humidity and (3) precipitation a complete analysis with thirty-two separate class columns there should be at least 200 items of data for valid results. This method is, therefore, not useful with small samples. In marketing and personnel it may be used effectively.

TABLE 44

SUMMARY TABLE

THE NUMBER OF THE 61 AMERICAN CITIES WHICH FALLS IN EACH OF THE FOUR CLASSES OF ELEVATION AND OF RELATIVE HUMIDITY, SHOWING THE REGRESSION BETWEEN THE TWO VARIABLES

Relative Humidity	Elevation				
	0–100	101–650	651–1300	1301–6200	Totals
38–59		1	1	9	11
60–65	1	4	3	4	12
66–71	3	7	9	2	21
72–77	11	5	1		17
Totals	15	17	14	15	61

The diagonal line of large frequencies 9, 9, 7, 5, 11 extending from the upper right-hand to the lower left-hand corners of the table indicate the high inverse or negative relationship between (1) elevation and (2) humidity as they affect precipitation. Rain is heavy where elevation is low and humidity high, while rain is low where elevation is high and humidity low.

TABLE 45

SUMMARY TABLE

THE MEAN PRECIPITATION FALLING IN EACH CLASS OF 61 AMERICAN CITIES AS RELATED TO THE COMBINED EFFECT OF ELEVATION AND RELATIVE HUMIDITY THROUGH 1947

Relative Humidity	Elevation				
	0–100	101–650	651–1300	1301–6200	Totals
38–59		9	8	12.9	12.1
60–65	43	33.5	30.7	18.5	28.8
66–71	41.7	40.6	33.8	27.0	36.5
72–77	46.1	41.2	30.0		43.7
Totals	45.0	37.2	31.0	16.5	32.6

Third Sample of Tabular Analysis

Tabular analysis is not only the most widely used method of statistical analysis involving two or more variables whenever large samples are available, but it has proved to be especially popular in the field of education. This is probably due to the facts that data in education are very abundant and are formed in such complex relationships that more highly mathematical methods often give rather artificial results. This is not to say that regression and multiple and partial correlation may not often be used with excellent results on educational and psychological data, but that often these data yield better results from the more plastic, variable and less rigid methods of tabular analysis. Table 46 gives the results in summary form for 733 cases of college students.

TABLE 46

SUMMARY TABLE

ANALYSIS SUMMARY OF COLLEGE APTITUDE AND ACHIEVEMENT TESTS OF 733 ARTS AND SCIENCES STUDENTS, OKLA. A. & M. COLLEGE, 1930–39. SHOWING TOTALS AND PERCENTAGES BY QUARTILES *

A.C.E. Percentile	Grade Point Averages				Quartiles Totals
	0.0–0.9	1.0–1.9	2.0–2.9	3.0–4.0	
0–24	9 11%	32 38%	38 46%	4 5%	83 100%
25–49	14 9%	49 31%	81 52%	13 8%	157 100%
50–74	12 6%	48 22%	111 52%	42 20%	213 100%
75–99	5 2%	38 14%	125 44%	112 40%	280 100%
Totals, Grade Average Groups	40	167	355	171	733

* Source: "Psychological Tests as Administrative Tools," Dean Schiller Scroggs.

Computation of Error in Tabular Analysis

The various methods and formulas for computing the error and reliability of the results in tabular analysis are taken up in their logical place in the following chapter in which the problems of sampling, probability, and error for all large samples are discussed in detail. The purpose of the present chapter is only to set up the detailed methods of making a tabular analysis. Chapter 13 presents the methods of measuring the dependability of the results.

Advantages of Tabular Analysis

1. In tabular analysis one is not forced to make any pre-judgments or assumptions as to the nature of the relationships in the data, either as to whether the regression is straight-line or curvilinear, or the relationships are arithmetic or geometric. It usually reveals whatever relationship that exists in the form in which it occurs, straight-line if they are straight and curved if they are actually curved.

2. It is a much shorter method than complex mathematical analysis and is an excellent device for discovering what relationships do exist before spending the time and money to make a long mathematical analysis. It greatly aids in deciding on what specific mathematical functions to employ if such analysis is later desired.

Disadvantages of Tabular Analysis

1. In tabular analysis, no ratio or coefficient of correlation is obtainable.

2. No least squares regression line is obtainable but a class-to-class trend is obtainable.

3. No measure of partial correlation is available and it is not possible to get a full, clear measure of the relative weight of each independent variable as it affects the total combined result. The relative weight of the independent variables is somewhat obscured if there are more than two.

With all of these mathematical limitations, tabular analysis is more widely used than any other complex method of statistical analysis.

SUMMARY

1. Tabular analysis is not just the making of tables but the use of statistical tables to analyze data.

2. Data tables are simply the associated values of the original data in unorganized form.

3. Organization tables are the arrangement of the original data into various combinations of columns and lines to reveal the relationships which obtain in the data. They may be designed in any form to suit any purpose.

4. The form of organization table to be used in the analysis should be clearly in mind as a part of the planning of the study before the data are collected.

5. Summary tables are brief tabulations of the results revealed in the organization tables. They may be carried to any desired degree of complexity.

6. Complete cross-classification tables are usually the most effective type of table for tabular analysis. The construction of such tables is explained in Chapter 3.

7. In making organization tables it is usually preferable to divide the range of data into class intervals which will contain approximately equal numbers of items. There may be any number of class intervals for each variable, but two, three, or four are the most common. More than five classes makes the table too large and complicated for good results.

8. The reliability of the means of the various classes in the table may be computed by the formulas given in Chapter 13.

9. Tabular analysis does not compel any pre-judgments as to the nature of the functions or relationships existing in the data; it simply reveals the facts as they are.

10. Tabular analysis is usually a shorter and less laborious type of analysis than mathematical analysis.

11. Tabular analysis has the defect that it does not supply a coefficient of correlation or a mathematical regression line.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter VIII, Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter III, Prentice-Hall, Inc., New York, 1948.

Ezekiel, M., *Methods of Correlation Analysis*, Chapters X, XI, John Wiley & Sons, New York, 1941.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to*

Agricultural Economics, Chapters VIII, XV, John Wiley & Sons, New York, 1942.

REVIEW QUESTIONS

1. What is the difference between tables and tabular analysis?
2. What is a "data table"? How must it be organized? Explain.
3. What is an "Organization Table"? On what principles should such a table be organized? Explain.
4. What is the difference between a *one-way* organization table and a *two-way* organization table? Explain.
5. What is the difference between a *two-way* and a *three-way* organization table? Explain in detail.
6. What is a summary table? On what factors does its form depend? What variety of forms may it take?
7. Should class intervals be of the same width? Why? Explain.
8. Should class frequencies be equal or approximately equal or does this condition make any difference in the analysis? Why?
9. How many columns would be required for a two-way four-class table that furnished complete sub-class totals?
10. How many columns would be required for a three-way table with sub-totals for one independent and the dependent variables if based on four classes for each variable?
11. Ordinarily how many classes should there be for each variable? Why?
12. What are the advantages and disadvantages of having only two class intervals for each variable?
13. What are the advantages and disadvantages of having six or eight class intervals for each variable?
14. What are the advantages and disadvantages of Tabular Analysis as compared with correlation?
15. For what types of data and problems is it better adapted? Why?

CHAPTER 13

PROBABILITY, SAMPLING, AND ERROR

In the preceding chapters we have learned how to compute many statistics. We know how to measure the central tendency, dispersion, skewness and kurtosis of a single variable and how to compute many relationships between two or more variables. The assumptions in all cases have been that although the samples and statistics were only representative and not perfect measures of the populations from which they were drawn, they did measure them quite well.

The assumption was that the arithmetic mean of $167.87 for the semi-monthly wages of 188 oil refinery employees was a fair indication of the earnings of these men. Likewise it was assumed that the standard deviation of $20.60 for the same sample was a good measure of the dispersion of the population from which that sample was drawn. It was merely assumed that the statistic was representative of the parameter of the universe but no measure of the degree of its accuracy was made.

Statistics are only approximations of the parameters of populations and vary widely from one to another as to their dependability. If a sample is not truly representative of a universe the statistics computed from it cannot measure that universe accurately. No statistical study, therefore, is complete until its degree of error is measured.

ERROR

The primary meaning of error is deviation or differences. Even a social error is only a deviation from the accepted mode of be-

havior. If the norm is fixed and accepted without question as the relations of mathematics are, error comes to mean mistake. If the product of two times two is stated as five, or anything other than four, the computer is said to have made a mistake, or an error. Error is not used in statistics in the sense of mistake, but is used only in the sense of difference or deviation. In computing the statistics of a sample as a means of measuring the parameters of a population, there is always the certainty that the statistics and the parameters will not be identical. If the exact mean of the heights of all the men over twenty years old in the world is 67.437 inches, it is most likely that no sample of this population would yield exactly an identical statistic. This deviation between the parameter and the statistic is the error, or deviation.

If one computes the mean of a random sample taken from a population or universe, the question always arises, how near does the mean of the sample (a statistic) approach the true mean of the universe (a parameter)? How can this question be answered? If we cannot know the parameter, how can we know how far the statistic misses it? No doubt, in many cases we can never know the exact deviation or difference. The most that can be done is to discover approximate limits to varying degrees of probability or areas or zones of confidence. If the probability of an event or statistic can be reduced to a ratio, so that one can be confident that such an event would occur 99 times out of 100 chances, or 98, or 95 times out of 100 chances, he has set secure limits to the probable size of the error and can judge the representativeness of the sample and its statistics accordingly.

The entire question of the representativeness of samples and their statistics turns on the adequacy of the measures of probability.

PROBABILITY

Probability is one of the oldest phases of statistics and was first developed as a mathematical theory in relation to games of chance. *Probability means the likelihood of the occurrence of an event.* One speaks of the probability of rain today or of snow tomorrow, or

the probability of winning the game Saturday, or of making a profit on a real estate venture, or of the rise in the price of a certain stock or of losing a battle or winning a war. Many of our questions have to be answered by "It probably will" or "It probably won't." In any case the probability of the event is the likelihood that it will occur.

The difficulty with the use of the term "probability" in everyday language is that which is common in the popular use of words. It lacks scientific precision and definiteness. Professor Richard von Mises[1] redefines probability in its relation to statistics with a scientific precision and clarity which makes this concept much more useful even to a beginning student. In a simple and abbreviated form one may say, *the probability of an event is the relative frequency with which this event recurs in an indefinitely prolonged sequence or series of observations.* If for instance in throwing a die a great many times, the number 3 came up in $\frac{1}{6}$ of all the throws, the probability of getting a "3" from throwing the die would be $\frac{1}{6}$, or one out of every six throws. If out of a very large number of birth registrations it appears that 52 out of every 100 births are males, the probability of a child being a male is .52. The probability is the *limiting value* of the ratio of this particular event (birth of a boy) out of all births.

The probability, the limiting value of the ratio, is established objectively by experimentation, or by repeated observations. Clearly to establish the limiting, or correct, ratio the number of observations must be large. For instance, among ten births 7 might be girls, but among 1,000 or among 10,000 births the correct, or limiting ratio of .52 would certainly appear. Out of 1,000 births the number of males would be near 520 in a normal population. Von Mises calls the long sequence of observations a "collective." We have learned to refer to such large fields of data as populations. By "collective," von Mises means a specific and clearly differentiated population of definite attributes. If out of a large sample taken from a population a certain attribute or event recurs a given number of times out of the total possible num-

[1] Mises, Richard von, *Probability Statistics and Truth*, The Macmillan Company, N.Y., 1939.

ber of times in the sample, that ratio is its probability. In this sense probability is the final or limiting value of a ratio based on a large number of observations from a population, or "collective."

This definition of probability is very explicit and is free from all subjective and philosophical content. It is a ratio determinable by sampling and objective experimentation. In any case the probability of an event is the likelihood that it will occur, this likelihood having been established as a dependable ratio by scientific statistical procedure.

Degrees of Probability

The world in which we live is a complex of probabilities which either singly or in combinations are neither equally certain nor equally measurable by statistical methods. As samples of the degrees of variation and uncertainty in probabilities three classes of events will be considered in this study.

1. *The probability of pure chance events.*
2. *The probability of variations in the occurrences in nature, physical and biological.*
3. *The probability of variations in human events and in institutions.*

These three classes of probabilities do not cover all cases but they do encompass much of the *range* of varying probabilities. The probability of events based on pure chance will be treated first.

Probability of Pure Chance Events

All chance events may either occur or not occur, or may occur in favorable or unfavorable ways.

If the total possible occurrences are divided into *favorable* and *unfavorable* events, then

n = total number of occurrences
p = number of favorable occurrences
q = number of unfavorable occurrences

and $p + q = n$ and $\frac{p}{n} + \frac{q}{n} = 1.$

This simple standard equation of probabilities should be kept clearly in mind throughout the following analysis.

Perhaps the simplest and most common pure chance experience of the average student is the flipping of coins. As two students go to the fountain one will say, "Let's match for drinks." "All right, I'll take heads." The coin is flipped and turns up tails. The student may say, "I'm always unlucky." But in the field of pure chance, *he can't be.* If these two students were to go to the fountain every day for a hundred days and "match for drinks," and the coin was normal and was fairly thrown the results would be about 50–50. At least in tossing a perfect coin a larger number of times approximately one-half of the tosses will turn tails and one-half will turn heads. It is not possible to predict with any degree of accuracy how any particular toss will turn, but it is possible to predict with nearly complete accuracy what a thousand fair, free tosses will give. The result would be approximately 500 heads and 500 tails. This is the principle of the probability of pure chance. The equation is $\frac{1}{2} + \frac{1}{2} = 1$. It is basic in statistical analysis and will be developed more fully later.

Unequal Probabilities

Not all probabilities fall into 50–50 or equal favorable and unfavorable distributions. The chance that one may draw the queen of hearts from a well-shuffled deck of fifty-two playing cards is only one out of fifty-two, or $\frac{1}{52}$. The chances are fifty-one out of fifty-two that one will not draw the queen of hearts, or $\frac{51}{52}$.

$$\frac{1}{52} + \frac{51}{52} = 1$$

These probabilities equal 1, but they are not equally divided.

Additions of Probabilities

The chances that one will draw a queen are four out of fifty-two, or $\frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{4}{52} = \frac{1}{13}$. The chances that one will draw a black-faced card are $\frac{26}{52}$. The probabilities that he will draw *a* spade are $\frac{13}{52}$. *Single alternative probabilities are additive.*

Multiplication of Probabilities

The chance that *two independent events will occur in unison is the product of their individual probabilities.* The chance that one

will at the same time, or in immediate succession, draw the king and queen of clubs is $\frac{1}{52} \times \frac{1}{52} = \frac{1}{2704}$. The probability that one will draw a club and a spade is $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$. The more complex the situation, the greater the number of independent events combined into one compound event, the smaller is the probability of such an event out of the total possible occurrences. This explains why one may play cards for hours and even for months without getting two hands during this period that are identical.

These relationships may be further illustrated by the throwing of dice. What are the chances, what is the probability, that if one throws two ordinary, well-formed, evenly balanced dice, A and B, he will get a 12 or a 7 or an 11? The following table indicates these probabilities.

Get	A	B	Total
12	6	6	$\frac{1}{36}$
11	5	6	
	6	5	$\frac{2}{36}$
10	4	6	
	6	4	
	5	5	$\frac{3}{36}$
9	4	5	
	5	4	
	3	6	
	6	3	$\frac{4}{36}$
8	4	4	
	3	5	
	5	3	
	2	6	
	6	2	$\frac{5}{36}$
7	3	4	
	4	3	
	2	5	
	5	2	
	6	1	
	1	6	$\frac{6}{77}$

Get	A	B	Total
6	5	1	
	1	5	
	3	3	
	2	4	
	4	2	$\frac{5}{36}$
5	4	1	
	1	4	
	2	3	
	3	2	$\frac{4}{36}$
4	2	2	
	1	3	
	3	1	$\frac{3}{36}$
3	1	2	
	2	1	$\frac{2}{36}$
2	1	1	$\frac{1}{36}$
Grand Total.....			$\frac{36}{36}$

The chances that one will throw a 7 is $\frac{6}{36}$, or one time out of 6. That he will throw a 7 or more than a 7 is $\frac{21}{36}$, and that he will throw 11 or more, are $\frac{3}{36}$.

Human beings everywhere are deeply interested in probability. What is the probability that it will rain today or tomorrow? that I will fail my examination or make a "B" grade? or that my investment will pay? or that our marriage will be successful? What are the chances that I shall escape from a shipwreck, a battle, or be killed in a car accident? What is the likelihood that I shall live to be seventy years old? Day by day we seek to know the probability of events in which we are interested. In statistical analysis an understanding and measurement of probability are of fundamental importance. *In statistics no measure is complete until its probability, or standard error, has been computed.*

MATHEMATICAL MEASUREMENT OF PROBABILITY

In order to make the principles and relationships of the probability of pure chance events available to the practical statistician, it was necessary to describe and measure them in mathematical equations. This work has been done by a number of brilliant mathematicians[1] over a long period of years. Of the various equations available two will be employed in this text. They are (1) The Binomial Expansion, and (2) The Normal Probability Curve, or Normal Curve of Error.

The Binomial Expansion

The student has learned in elementary algebra how the quantity of $(a + b)^2$ may be expanded to any desired power, such as

$$(a+b)^n=a^n+na^{n-1}b+\frac{n(n-1)}{1\cdot 2}a^{n-2}b^2+n\frac{(n-1)(n-2)}{1\cdot 2\cdot 3}a^{n-3}b^3+\cdots+b^n.$$

Or if the computation is for $(a + b)^2$ the result is $1a^2 + 2ab + 1b^2$,

and $$(a+b)^3 = 1a^3 + 3a^2b + 3ab^2 + 1b^3,$$

and $$(a + b)^4 = 1a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + 1b^4.$$

[1] Leplace, Pierre Simon, 1749–1827, greatest of French astronomers and mathematicians. Also, Gauss, Karl Friedrich, 1777–1855. German mathematician of the University of Göttingen. See *Encyclopedia Britannica*, or *New International Encyclopedia.*

For any power of $(a + b)$ the coefficients of the several terms express and describe the total number of possible chance combinations for that particular power. For $(a + b)^2$ these are $1+2+1=4$, for $(a + b)^3$ they are $1 + 3 + 3 + 1 = 8$, and for $(a + b)^4$ they are $1 + 4 + 6 + 4 + 1 = 16$. The number of possible chance combinations doubles for each successive power.

Since in an occurrence based on pure chance such as the tossing of a coin or the throw of a die, the possibilities of favorable and unfavorable events are equal (the coin has an equal chance to turn heads or tails), the equation may be written:

$$p = \tfrac{1}{2},\ q = \tfrac{1}{2},\ p + q = 1,\ (\tfrac{1}{2} + \tfrac{1}{2}) = 1.$$

Possible Occurrences For Various Numbers of Coins

If two coins, a and b, are thrown the total possible combinations of occurrences are:

a	b	a	b	a	b
H	H	H	T	T	T
		T	H		

or $(\tfrac{1}{2} + \tfrac{1}{2})^2 = \tfrac{1}{2}^2 + 2(\tfrac{1}{2})(\tfrac{1}{2}) + (\tfrac{1}{2})^2$

$\tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4} = \tfrac{4}{4} = 1.$

If three coins a, b, and c are tossed the possible results are:

a	b	c	a	b	c	a	b	c	a	b	c
H	H	H	H	H	T	T	T	H	T	T	T
			H	T	H	T	H	T			
			T	H	H	H	T	T			

$$(\tfrac{1}{2} + \tfrac{1}{2})^3 = (\tfrac{1}{2})^3 + 3(\tfrac{1}{2})^2(\tfrac{1}{2}) + 3(\tfrac{1}{2})(\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8} = \tfrac{8}{8} = 1.$$

If four coins a, b, c, and d are tossed the total possible results are:

a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
H	H	H	H	H	H	H	T	H	H	T	T	T	T	T	H	T	T	T	T
				H	H	T	H	H	T	H	T	T	T	H	T				
				H	T	H	H	T	H	T	H	T	H	T	T				
				T	H	H	H	T	T	H	H	H	T	T	T				
								T	H	H	T								
								H	T	T	H								

$$(\tfrac{1}{2} + \tfrac{1}{2})^4 = (\tfrac{1}{2})^4 + 4(\tfrac{1}{2})^3(\tfrac{1}{2}) + 6(\tfrac{1}{2})^2(\tfrac{1}{2})^2 + 4(\tfrac{1}{2})(\tfrac{1}{2})^3 + (\tfrac{1}{2})^4$$
$$= \tfrac{1}{16} + \tfrac{4}{16} + \tfrac{6}{16} + \tfrac{4}{16} + \tfrac{1}{16} = \tfrac{16}{16} = 1.$$

In like manner this binomial equation of pure chance occurrences may be expanded to any desired powers.

Binomial Triangle

The coefficients of the terms of the binomial expanded to successive powers may be recorded in the form of a right triangle which materially reduces the labor of computing the successive equations.

TABLE 47

BINOMIAL TRIANGLE

Exponents	Totals	Coefficients											
1st	2	1	1										
2nd	4	1	2	1									
3rd	8	1	3	3	1								
4th	16	1	4	6	4	1							
5th	32	1	5	10	10	5	1						
6th	64	1	6	15	20	15	6	1					
7th	128	1	7	21	35	35	21	7	1				
8th	256	1	8	28	56	70	56	28	8	1			
9th	512	1	9	36	84	126	126	84	36	9	1		
10th	1,024	1	10	45	120	210	252	210	120	45	10	1	
11th	2,048	1	11	55	165	330	462	462	330	165	55	11	1
12th	4,096	1	12	66	220	495	792	924	792	495	220	66	12	1

To arrange this table (1) set down as left-hand stubs the successive exponents desired under the heading of *Exponents*. (2) Write down 1 as the first coefficient of each power. (3) Form the next column by writing 1 in the first line, and to form the number (coefficient) below it, add to it the number directly to its left, as $1 + 1 = 2$, $2 + 1 = 3$, $3 + 1 = 4$, etc. (4) Begin each succeeding column one line lower with 1 as its first number and obtain the successive numbers (coefficients) below it by adding to each number in that column the number in the same line immediately to its left. Begin the third column opposite exponent 2 in the second line with 1. To get the 3 below it add the 2 beside it to the 1, as $1 + 2 = 3$, then $3 + 3 = 6$, and $6 + 4 = 10$, and $10 + 5 = 15$, and so on for all other coefficients. By this method the student can quickly obtain whatever set of coefficients he may

require for a given frequency distribution. The column headed *Totals* next to the column headed *Exponents* contains the total of all coefficients for that power. They double for each successive power.

WORKSHEET NO. 60

COMPARISON OF THEORETICAL AND ACTUAL DISTRIBUTIONS OF THROWS OF TWELVE COINS (PENNIES) 4,096 TIMES

Number of Heads up in Throw	Frequency of Throws	
	Theoretical Number	Actual Number
0	1	1
1	12	15
2	66	61
3	220	205
4	495	506
5	792	783
6	924	926
7	792	805
8	495	504
9	220	212
10	66	60
11	12	16
12	1	2
Totals	4,096	4,096

These two frequencies are plotted on Fig. 73. It should be observed that the binomial expanded to the 12th power does *not* give a continuous smooth curve, but only thirteen points on such a curve. If, however, the number of coins and throws were increased to a very large number these plotted points would become more numerous and closer together, and as the number of coins and throws continued to increase the curve would ultimately approach a smooth continuous line. Only with this reservation, that the sample be infinitely large, that is, that it approach the limits of an infinite population, can the binomial expansion ac-

tually be a full and accurate measure of the probabilities of pure chance events. However, for practical purposes an ideal binomial expansion frequency curve may be fitted to a finite or limited sample to discover whether and how near the sample has the form of a true binomial curve of a given power. It is a useful device for describing and measuring a frequency distribution. Such an equation is called the *point binomial*, because it gives only a finite or limited number of points and not a smooth curve.

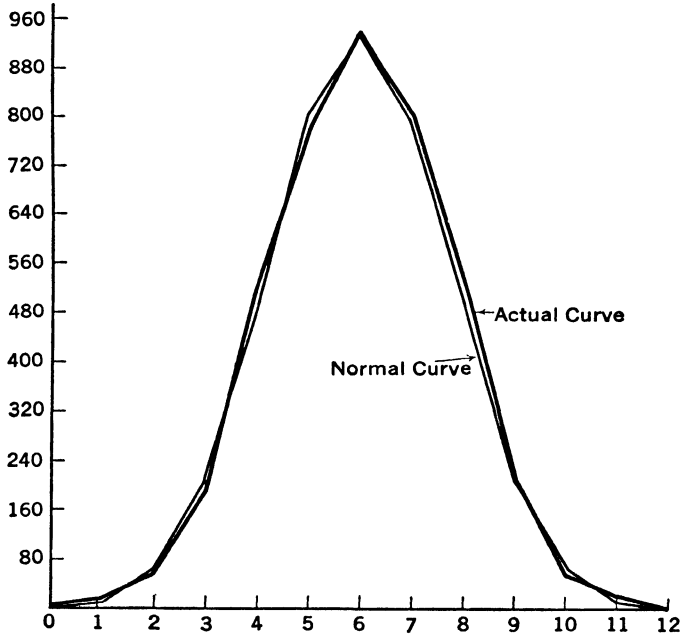


FIG. 73. Comparison of theoretical and actual frequencies of 12 coins thrown 4,096 times, as shown in Worksheet No. 58.

THE LAW OF PURE CHANCE OCCURRENCES

From the relationship revealed in the equations above and illustrated in Fig. 73 the following generalization may be made.

Variate values of pure chance occurrences tend to be distributed

about their mean approximately in proportion to the values of coefficients to the nth power of a suitable binomial.

The twelve-sided polygon shown in Fig. 73 would change to a figure of more and shorter sides as the powers increased until at infinity it would constitute a perfectly continuous smooth curve.

THE NORMAL CURVE

This perfectly smooth curve which results from the expansion of n, the index of the binomial $(\frac{1}{2} + \frac{1}{2})^n$ is called the *normal curve*, or the *normal probability curve*, or the *normal curve of error*. It is sometimes referred to as the Gaussian curve or the Laplacian curve from the names of two prominent mathematicians who contributed to its development.

The equation of this curve may be written in several forms the following one of which is perhaps the simplest and most useful for the elementary student.

Meaning of symbols:

Y_c = the computed height of an ordinate at a distance x from the arithmetic mean
i = class interval
σ = standard deviation of the sample
π = the constant, 3.1416; $\sqrt{2\pi} = 2.5066$
e = the constant, 2.71828, natural log base
x = any selected deviation from the arithmetic mean
N = number of observations in the sample

FORMULA NO. 99

$$Y_c = \frac{Ni}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

If we substitute the two constants, $\sqrt{2\pi} = 2.5066$ and 2.71828 in the equation, it becomes

$$Y_c = \frac{Ni}{2.5066\sigma} 2.71828^{-\frac{x^2}{2\sigma^2}}$$

At the mean $-\frac{x^2}{2\sigma^2} = 0$. $Y_c = \frac{Ni}{2.5066\sigma}$ at the mean.

By substituting in this equation Ni, the number of items in the sample, times the class interval, the standard deviation of the sample and any particular x, one may compute the ordinate Y_c, and by varying x may compute as many ordinates as desired. This formula is rather complex and it is not expected that the elementary student will spend much time with it. Its form is shown in Fig. 74.

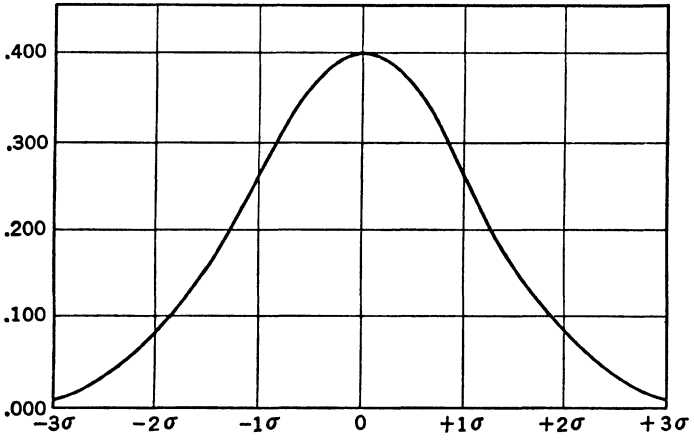


FIG. 74. Normal curve of error

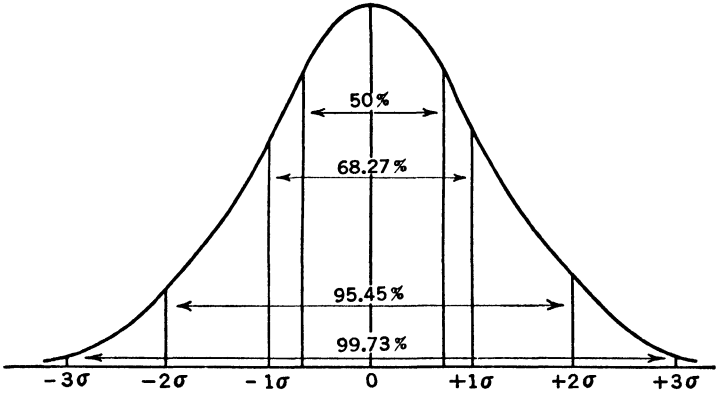


FIG. 75. Showing normal curve of error and percentages of items included within various ranges of error

Ordinates of the normal probability curve are shown in Appendix Table I. This table is quite useful in fitting normal curves to frequency distributions. The normal curve may also be fitted to any given frequency distribution by the use of the formula itself. If we make a histogram of the frequency distribution of 106 Stillwater school children, we may also plot on this same graph a normal curve as shown in Fig. 76. The ordinates of the normal curve are obtained as follows:

1. Take the formula $Y_c = \frac{Ni}{2.5066\sigma} 2.71828^{-\frac{x^2}{2\sigma^2}}$

2. Select x's which are some exact simple fractions of the standard deviation of the sample, such as $\frac{1}{4}\sigma$, $\frac{1}{2}\sigma$, etc., and substitute them in the above formula successively until sufficient ordinates are obtained to outline a smooth curve $\sigma = 4.7$.

If in the equation

$$Y_c = \frac{Ni}{2.5066\sigma} 2.71828^{-\frac{x}{2\sigma^2}}$$

are substituted the values for N and i in the problem on heights of school children and x is set equal to one standard deviation, 4.7 in this case, the solution of the equation becomes

$$Y_c = \frac{106 \times 2}{2.5066 \times 4.7} 2.71828^{-\frac{(4.7)^2}{2(4.7)^2}}$$

$$= \frac{212}{11.78102} 2.71828^{-\frac{(22.09)}{2(22.09)}}$$

$$= 17.993 \times \frac{1}{2.71828^{\frac{1}{2}}} = 18.0 \times \frac{1}{\sqrt{2.71828}}$$

$$= 18.0 \times \frac{1}{1.6489} = 18.0 \times .60653 = 10.9175$$

which is the ordinate of the normal curve distant one σ on either side of the mean. By substituting other values for x in terms of σ as many ordinates may be computed as desired.

For x = one-half σ, or 2.35 in this case the solution is

$$Y_c = \frac{Ni}{2.5066\sigma} 2.71828^{-\frac{(2.35)^2}{2(4.7)^2}}$$

$$= 18.0 \times 2.71828^{-\frac{(5.5225)}{2(22.0900)}}$$

$$= 18.0 \times 2.71828^{-\frac{1}{2}\times\frac{1}{4}} = 18.0 \times 2.71828^{-\frac{1}{8}}$$

$$= 18.0 \times \frac{1}{2.71828^{\frac{1}{8}}} = 18.0 \times \frac{1}{2.71828^{\frac{1}{8}}}$$

$$= 18.0 \times \frac{1}{\sqrt[8]{2.71828}} = 18.0 \times .88250 = 15.880$$

which is the ordinate at a distance of one-half σ on either side of $\overline{X}$ in this sample of heights of school children.

This method of computation of ordinates is so tedious that tables of ratios have been developed which greatly facilitate the computations.

If in the equation of the nominal curve x is set to equal zero (0) the solution is simplified to

$$Y_c = \frac{Ni}{2.5066\sigma} 2.71828^{-\frac{(0)^2}{2(4.7)^2}}$$

Since any quantity at a *power of zero equals* 1, or $2.71828^{-\frac{(0)^2}{2(4.7)^2}}$ becomes one (1), the maximum ordinate for any distribution may be computed from the formula

$$Y_c = \frac{Ni}{2.5066\sigma},$$

which in this case becomes

$$\frac{212}{11.78102} = 17.993 = 18.0$$

By selecting the ratios in Appendix Table I for desired values of $\frac{x}{\sigma}$ such as

$$\frac{1.175}{4.7} = .25, \quad \frac{2.35}{4.7} = .5, \quad \frac{4.7}{4.7} = 1.0, \text{ etc.},$$

the student may quickly compute all the desired ordinates, by multiplying the maximum ordinate by the desired ratio in Appendix Table I. For this sample of heights of school children they are shown in Worksheet No. 61.

The superimposed normal curve must be centered at the mean of the histogram on frequency polygon of the sample as is shown in Fig. 76.

WORKSHEET NO. 61

COMPUTATION OF ORDINATE OF NORMAL CURVE AT INTERVALS OF .25σ FROM MEAN. BASED ON APPENDIX III

Distance from Mean in .25σ	Ratio to Ordinate at Mean		Mean Ordinate		Ordinate in Terms of Data
.00	1.00000	×	18	=	18.0000
.25	.96924	×	18	=	17.4463
.50	.88250	×	18	=	15.8850
.75	.75484	×	18	=	13.5871
1.00	.60653	×	18	=	10.9175
1.25	.45783	×	18	=	8.2409
1.50	.32465	×	18	=	5.8437
1.75	.21627	×	18	=	3.8929
2.00	.13534	×	18	=	2.4361
2.25	.07956	×	18	=	1.4321
2.50	.04394	×	18	=	.7909
2.75	.02280	×	18	=	.4104
3.00	.01111	×	18	=	.1999

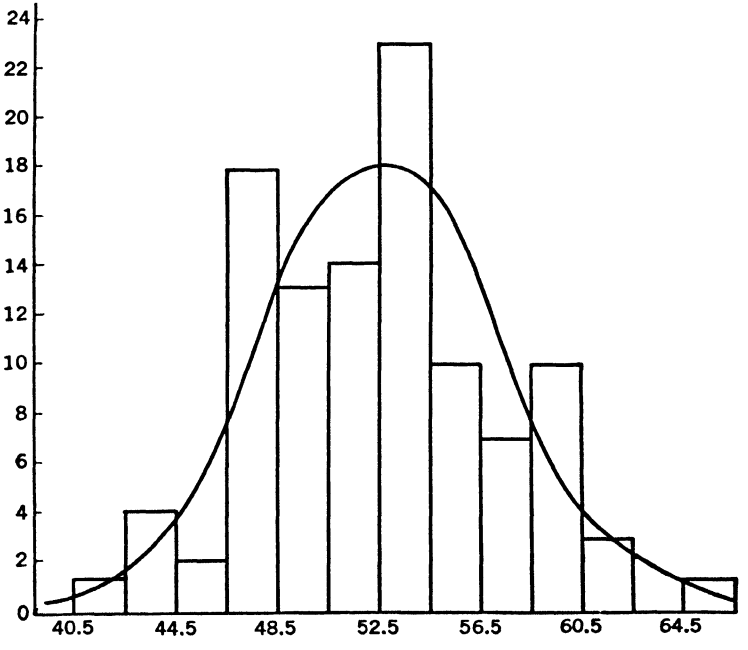


FIG. 76. Histogram of frequency distribution of heights of 106 Stillwater, Oklahoma, grade-school children and superimposed normal probability curve

The Normal Curve with Physical and Biological Data

The normal curve as explained in the previous sections of this chapter deals with occurrences based on pure chance. The logical principles underlying this curve are:

1. The forces affecting individual events are independent of each other.
2. The forces affecting individual events are many and of approximately equal weight.
3. These forces operate so as to produce deviations from the mean that are about equal in size and number.

In the realm of pure chance these statements are correct, and fortunately they also apply with a large degree of accuracy in the physical and biological fields. The leaves of a tree ordinarily vary in length approximately according to these principles. There will be a few very short leaves and very long leaves. As the lengths of the leaves approach the mean their numbers increase according to the frequencies of the normal curve. The heights of trees of the same variety and age and location and the heights of mature men of the same race as well as the size and weights of animals of the same species and age as well as most other natural phenomena tend to conform to the normal curve. This conformity of natural phenomena to the normal curve of pure chance occurrences is so close that the general law may be stated as

Variate values of a natural phenomenon tend to be symmetrically distributed about the mean in proportions conforming to the law of chance distributions.[1]

This similarity between pure chance occurrences and the distribution of most natural phenomena widens the scientific and mathematical application of statistical analyses to a large portion of the physical and biological universe. Since the same three laws of (1) independent forces, (2) numerous and equal forces, and (3) equal deviations from the mean obtain and operate in the

[1] Jerome, Harry, *Statistical Methods*, Harper & Brothers, New York, 1924, p. 150.

physical and biological fields there is a sound logical basis for expecting such phenomena to conform to the normal curve. The wide prevalence of these basic relationships in all these fields lays the foundation for logical statistical inferences from samples in these fields. *In fact, wherever these laws apply statistical measures are more than descriptive devices. Under such favorable conditions statistics become the secure bases of scientific inference and forecasting.* Out of such studies science is built.

Analysis of Normal Curve

Within the normal probability curve there exist certain fixed ratios, relations, proportions and distributions which may be stated in standard statistical and mathematical terms and universally applied with a high degree of accuracy for all distributions of data approximating the form of the normal curve.

Relationships in Normal Curve

1. The arithmetic mean, median, and mode coincide.
2. The standard deviation cuts the curve at the points of inflection (the points on the steep slope where the curvature reverses).
3. The first and third quartiles are equally distant from the median.
4. Within one standard deviation plus and minus from the mean, 68.268% of the items fall.
5. Two standard deviations taken plus and minus from the mean contain 95.45% of the items.
6. The average deviation is .7979 of the standard deviation.
7. The semi-interquartile range equals the probable error, which equals .6745 of the standard deviation.

For any data which fall approximately in the form of the normal curve and are correctly based on the three logical principles underlying the normal curve, the seven ratios stated above will hold true and may logically be deduced from the samples as applicable to the population from which they are taken.

For example, one may logically expect these ratios to hold as to variations in the heights of men, length of tree leaves, sizes of grains of sand, and velocity and forces of molecules of gas striking

the walls of the container. They would also apply to any relationship in the social sciences, economics, psychology, sociology, and education in which it may be established that the basic relationships of the normal frequency curve apply.

PROBABILITY OF VARIATION IN HUMAN BEHAVIOR AND IN SOCIAL EVENTS

The major difficulty in applying the analysis of the normal curve to social data is that in this field the three basic logical principles on which the normal curve is based are not entirely true. A coin or a die when freely thrown has an equal chance to fall either way, but in human and social relationships one is *not fully free* either in heredity or environment. The children of the mentally deficient do *not* have a full free chance to be mentally normal or superior. They are to some extent bound by the genes of their parents and are more likely to be like them than different from them, *Like begets like.* The children of poor parents who lack education and culture are more likely to lack culture and education, even in a democracy, than the children of well-to-do parents. *Culture, too, tends to perpetuate itself.* Instead of a child's being as free as a thrown penny to turn whichever way it falls, it is held and biased in its life by its heredity and environment to some extent, perhaps in many cases to a large extent. The same principle affects groups of persons from the local communities to the state and nation. Certain biases, traits, characteristics, "bents," prejudices, handicaps, and peculiarities tend to persist and to determine the lives of many succeeding generations.

These traits and biases tend to crystallize into customs, traditions, and institutions which are imposed on children from birth and mold and color their thinking, characters, and lives in spite of even strong individual tendencies to "fall" some other way.

These forces affecting human beings *are numerous, but they are often neither independent nor evenly balanced.* They frequently do *not* tend to produce *variations of equal extent and weight on either side of their mode.* Their mode, median, and mean do *not* coincide, neither are their quartiles equally distant from their median. In

other words they are not normal but are definitely skewed either to the right or to the left. The sons and daughters of professional families tend to become professional people. The children of day laborers tend to become day laborers. The sons of planters tend to become planters and the sons of share croppers tend to follow in their fathers' economic status.

Three factors make it much more difficult to generalize from the normal curve in the social sciences than in the physical and biological fields.

1. The data are much more complex and variable.
2. The data are usually skewed to a marked degree.
3. The complexity, variability and skewness of the data are not uniform and permanent but are subject to trends, cycles and complex shifts.

These conditions make dependable research more difficult in the social science field than in the physical and biological domains; but the student must not conclude that statistics are useless in social studies, because it is here that they are most necessary and relatively fruitful. It simply means that in this more difficult field the statistician must be more careful and better equipped for his work. He must become a master of both statistical techniques and subject matter.

One alleviating condition in the social sciences is the relative abundance of data. The records of human behavior and activities are so extensive that large samples are usually easy to obtain. Business, government, and educational records are quite complete and yield themselves readily to statistical analysis.

Improved methods of analysis have been developed which make more accurate results possible. The methods of multiple and partial correlation developed in Chapter 11 are applicable to social and business data. A quite flexible and profitable approach in this area is given by tabular analysis which is developed in Chapter 12. With the improved methods of sampling now available, and the more accurate measure of error now possible, excellent results are obtainable in the social sciences.

In most cases it is impossible to measure an entire universe

or population. To study the yield of all the millions of acres in wheat, or any particular characteristic of all the millions of school children, or the tensile strength of all wood, or the temperature and rainfall of every square mile of the earth's surface would not only be impossible in most cases but also very expensive in all cases. Such extensive studies would be a foolish waste of time and money. Equally accurate results for all practical purposes may be obtained from a relatively small, properly chosen sample with a minimum expenditure of time and money. The sampling of large populations is the most economical device known to man. To measure thoroughly all important characteristics of every bushel of wheat in a carload would cost more than the wheat is worth. A small well-mixed sample taken at each end and in the middle of the car of wheat will give approximately the same results as a total study of every bushel and give it at one-thousandth part of the cost. Sampling is literally the basis of all human decisions and is the foundation of much scientific analysis and knowledge.

THEORY OF SAMPLING

The possibility of effective sampling depends on the permanence, dependability, and uniformity of the nature and characteristics of the population sampled. If all populations varied without limit within themselves, no sampling would be valid. *It is only because we live in a world with relatively fixed characteristics and definitely limited variations that we can select a small portion of a population and still be certain that we have a fairly accurate picture of that population.* Wheat varies to some extent within itself in moisture content, weight, protein, etc., but it is always wheat. It is never cotton or grass or even oats. There are certain permanent and dependable characteristics of boys and girls which are universal and can be depended on to remain constant in a population of adolescents.

The dependability of samples and the adequacy of the sampling method in statistics depend upon at least five permanent characteristics of numbers. These are:

The Inertia of Large Numbers

If a number of relatively small samples are taken from a large population, all the samples will be found to be quite similar. This result proves that the large body of numbers, or the population, is a quite constant body. Its character, its movement, and its inertia are not easily changed.

The Permanence or Persistence of Small Numbers

If a population contains a variety of characteristics, a little of most or all of these several characteristics will be found in any random sample of that population. As illustrations of this principle, one may point out that if there is a mixture of red corn in a field of white corn, a red ear will be found at somewhat regular intervals throughout the field. The Mendelian distribution of recessive characteristics is a clear illustration of this principle.

Multiplicity of Forces

We live in a world in which a multiplicity of forces affects each observation. Every grain of wheat and every child is the result of a vast number of forces of heredity and environment. This fact explains why there is so much variation within families, communities, groups and samples, and why no sample will be a straight uniformity. "Variety is the spice of life," is a homely way of recognizing this universal characteristic of populations.

Independence of Forces

This vast complex of forces affecting all life and existence, although related, acts with a large degree of independence on individual observations and specimens. This fact explains why the children of the same parents, while quite similar in most respects, may vary considerably in some particular characteristics. It explains why a series of samples, while having much in common, may contain individual observations that are quite different. The basic law of existence is *diversity in unity.*

Equality of Forces

These forces are so balanced and related that they tend to generate equal values above and below the average. There appears to be no long-time bias toward smaller or larger values. If in a particular species, such as the horse, the evolution through the ages produces a larger animal, it seems to be not that there is a bias toward larger animals in the forces of heredity, but that only the larger ones actually survive in the difficult environment in which they must live. *In any case, at any particular time those individual cases which fall below the average are about the same in number as those which rise above the average.* The validity of sampling rests upon this universal tendency for the items of a population to cluster about the center of the range. They pile up in the middle with about equal scatter on either side. Of course, sampling procedure is appropriate to non-normal as well as to normal distributions.

Because of this random movement of these forces, it is possible by means of a random sample to obtain a reliable and realistic cross-section of a population. This is basic in statistics. *The foundation of all statistical analysis is the representativeness and dependability of random samples.*

METHOD OF SAMPLING

Statisticians generally recognize four different methods of sampling. The first is the *random sample.* By random sampling we mean the selection of a sample in such a manner *that every item in the population has an equal chance of being included in the sample.* There is no bias in true random sampling. No particular class or group is barred, minimized or magnified by any means. An example of bias is the selection of a sample of a city from its telephone directory. Most very poor families do not have telephones and are automatically eliminated. Persons living in hotels or rooming houses may not have a telephone. Transients would be excluded. Such a bias is fatal to the adequacy of the sample. The failure of the *Literary Digest* poll on the 1936 Presidential

election was caused by such a bias. It failed to include the lower economic levels of the population. Bias in sampling does not mean that there is a desire to misrepresent the facts. It means that there is some hidden and unwanted fault in the sampling that makes it unrepresentative of the population. The researcher must always be on the alert to discover such pitfalls, for they invalidate his entire work if they are large, and injure it under any conditions.

There are many situations when one must depend on getting a random sample. If our present knowledge of the universe is quite limited, if we are unacquainted with the subgroups of the population, we can do no more than take the best random sample possible under the circumstances. Suppose that we wish a sample of the workers on a large defense project. No enumeration has previously been made. We do not know whether they are young or old, rural or urban, white or Negro, native born or foreign born, citizens or aliens, skilled or unskilled. In fact, the reason for taking the sample is to discover as much information about them as possible. At present we have none. The most we can do is to use every precaution to get as good a random sample as possible. To do this the sample should not be taken at only one place or at only one time; for instance, at noon, or in the carpenter shop. The night workers may be quite different from the day workers, and the carpenters be very dissimilar to the welders. It should be taken from all over the plant during a twenty-four-hour period. A certain percentage or number of the total should be taken, as every fifth, or every tenth, or if the total is large, every fiftieth man, or one percent, or one-half, or one-tenth of one percent of the total. There should be an equal chance to include every one of the workers in the sample. A simple illustration would be the selection of red and white marbles from a box composed of 50 red and 50 white marbles of equal size. By shaking the box thoroughly, selecting one, recording its color, replacing it in the box to keep the total constant, shaking and drawing again; in a sample of 30 marbles one should have approximately half white and half red marbles. In dealing with human beings or anything less simple and uniform than marbles,

the process is somewhat more difficult, but should approximate pure random sampling as nearly as possible.

A second type is the *stratified sample.* When little or nothing is known about the internal structure of the universe, only general random sampling can be used. Often, however, there is available a considerable amount of information about certain characteristics of the population. Good illustrations of the former type are the crowds present at Independence Day celebrations, county fairs, general holiday picnics, the Mardi Gras, and the like. Little is known beforehand of the composition of such populations. Good examples of the latter type are student bodies of colleges, the membership of clubs, and the total census of a nation and its subdivisions. Whenever such information is available concerning a population, it should be used in making a sample of the population. By dividing the universe into strata according to known facts and sampling each stratum in proportion to its relative size in the population, not only can a more representative sample be obtained, but also a smaller sample will suffice.

Let us suppose that we are sampling the student body of a university to determine the average annual expenditure of students on picture shows. The catalogue or the report of the registrar's office will reveal the portion of the student body which falls into each of the following groups, freshmen, sophomores, juniors, seniors, graduate students, fraternity and sorority members, etc. If the total number of students is 10,000 and we design our sample to include 1% of each student group, our sample will contain exactly the same proportions of each group. A total sample of 100 chosen by this method would probably represent the entire student body better than a sample of two or three times that size chosen at random without any consideration of the internal structure of the population. It is sampling of this type which enables the *Gallup Poll* and *Fortune* so accurately to measure public opinion from a relatively small percentage of the total population. If our sample of the university student body was not organized in this way, it is possible that we might not get one graduate student in a sample of 100, and perhaps very few seniors.

The ubiquitous freshmen and sophomores might completely dominate it. *The statistician, therefore, should remember that whenever exact, specific knowledge is available concerning the internal structure of any population, he should take advantage of it in designing a sample for that population.* THE SAMPLING WITHIN THE SUBDIVISIONS OF THE POPULATION SHOULD BE TAKEN AT RANDOM.

A *purposive sample* is one selected to measure some particular phase of a population. A definite purpose or result is in mind. To attain this result, specific controls are required. Suppose that we wished a sample of a student body which would exactly conform to the entire student body on the measure of grade points. We would have to know first the number of students having an average of 1 grade point, the number having an average of 2 grade points, and so on, to the highest level. We would then select out of each grade-point level the same ratio of students. If the number of students on the four levels were 500, 1 point; 6,000, 2 points; 3,600, 3 points; and 400, 4 points, and if we chose a sample of 100 it would contain 5, 1 point; 60, 2 points; 36, 3 points; and 4, 4 points, such a sample as far as this one purpose of grade points is concerned would represent the entire population perfectly. This method has been used to good advantage in agricultural statistics where very exact knowledge existed concerning the population and one specific purpose was in view.

The fourth method, the *stratified purposive sample* is a combination of the two methods previously discussed, the *purposive* and the *stratified*. If we selected our sample to show grade-point averages according to the strata of the student body according to classes, societies, and other internal structures, it would be a *stratified purposive* sample. When it is the purpose of the statistician to make exact, detailed microscopic studies of a few limited characteristics of a universe, and sufficiently exact information is available, this method gives excellent results. Of course, within the limits of the strata and purposes, the *sampling is still to be taken at random and without bias.*

In any case a researcher, before he spends any considerable amount of time or money on a complex statistical analysis, should satisfy himself that the sample is adequate and dependable.

RANDOM SAMPLING BY USE OF A TABLE OF RANDOM NUMBERS

A table of random numbers such as Table 48 is a useful device in obtaining a sample with a maximum of pure random selection. This device is used as follows:

1. Design some system of consecutive numbers for each item in the population from which the sample is to be chosen. The numbering of the items of the population may be based on (1) size, (2) location, (3) time sequence, or (4) any suitable combination of these three bases. The names in a directory may be numbered in their alphabetical sequence, the houses in a city in their successive distances from a central point or central street, the sales slips in a store by the time sequence of the sales, etc.

2. Since the numbers in Table 48 are all distributed in purely random locations, one may begin by placing his pencil by chance on any number. For instance the pencil point might fall on the second digit (5) in the number 95348, in column (8) line 77 on page 375. This is the point of beginning for this problem.

3. If the numbers in the population run as high as four digits, 9999, five and the three digits in the number after 5 or 5348 may be taken as the first number in the sample. Following this each four successive digits in this line may be taken, as, 7866, 2112, 6381, 6515, 0245, etc., through all the lines in the table. Or the numbers may be taken on a diagonal line, as 5348, (dropping down one line and to the right) take 4534, 4079, 3180, 8705, 2799, etc. The random numbers may be taken from the table in any systematic orderly sequence, vertically, horizontally, or diagonally in any direction.

If the highest number in the population is 999, three digits may be used, or any other number that is sufficient for populations large or small. Whatever the random number selected may be, that number in the universe is taken.

4. Continue the process of selection until a sample of the desired size is taken. Such samples should have a high degree of randomness because they are selected according to a purely random pattern of numbers.

TABLE 48

A Table of 14,000 Random Units

Line \ Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	53342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953

Source: Table 48 is the first 4 pages of a 30-page table of 105,000 random digits prepared by the Bureau of Transport Economics and Statistics of the Interstate Commerce Commission, Washington, D.C., Mr. W. H. S. Stevens, Director. It is used in this text with their permission.

Line \ Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65795	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	73577	12908	30883	18317	28290	35797	05998	41688	34952	37888	38917	88050
36	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
37	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	83517	36103	42791
38	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
39	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89656	20103	77490	18062
41	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
42	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
43	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
44	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583
46	90725	52210	83974	29992	65831	38857	50490	83765	55657	14361	31720	57375	56228	41546
47	64364	67412	33339	31926	14883	24413	59744	92351	97473	89286	35931	04110	23726	51900
48	08962	00358	31662	25388	61642	34072	81249	35648	56891	69352	48373	45578	78547	81788
49	95012	68379	93526	70765	10592	04542	76463	54328	02349	17247	28865	14777	62730	92277
50	15664	10493	20492	38391	91132	21999	59516	81652	27195	48223	46751	22923	32261	85653
51	16408	81899	04153	53381	79401	21438	83035	92350	36693	31238	59649	91754	72772	02338
52	18629	81953	05520	91962	04739	13092	97662	24822	94730	06496	35090	04822	86774	98289
53	73115	35101	47498	87637	99016	71060	88824	71013	18735	20286	23153	72924	35165	43040
54	57491	16703	23167	49323	45021	33132	12544	41035	80780	45393	44812	12515	98931	91202
55	30405	83946	23792	14422	15059	45799	22716	19792	09983	74353	68668	30429	70735	25499
56	16631	35006	85900	98275	32388	52390	16815	69298	82732	38480	73817	32523	41961	44437
57	96773	20206	42559	78985	05300	22164	24369	54224	35083	19687	11052	91491	60383	19746
58	38935	64202	14349	82674	66523	44133	00697	35552	35970	19124	63318	29686	03387	59846
59	31624	76384	17403	53363	44167	64486	64758	75366	76554	31601	12614	33072	60332	92325
60	78919	19474	23632	27889	47914	02584	37680	20801	72152	39339	34806	08930	85001	87820

61	03931	33309	57047	74211	63445	17361	62825	39908	05607	91284	68833	25570	38818	46920
62	74426	33278	43972	10119	89917	15665	52872	73823	73144	88662	88970	74492	51805	99378
63	09066	00903	20795	95452	92648	45454	09552	88815	16553	51125	79375	97596	16296	66092
64	42238	12426	87025	14267	20979	04508	64535	31355	86064	29472	47689	05974	52468	16834
65	16153	08002	26504	41744	81959	65642	74240	56302	00033	67107	77510	70625	28725	34191
66	21457	40742	29820	96783	29400	21840	15035	34537	33310	06116	95240	15957	16572	06004
67	21581	57802	02050	89728	17937	37621	47075	42080	97403	48626	68995	43805	33386	21597
68	55612	78095	83197	33732	05810	24813	86902	60397	16489	03264	88525	42786	05269	92532
69	44657	66999	99324	51281	84463	60563	79312	93454	68876	25471	93911	25650	12682	73572
70	91340	84979	46949	81973	37949	61023	43997	15263	80644	43942	89203	71795	99533	50501
71	91227	21199	31935	27022	84067	05462	35216	14486	29891	68607	41867	14951	91696	85065
72	50001	38140	66321	19924	72163	09538	12151	06878	91903	18749	34405	56087	82790	70925
73	65390	05224	72958	28609	81406	39147	25549	48542	42627	45233	57202	94617	23772	07896
74	27504	96131	83944	41575	10573	08619	64482	73923	36152	05184	94142	25299	84387	34925
75	37169	94851	39117	89632	00959	16487	65536	49071	39782	17095	02330	74301	00275	48280
76	11508	70225	51111	38351	19444	66499	71945	05422	13442	78675	84081	66938	93654	59894
77	37449	30362	06694	54690	04052	53115	62757	95348	78662	11163	81651	50245	34971	52924
78	46515	70331	85922	38329	57015	15765	97161	17869	45349	61796	66345	81073	49106	79860
79	30986	81223	42416	58353	21532	30502	32305	86482	05174	07901	54339	58861	74818	46942
80	63798	64995	46583	09785	44160	78128	83991	42865	92520	83531	80377	35909	81250	54238
81	82486	84846	99254	67632	43218	50076	21361	64816	51202	88124	41870	52689	51275	83556
82	21885	32906	92431	09060	64297	51674	64126	62570	26123	05155	59194	52799	28225	85762
83	60336	98782	07408	53458	13564	59089	26445	29789	85205	41001	12535	12133	14645	23541
84	43937	46891	24010	25560	86355	33941	25786	54990	71899	15475	95434	98227	21824	19585
85	97656	63175	89303	16275	07100	92063	21942	18611	47348	20203	18534	03862	78095	50136
86	03299	01221	05418	38982	55758	92237	26759	86367	21216	98442	08303	56613	91511	75928
87	79626	06486	03574	17668	07785	76020	79924	25651	83325	88428	85076	72811	22717	50585
88	85636	68335	47539	03129	65651	11977	02510	26113	99447	68645	34327	15152	55230	93448
89	18039	14367	61337	06177	12143	46609	32989	74014	64708	00533	35398	58408	13261	47908
90	08362	15656	60627	36478	65648	16764	53412	09013	07832	41574	17639	82163	60859	75567
91	79556	29068	04142	16268	15387	12856	66227	38358	22478	73373	88732	09443	82558	05250
92	92608	82674	27072	32534	17075	27698	98204	63863	11951	34648	88022	56148	34925	57031
93	23982	25835	40055	67006	12293	02753	14827	23235	35071	99704	37543	11601	35503	85171
94	09915	96306	05908	97901	28395	14186	00821	80703	70426	75647	76310	88717	37890	40129
95	59037	33300	26695	62247	69927	76123	50842	43834	86654	70959	79725	93872	28117	19233

Line \ Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
96	42488	78077	69882	61657	34136	79180	97526	43092	04098	73571	80799	76536	71255	64239
97	46764	86273	63003	93017	31204	36692	40202	35275	57306	55543	53203	18098	47625	88684
98	03237	45430	55417	63282	90816	17349	88298	90183	36600	78406	06216	95787	42579	90730
99	86591	81482	52667	61582	14972	90053	89534	76036	49199	43716	97548	04379	46370	28672
100	38534	01715	94964	87288	65680	43772	39560	12918	86537	62738	19636	51132	25739	56947
101	13284	16834	74151	92027	24670	36665	00770	22878	02179	51602	07270	76517	97275	45960
102	21224	00370	30420	03883	94648	89428	41583	17564	27395	63904	41548	49197	82277	24120
103	99052	47887	81085	64933	66279	80432	65793	83287	34142	13241	30590	97760	35848	91983
104	00199	50993	98603	38452	87890	94624	69721	57484	67501	77638	44331	11257	71131	11059
105	60578	06483	28733	37867	07936	98710	98539	27186	31237	80612	44488	97819	70401	95419
106	91240	18312	17441	01929	18163	69201	31211	54288	39296	37318	65724	90401	79017	62077
107	97458	14229	12063	59611	32249	90466	33216	19358	02591	54263	88449	01912	07436	50813
108	35249	38646	34475	72417	60514	69257	12489	51924	86871	92446	36607	11458	30440	52639
109	38980	46600	11759	11900	46743	27860	77940	39298	97838	95145	32378	68038	89351	37005
110	10750	52745	38749	87365	58959	53731	89295	59062	39404	13198	59960	70408	29812	83126
111	36247	27850	73958	20673	37800	63835	71051	84724	52492	22342	78071	17456	96104	18327
112	70994	66986	99744	72438	01174	42159	11392	20724	54322	36923	70009	23233	65438	59685
113	99638	94702	11463	18148	81386	80431	90628	52506	02016	85151	88598	47821	00265	82525
114	72055	15774	43857	99805	10419	76939	25993	03544	21560	83471	43989	90770	22965	44247
115	24038	65541	85788	55835	38835	59399	13790	35112	01324	39520	76210	22467	83275	32286
116	74976	14631	35908	28221	39470	91548	12854	30166	09073	75887	36782	00268	97121	57676
117	35553	71628	70189	26436	63407	91178	90348	55359	80392	41012	36270	77786	89578	21059
118	35676	12797	51434	82976	42010	26344	92920	92155	58807	54644	58581	95331	78629	73344
119	74815	67523	72985	23183	02446	63594	98924	20633	58842	85961	07648	70164	34994	67662
120	45246	88048	65173	50989	91060	89894	36036	32819	68559	99221	49475	50558	34698	71800
121	76509	47069	86378	41797	11910	49672	88575	97966	32466	10083	54728	81972	58975	30761
122	19689	90332	04315	21358	97248	11188	39062	63312	52496	07349	79178	33692	57352	72862
123	42751	35318	97513	61537	54955	08159	00337	80778	27507	95478	21252	12746	37554	97775
124	11946	22681	45045	13964	57517	59419	58045	44067	58716	58840	45557	96345	33271	53464
125	96518	48688	20996	11090	48396	57177	83867	86464	14342	21545	46717	72364	86954	55580
126	35726	58643	76869	84622	39098	36083	72505	92265	23107	60278	05822	46760	44294	07672
127	39737	42750	48968	70536	84864	64952	38404	94317	65402	13589	01055	79044	19308	83623
128	97025	66492	56177	04049	80312	48028	26408	43591	75528	65341	49044	95495	81256	53214
129	62814	08075	09788	56350	76787	51591	54509	49295	85830	59860	30883	89660	96142	18354
130	25578	22950	15227	83291	41737	79599	96191	71845	86899	70694	24290	01551	80092	82118

131	68763	69576	88991	49662	46704	63362	56625	00481	73323	91427	15264	06969	57048	54149
132	17900	00813	64361	60725	88974	61005	99709	30666	26451	11528	44323	34778	60342	60388
133	71944	60227	63551	71109	05624	43836	58254	26160	32116	63403	35404	57146	10909	07346
134	54684	93691	85132	64399	29182	44324	14491	55226	78793	34107	30374	48429	51376	09559
135	25946	27623	11258	65204	52832	50880	22273	05554	99521	73791	85744	29276	70326	60251
136	01353	39318	44961	44972	91766	90262	56073	06606	51826	18893	83448	31915	97764	75091
137	99083	88191	27662	99113	57174	35571	99884	13951	71057	53961	61448	74909	07322	80960
138	52021	45406	37945	75234	24327	86978	22644	87779	23753	99926	63898	54886	18051	96314
139	78755	47744	43776	83098	03225	14281	83637	55984	13300	52212	58781	14905	46502	04472
140	25282	69106	59180	16257	22810	43609	12224	25643	89884	31149	85423	32581	34374	70873
141	11959	94202	02743	86847	79725	51811	12998	76844	05320	54236	53891	70226	38632	84776
142	11644	13792	98190	01424	30078	28197	55583	05197	47714	68440	22016	79204	06862	94451
143	06307	97912	68110	59812	95448	43244	31262	88880	13040	16458	43813	89416	42482	33939
144	76285	75714	89585	99296	52640	46518	55486	90754	88932	19937	57119	23251	55619	23679
145	55322	07598	39600	60866	63007	20007	66819	84164	61131	81429	60676	42807	78286	29015
146	78017	90928	90220	92503	83375	26986	74399	30885	88567	29169	72816	53357	15428	86932
147	44768	43342	20696	26331	43140	69744	82928	24988	94237	46138	77426	39039	55596	12655
148	25100	19336	14605	86603	51680	97678	24261	02464	86563	74812	60069	71674	15478	47642
149	83612	46623	62876	85197	07824	91392	58317	37726	84628	42221	10268	20692	15699	29167
150	41347	81666	82961	60413	71020	83658	02415	33322	66036	98712	46795	16308	28413	05417
151	38128	51178	75096	13609	16110	73533	42564	59870	29399	67834	91055	89917	51096	89011
152	60950	00455	73254	96067	50717	13878	03216	78274	65863	37011	91283	33914	91303	49326
153	90524	17320	29832	96118	75792	25326	22940	24904	80523	38928	91374	55597	97567	38914
154	49897	18278	67160	39408	97056	43517	84426	59650	20247	19293	02019	14790	02852	05819
155	18494	99209	81060	19488	65596	59787	47939	91225	98768	43688	00438	05548	09443	82897
156	65373	72984	30171	37741	70203	94094	87261	30056	58124	70133	18936	02138	59372	09075
157	40653	12843	04213	70925	95360	55774	76439	61768	52817	81151	52188	31940	54273	49032
158	51638	22238	56344	44587	83231	50317	74541	07719	25472	41602	77318	15145	57515	07633
159	69742	99303	62578	83575	30337	07488	51941	84316	42067	49692	28616	29101	03013	73449
160	58012	74072	67488	74580	47992	69482	58624	17106	47538	13452	22620	24260	40155	74716
161	18348	19855	42887	08279	43206	47077	42637	45606	00011	20662	14642	49984	94509	56380
162	59614	09193	58064	29086	44385	45740	70752	05663	49081	26960	57454	99264	24142	74648
163	75688	28630	39210	52897	62748	72658	98059	67202	72789	01869	13496	14663	87645	89713
164	13941	77802	69101	70061	35460	34576	15412	81304	58757	35498	94830	75521	00603	97701
165	96656	86420	96475	86458	54463	96419	55417	41375	76886	19008	66877	35934	59801	00497

Line / Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
166	03363	82042	15942	14549	38324	87094	19069	67590	11087	68570	22591	65232	85915	91499
167	70366	08390	69155	25496	13240	57407	91407	49160	07379	34444	94567	66035	38918	65708
168	47870	36605	12927	16043	53257	93796	52721	73120	48025	76074	95605	67422	41646	14557
169	79504	77606	22761	30518	28373	73898	30550	76684	77366	32276	04690	61667	64798	66276
170	46967	74841	50923	15339	37755	98995	40162	89561	69199	42257	11647	47603	48779	97907
171	14558	50769	35444	59030	87516	48193	02945	00922	48189	04724	21263	20892	92955	90251
172	12440	25057	01132	38611	28135	68089	10954	10097	54243	06460	50856	65435	79377	53890
173	32293	29938	68653	10497	98919	46587	77701	99119	93165	67788	17638	23097	21468	36992
174	10640	21875	72462	77981	56550	55999	87310	69643	45124	00349	25748	00844	96831	30651
175	47615	23169	39571	56972	20628	21788	51736	33133	72696	32605	41569	76148	91544	21121
176	16948	11128	71624	72754	49084	96303	27830	45817	67867	18062	87453	17226	72904	71474
177	21258	61092	66634	70335	92448	17354	83432	49608	66520	06442	59664	20420	39201	69549
178	15072	48853	15178	30730	47481	48490	41436	25015	49932	20474	53821	51015	79841	32405
179	99154	57412	09858	65671	70655	71479	63520	31357	56968	06729	34465	70685	04184	25250
180	08759	61089	23706	32994	35426	36666	63988	98844	37533	08269	27021	45886	22835	78451
181	67323	57839	61114	62192	47547	58023	64630	34886	98777	75442	95592	06141	45096	73117
182	09255	13986	84834	20764	72206	89393	34548	93438	88730	61805	78955	18952	46436	58740
183	36304	74712	00374	10107	85061	69228	81969	92216	03568	39630	81869	52824	50937	27954
184	15884	67429	86612	47367	10242	44880	12060	44309	46629	55105	66793	93173	00480	13311
185	18745	32031	35303	08134	33925	03004	59929	95418	04917	57596	24878	61733	92834	64454
186	72934	40086	88292	65728	38300	42323	64068	98373	48971	09049	59943	36538	05976	82118
187	17626	02944	20910	57662	80181	38579	24580	90529	52303	50436	29401	57824	86039	81062
188	27117	61399	50967	41399	81636	16663	15634	79717	94696	59240	25543	97989	63306	90946
189	93995	18678	90012	63645	85701	85269	62263	68331	00389	72571	15210	20769	44686	96176
190	67392	89421	09623	80725	62620	84162	87368	29560	00519	84545	08004	24526	41252	14521
191	04910	12261	37566	80016	21245	69377	50420	85658	55263	68667	78770	04533	14513	18099
192	81453	20283	79929	59839	23875	13245	46808	74124	74703	35769	95588	21014	37078	39170
193	19480	75790	48539	23703	15537	48885	02861	86587	74539	65227	90799	58789	96257	02708
194	21456	13162	74608	81011	55512	07481	93551	72189	76261	91206	89941	15132	37738	59284
195	89406	20912	46189	76376	25538	87212	20748	12831	57166	35026	16817	79121	18929	40628
196	09866	07414	55977	16419	01101	69343	13305	94302	80703	57910	36933	57771	42546	03003
197	86541	24681	23421	13521	28000	94917	07423	57523	97234	63951	42876	46829	09781	58160
198	10414	96941	06205	72222	57167	83902	07460	69507	10600	08858	07685	44472	64220	27040
199	49942	06683	41479	58982	56288	42853	92196	20632	62045	78812	35895	51851	83534	10689
200	23995	68882	42291	23374	24299	27024	67460	94783	40937	16961	26053	78749	46704	21983

ADEQUACY AND DEPENDABILITY OF SAMPLES

Two devices have been developed for determining the degree to which a sample represents a population, (1) the degree of uniformity of means as samples are increased in size, and (2) the standard error of the mean of the samples.

Increasing the size of the sample. It is quite evident that one item taken from a large universe, one child from ten million, one shovel of ore from a large mine, one quart of wheat from an elevator filled with two million bushels, will not likely represent the large universe accurately. Likewise, a sample of two items, or ten or even fifty items, may not measure accurately the many qualities and variations of the large population. But as the sample is increased in size from 100 to 1000 to 10,000 or 100,000 items, the degree of accuracy with which the statistic computed from the sample approaches the parameter of the universe will increase. The deviation or error between them will decline. If the sample can be expanded to include the entire universe, the two measures would be identical.

This device of measuring the dependability or representativeness of a statistic is frequently employed as follows: a large sample will be taken from the population and its mean and standard deviation computed. The sample will then be split in half by random choice of items, and the mean and standard deviation of the two smaller samples computed. If the errors or variations among the three samples, the original large sample and its two halves, are small, the assumption is that the sample represents the population well. If the errors are large, doubt as to the validity of the statistics arises. This device may be reversed and after one sample of quite large size is taken another sample twice or even four times as large may be taken and their statistics compared. If it is possible still larger samples may be taken and the results compared until a size of sample is reached after which an increase in sample size changes the values of the statistics but little. In certain types of farm sampling by the Department of Agriculture, it was discovered that after a sample of 50,000 items was taken little change in the results followed although other samples of much larger size were used.

This device of reducing the error in samples by increasing their size involves three difficulties. In the first place the principal purpose in using samples is to reduce costs. Since the reliability of samples increases in direct ratio to the square root of their size, a sample must be increased a large amount to reduce its error by a small degree.

Sample.	Root	Ratio
$\sqrt{100}$	10	1
$\sqrt{400}$	20	2
$\sqrt{900}$	30	3
$\sqrt{1600}$	40	4

One must add 1500 items to a sample of 100 items to increase its reliability by four times. Small samples are therefore much more dependable in direct proportion to their size than are large ones. Since the cost of obtaining a sample tends to increase in direct proportion to its size, cost rises faster than reliability. Except in cases in which usable reliable secondary data are easily available in large quantities, large samples are prohibitive as to cost.

The second difficulty is that there are many cases in which it is all but impossible to get large samples regardless of cost. In laboratory experiments in chemistry, physics, and psychology and in field tests in agriculture, advertising, and marketing, as well as in many types of personnel, educational and managerial problems, samples of even a few hundred or a thousand items could not be obtained in months or even years. If one had to depend on size of sample as a test of reliability in such cases, no adequate measure could be obtained.

The third inadequacy of this device is that there is no stable ratio against which one can compare sample dependability. Because of the inherent nature of chance sampling variations, one cannot be confident that any given size of sample has reduced the error of the statistic of the sample to a definite ratio to the parameter of the population. Two or more samples of successively increasing size often show a high degree of stability in means, only to have larger samples show higher or lower values. One often cannot be certain whether he has attained final stability or is just

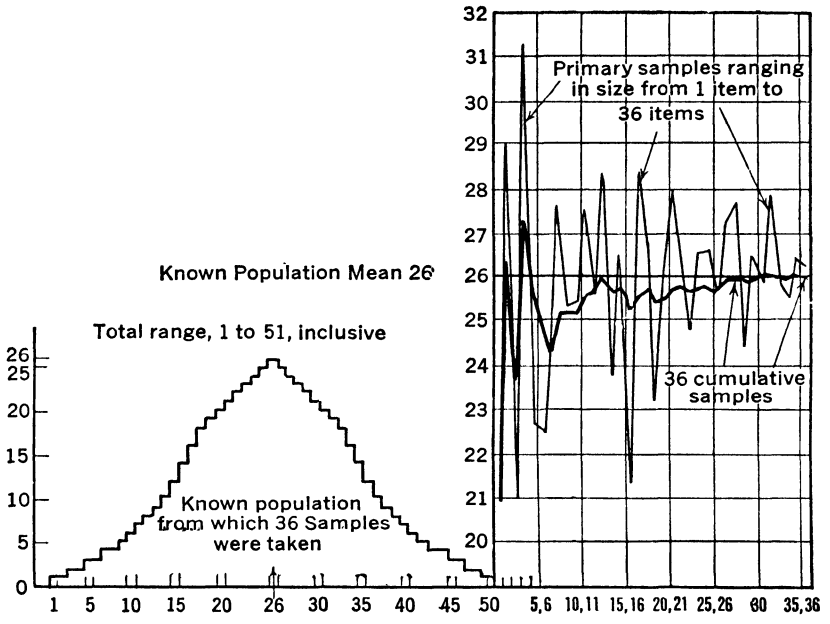


FIG. 77. Variation in size of successive samples from the same known population

on a temporary plateau which could be invalidated by still increasing sample size. These variations are illustrated in Figure 77.

One may illustrate the difficulty of measuring representativeness of samples from increasing size by creating an artificial universe of known size and parameters and by selecting from it successive samples of increasing size.

The universe in Figure 77 is symmetrical in the approximate form of a normal curve and contains 579 items whose sum is 15,054 and mean is 26. From this universe, 36 samples were drawn. The first sample contained one item, the second contained two and so on, each successive sample being one item larger than the previous one until 36 were drawn. The means of the first five samples range from 21.00 to 32.25. The means of the next five samples range from 22.50 to 27.62. The third five have a narrower spread of 23.71 to 27.36. The fourth five are still wider, 21.25 to 28.24. The fifth five scatter from 24.74 to 27.86. The sixth five are a little

closer, 25.44 to 27.78. But the thirty-second mean of thirty-two items is the largest mean after the seventeenth and is the sixth largest of the thirty-six. The twentieth and thirty-sixth means are nearest to the parameter of the universe which is twenty-six. There is no similarity of successive means sufficient to determine from that source *alone* the probable representativeness of any particular mean. To arrive at that decision, still larger samples would have to be drawn.

A second set of samples based on cumulation of previous samples indicates that (1) the larger successive samples are the less their means vary and (2) the nearer their means approach the mean of the universe. In this array of cumulative samples, the variation is wide until after the twentieth sample is passed. A further difficulty in judging sample dependability by this device is that four successive samples fall within a range of .09 between sample 21 and sample 24. One might well be tempted to conclude that approximately 25.7 was the true mean of the universe if this were the only information available. This presumption is destroyed, however, when the last five cumulative samples show means of from 25.96 to 26.00.

The last one of these cumulative samples contains 666 items which is a large sample. Not until after the twentieth cumulative sample which contains 210 items does the error fall below 2.0%. The sixteenth cumulative sample with 136 items shows a deviation of 3.26%.

One may conclude from the above facts and relationships that this method of attaining and measuring representativeness of samples and statistics is not practicable except in cases in which large quantities of secondary data are easily available. The cost of primary data especially if obtained by laboratory methods would be prohibitive.

If a second set of 36 samples of identical size with those above were taken from this population, there is no assurance that any one of the second set would have a mean identical with the mean of any one of the first set. Two or three of the means might be identical but there is no assurance that they would be. One is certain that there would be wide variations among many of them.

WORKSHEET NO. 62

SAMPLE SUMS AND MEANS, SIMPLE AND CUMULATIVE, DRAWN AT RANDOM FROM THE UNIVERSE IN FIGURE

Sample N	Sums ΣX	Means X	Cumulative Samples	Cumulative N	Means of Cumulative Totals
1	21	21.00	21	1	21.00
2	58	29.00	79	3	26.33
3	63	21.00	142	6	23.67
4	129	32.25	271	10	27.10
5	113	22.60	384	15	25.60
6	135	22.50	519	21	24.71
7	163	23.29	682	28	24.36
8	221	27.62	903	36	25.08
9	227	25.22	1,130	45	25.11
10	255	25.50	1,385	55	25.18
11	301	27.36	1,686	66	25.54
12	305	25.42	1,991	78	25.53
13	366	28.15	2,357	91	25.90
14	332	23.71	2,689	105	25.61
15	394	26.27	3,083	120	25.69
16	340	21.25	3,423	136	25.17
17	480	28.24	3,903	153	25.51
18	481	26.72	4,384	171	25.64
19	439	23.10	4,823	190	25.38
20	523	26.15	5,346	210	25.46
21	585	27.86	5,931	231	25.67
22	569	25.86	6,500	253	25.69
23	569	24.74	7,069	276	25.61
24	636	26.50	7,705	300	25.68
25	664	26.56	8,369	325	25.75
26	667	25.65	9,036	351	25.74
27	731	27.07	9,767	378	25.84
28	777	27.75	10,544	406	25.97
29	705	24.31	11,249	435	25.85
30	791	26.37	12,040	465	25.89
31	801	25.84	12,841	496	25.88
32	889	27.78	13,730	528	26.00
33	853	25.85	14,583	561	25.99
34	865	25.44	15,448	595	25.96
35	922	26.34	16,370	630	25.98
36	941	26.14	17,311	666	25.99

What probability is there then that the mean of any particular sample represents the population well?

The mathematical illustrations and measurements of probability presented above including the binomial expansion are applicable primarily to large numbers and to groups of statistics rather than to any specific sample or statistic. In the case of $(a + b)^2$, or of two coins thrown, the formula is $1a^2 + 2ab + 1b^2$ in which there are four and only four possible combinations of falls. These combinations are (1) both coins might fall heads, (2) both coins might fall tails, (3) coin a might fall heads and coin b fall tails, or (4) coin a might fall tails and coin b fall heads. There are and can be only these four relationships.

If these two coins are thrown a large number of times, 100 or 500 or whatever number of times that is necessary to balance out the opposite but compensating chance variations of sampling, the ratios of falls will closely follow the binomial coefficients. But there is little likelihood that only four throws would produce the four relationships. Instead of a normal distribution, such results as $2 + 1 + 1$, or $0 + 2 + 2$, or possibly $0 + 4 + 0$, or $3 + 0 + 1$ might result in any single series of four throws. If five or six or even ten coins were thrown, one single series of throws would most likely depart radically from the specific possible combinations of the expanded binomial. There is no compulsion in the binomial expansion necessitating that the chance events of a given sample shall fall according to its limited and fixed relationship. To return to von Mises postulate, *the probability of an event is the limit, or ratio, which is approached in a prolonged series of observations.* Measures of probability are, therefore, dependent on *prolonged series* of events and are applicable to single specific samples only as parts of such series.

Measurements of Standard Error

Although the continuous expansion of the size of sample is usually not a satisfactory device for determining the representativeness of a sample or the statistics computed from it, there is available in Worksheet No. 62 information which may be used to compute a satisfactory measure of the dependability of a sample.

WORKSHEET NO. 63

ARRAY OF MEANS OF 36 SUCCESSIVE SAMPLES DRAWN FROM WORKSHEET NO. 63

21.00	23.29	25.44	26.14	26.56	27.78
21.00	23.71	25.50	26.15	26.72	27.86
21.25	24.31	25.65	26.27	27.07	28.15
22.50	24.74	25.84	26.34	27.36	28.24
22.60	25.22	25.85	26.37	27.62	29.00
23.10	25.42	25.86	26.50	27.75	32.25

WORKSHEET NO. 64

FREQUENCY DISTRIBUTION OF SAMPLE MEANS FROM WORKSHEET NO. 63

20.99 to 22.99	𝍸	5
22.99 to 24.99	𝍸	5
24.99 to 26.99	𝍸𝍸𝍸 1	16
26.99 to 28.99	𝍸 111	8
28.99 to 30.99	1	1
30.99 to 32.99	1	1

The two significant facts about this information are; (1) that these sample means do fall in a frequency distribution, and (2) that the central concentration is high. Here is dependable information which when fully developed will enable us to measure the representativeness of a mean.

This particular series of arithmetic means is partially inadequate for this purpose because they are for samples of varying size, and are, therefore, unequally representative of the population for two reasons; (1) chance sampling, and (2) variation in size. This limitation can be largely overcome by selecting a series of samples of identical size from the same universe.

In Worksheet No. 65 are recorded the means of sixty random samples of fifty items each taken from the population of semi-monthly wage payments of 269 oil refinery employees. These samples are identical in size and origin. They, therefore, give a good demonstration of the tendency of a series of sample means to fall in a normal curve.

WORKSHEET NO. 65

THE ARITHMETIC MEANS OF SIXTY RANDOM SAMPLES OF FIFTY ITEMS, EACH DRAWN FROM THE POPULATION OF SEMI-MONTHLY WAGES OF 269 OIL REFINERY EMPLOYEES

164.60	166.35	167.19	167.72	168.26	169.18
164.85	166.39	167.30	167.73	168.31	169.26
165.06	166.48	167.33	167.82	168.39	169.41
165.33	166.61	167.37	167.84	168.43	169.51
165.50	166.69	167.40	167.92	168.54	169.59
165.62	166.76	167.44	168.03	168.56	169.67
165.86	166.84	167.48	168.05	168.84	170.04
166.00	166.89	167.56	168.14	168.92	170.49
166.18	166.95	167.62	168.22	168.97	170.83
166.26	166.97	167.64	168.24	169.13	171.13

The mean of the 60 means = 167.73

The sum of the squared deviations of the 60 means from their mean (167.73) = 125.9528

$\sigma_{\bar{x}}$ of the series of 60 means = 1.449

$$\sigma_{\bar{x}} = \frac{\sqrt{125.9528}}{60} = \sqrt{2.099213} = 1.449$$

WORKSHEET NO. 66

FREQUENCY DISTRIBUTION OF THE 60 MEANS IN WORKSHEET NO. 65 (See Fig. 78)

164.20 to 165.20	111	3
165.20 to 166.20	𝍸 1	6
166.20 to 167.20	𝍸 𝍸 111	13
167.20 to 168.20	𝍸 𝍸 𝍸 11	17
168.20 to 169.20	𝍸 𝍸 11	12
169.20 to 170.20	𝍸 1	6
167.20 to 171.20	111	3
		60

The most probable value of the mean of a population is the mean of several means of random samples taken from that population.

Since the means of a number of samples of identical size taken

from the same population fall in a small smooth normal curve, the standard deviation of that small frequency distribution of means may be used as a measure of the error of those means. One standard deviation of those means plus and minus from the population mean will include about 68.27 percent of the sample means. Two $\sigma_{\bar{x}} \pm$ from the mean will measure the range in which 95.45 percent of the sample means will fall, and $3\sigma_{\bar{x}} \pm$ will cover the range within which 99.73 percent of such means will be.

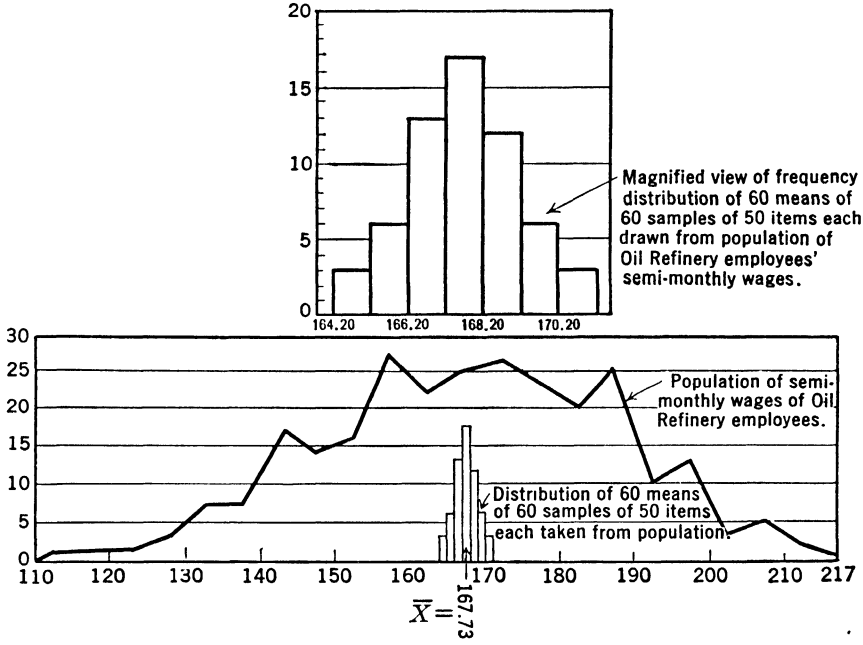


FIG. 78. Relationship of distribution of 60 sample means of 60 samples of 50 items each taken from the population of oil refinery semi-monthly wages, to that population

The two facts illustrated in Figure 78, that (1) the mean of the series of sample means is the most probable value of the mean of the population, and (2) that these means fall in a narrow normal curve about the true mean make it possible to compute the degree of probability that any sample mean lies within a given portion of the range, or deviates from the population mean by only as

much or less than a given value for a specific degree or percentage of confidence.

Standard Error of the Mean

FORMULA NO. 100

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N-1}}$$

In this case the values are:

$$\sigma_{\bar{x}} = \frac{20.60}{\sqrt{188-1}} = \frac{20.60}{13.675} = 1.506$$

This value computed from the formula for the standard error of the mean is within .007 of the standard deviation of the 60 means which is 1.449.

$$1.506 - 1.449 = .007$$

A close approximation of the true standard error can be obtained by computing the mean and standard deviation for any adequate random sample of the population. The process introduces the error of substituting the standard deviation of the sample for the true standard deviation of the universe, as is done above; but with a sample of fair size this error is small.

In the case of the wages of the oil refinery employees in Chapter 6, the explanations are:

$$\begin{aligned} \bar{X} &= \$167.87 \\ \sigma_x &= \$\ 20.60 \\ \sigma_{\bar{x}} &= \quad 1.506 \\ \bar{X} + \sigma_{\bar{x}} &= \quad 166.37 \text{ to } 169.37 \end{aligned}$$

Since there is no way of knowing from this single sample whether its mean of $167.87 is one of the few extreme means that might be drawn from the universe, or is one of the middle means near the true mean of the universe, one cannot say that the computation $1.5\sigma_{\bar{x}} \pm$, or, 166.37 to 169.37 indicates that the chances are 68.27 out of 100 that the true mean lies within this range. The standard

error of the mean computed from any sample of the same size from this population would be approximately the same whether the sample were a rare and extreme one or not. In any case for a sample of 188 items from this universe it would be approximately 1.5; but if the mean of the particular sample is either an extremely small one or extremely large one, a single $\sigma_{\bar{x}} \pm$ may not extend from the sample mean to include the population mean.

If we take either one of two extreme samples, 164.5 and 170.5, the range of error plus and minus would be,

$$164.5 \pm 1.5 = 163.00 \text{ to } 166.00$$
$$170.5 \pm 1.5 = 169.00 \text{ to } 172.00$$

In neither case does one standard error reach to the true mean which is known in this case to be approximately 167.73. There is a gap in the range from 166 to 169 which is not covered. To be certain of extending our computation of error to the mean of the population one should measure $2\sigma_{\bar{x}} \pm$, as

$$164.5 \pm 2 \times 1.5 - 161.5 \text{ to } 167.5$$
$$170.5 \pm 2 \times 1.5 - 167.5 \text{ to } 173.5$$

For the smaller mean the top range of error 167.5 does not reach the true mean of 167.73 but for the large mean it does.

Since the true mean of the population is not a variable but is fixed (in this case at 167.73), one should not say that the parameter falls within a given range of error but rather that the probabilities are, so and so, that the sample mean lies within a certain distance of the parameter with a specific degree of confidence. In the case of the larger mean above, it is probable that with a range of confidence of $2\sigma_{\bar{x}} \pm$, or 95.45 out of 100 that the sample mean lies within $3.\sigma_{\bar{x}}$ of the parameter. If one chose a higher degree of certainty, 99.73 out of 100, that the sample mean falls within $4.5\sigma_{\bar{x}}$ of the parameter, there would be very little chance that the measure of error was inadequate.

In fact a range of error of 1%, which is covered by $2.58\sigma_{\bar{x}} \pm$, is considered quite adequate for most problems. A degree of 2% error is covered by $2.33\sigma_{\bar{x}} \pm$ and a 5% error by $1.96\sigma_{\bar{x}} \pm$.

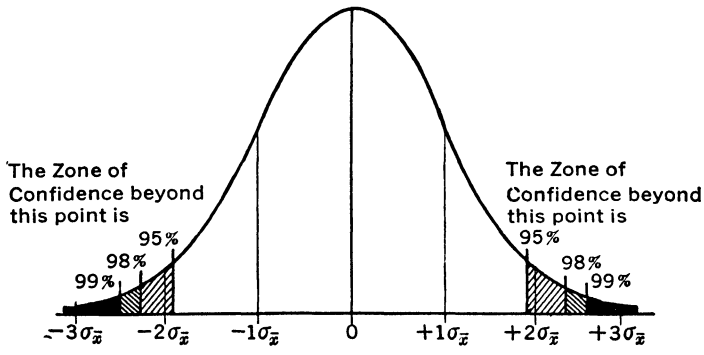


FIG. 79. Showing the significance of the range of Standard Errors about the mean and indicating location and size of Zones of Confidence. See Appendix Table II

Standard Error of Standard Deviation [1]

A similar measure may be computed for the standard deviation of a sample as follows:

FORMULA NO. 101

$$\sigma_\sigma = \frac{\sigma_x}{\sqrt{2(N-1)}} = \frac{20.60}{\sqrt{2(188-1)}} = \frac{20.60}{19.339} = 1.065$$

Standard Error of Coefficient of Correlation

The dependability of the coefficient of correlation may be roughly computed as follows:

FORMULA NO. 102

$$\sigma_r = \frac{1-r^2}{\sqrt{N-1}}$$

[1] The symbol, t, indicates the ratio of a statistical measure which is distributed normally around a hypothetical mean usually taken as zero to an estimate of the standard error of that measure based on the number of degrees of freedom involved. It has many uses of which that indicated here is only one. The point that the student should remember here is that t indicates a *ratio between a statistical measure and its standard error.* Fuller explanation will be given in Chapter 20, which treats of small samples.

The large T indicates a *normal deviate.* A normal deviate is the deviation of a statistical measurement from the mean of a normal distribution expressed in units of its *standard deviation.*

In the case of the correlation between Per Capita Bank Deposits, (X) and Per Capita Income Payments, (Y) in Worksheet No. 31 the value is:

$$\frac{1. - .8046}{\sqrt{25 - 1}} = \frac{.1954}{4.8990} = .04. \quad .8964 \pm .04, \text{ or } .8564 - .9364.$$

LESS FREQUENTLY USED STANDARD ERRORS

1. Standard Error of the Median:

$$\sigma_{Md} = \sqrt{\frac{\pi}{2N}}\,\sigma_x = 1.25331\,\sigma_{\bar{x}}$$

2. Standard Error of a Percentage:

$$\sigma_{\%} = \sqrt{\frac{Pq}{N}} = \sqrt{\frac{Pq}{N}}$$

3. Standard Error of Skewness:

$$\sigma_{SK} = \frac{1.225}{\sqrt{N}}$$

4. Standard Error of The Alphas:

$$\sigma_{\alpha_3} = \sqrt{\frac{6}{N}}$$

$$\sigma_{\alpha_4} = \sqrt{\frac{24}{N}} = 2\sigma_{\alpha_3}$$

This method of computing the standard error of the coefficient of correlation is not valid unless (1) the value of r is small (near zero) and (2) the size of the sample is large. As was pointed out in the explanation of probability any valid use of a standard error rests on the assumption that when any given statistic is computed for a series of samples from the same population these successive statistics fall in the form of a normal distribution. This result cannot occur unless the correlation of the population approaches zero. Since the maximum of correlation is either a + 1 or a − 1, the distribution must be highly skewed if r is large. If the true r of a population is + .94, the distribution of the r's

of 100 samples from this population must be highly skewed to the left, because r cannot be more than a + 1.0 but can be a − 1.0. On the right hand side the distribution can vary in a range of only .06 or from + .94 to + 1.0, but on the left side it may vary through a range of 1.94, or from − 1.0 to + .94. Such a distribution is too highly skewed for the relationship of the normal curve to be valid in its measurement. Also if the sample is quite small, the distribution of the r's will be erratic.

In such cases, (1) when r is large, or (2) when N is small, the measurement of the significance of r may be obtained by the computation of t[1] from the formula

FORMULA NO. 103

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

By this method one may compute the probability of r differing significantly from zero.

$$t = \frac{.94\sqrt{27-2}}{\sqrt{1-.94^2}} = \frac{4.70}{.34} = 13.82$$

but

$$t = \frac{.30\sqrt{27-2}}{\sqrt{1.-.30^2}} = \frac{1.50}{.974} = 1.54$$

For a sample of 27 items and $r = .96$ the highly significant[1] value of $t = 2.787$. The computed value of 13.82 is five times larger than the required 2.787. But when $N = 27$ and $r = .30$ the computed value of t is only 1.54 which is too small to prove that r is significant.

But when N is large,

$$t = \frac{.30\sqrt{402-2}}{\sqrt{1-.30^2}} = \frac{6.00}{.974} = 6.16$$

a correlation of .30 is highly significant. The required value in this table is 2.576 which is less than one-half of our computed value.

[1] See Table 50, Chapter 20.

Transformation of r *to* z

The *t*-test is applicable only for determining whether the computed *r* is significantly different from zero. R. A. Fisher has devised a method for testing the significance (1) between any two correlation coefficients, and (2) between a computed *r* and any theoretical value.

A simplified statement of the formula is

$$z = 1.15129 \text{ Log } 10 \left(\frac{1+r}{1-r}\right)$$

$$\text{If } r = .50,\ z = 1.15129 \text{ Log} \left(\frac{1+.50}{1-.50}\right) = 1.15129 \text{ Log} \left(\frac{1.50}{.50}\right)$$

$$= 1.5129 \text{ Log } 3.0 = 1.15129 \times .477121 = .5493$$

$$\text{If } r = .80,\ z = 1.15129 \text{ Log} \left(\frac{1+.80}{1-.80}\right) = 1.15129 \text{ Log} \left(\frac{1.80}{.20}\right)$$

$$= 1.15129 \text{ Log } 9.0 = 1.15129 \times .954243 = 1.0986$$

When $r = 0$, $z = 0$, but because z is based on logarithms the z-values increase at a faster rate than the r-values.

z-values Uniform			*r-values Uniform*		
z-values	for	*r*-values	*z*-values	for	*r*-values
.5		.4621	.2025		.20
1.0		.7616	.4237		.40
1.5		.9052	.6931		.60
2.0		.9640	1.0986		.80
2.5		.9866	1.4721		.90
3.0		.9951	1.8460		.95
			Infinity		1.00

The z-values conform to the normal curve. $\sigma_z = \frac{1}{\sqrt{N-3}}$ is the formula for the standard error of z. The z-equivalent for most r-values may be read from Appendix Table III.

Difference of Two Means [1]

In comparing the means of two samples the question arises as to whether the two samples could have been obtained from the same

[1] See Chapter 20 for full analysis.

parent population or whether they represent two separate populations. If they are from the same population, the means will differ no more than the variation that is determined by chance random sampling, but if their difference is greater than that which could result from chance sampling the samples must be considered as having been drawn from two separate populations.

The method of testing this similarity of sample means is to compare the difference of the two means with the square root of the sum of their two squared standard errors of their means.

FORMULA No. 104

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_D}$$

Meaning of symbols:

T = the difference of two means expressed as a normal deviate

σ_D = square root of the sum of the two squared standard errors of the means of the samples

FORMULA No. 105

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{N-1} + \frac{\sigma_2^2}{N-1}}, \quad \text{or} \quad \sigma_D = \sqrt{\sigma_{\bar{x}1}^2 + \sigma_{\bar{x}2}^2}$$

If T is more than 2.63, the chances are only 1 out of 100 that the two samples are from the same parent population. If T is only 2.0, the chances are only 1 out of 21 that the two samples were drawn from the same universe. Values of less than 2.0 are not very dependable. They do not indicate clearly that the samples are from separate populations. These ratios apply only to large samples of 100 or more items. For smaller samples the ratios must be larger.

Probable Error

The probable error of a mean is 0.6745 $\sigma_{\bar{x}}$. It measures a 50–50 probability while the standard error measures 68–32 probability. The latter is the more desirable measure.

These are the more common measures of error. Other and more complex measures will be taken up in later chapters.

SUMMARY

Sampling

1. Sampling is the basis of most statistical studies.
2. The value of statistical studies tends to be limited by the size and representativeness of the sample.
3. Samples tend to improve in the ratio of the square root of their size.
4. A group of samples from the same population tends to possess and exhibit the same characteristics.
5. All samples should be taken at random. They should be as free from all bias as possible.
6. A stratified sample is one which conforms to certain limited known characteristics of the population as to internal proportions, but within these proportions is taken at random.
7. A parameter is the true measure of a population.
8. A statistic is a computed measure of a sample.

Probability

1. The probability of an event is the relative frequency with which this event recurs in a prolonged sequence or series of observations.
2. The probabilities of mutually exclusive events are additive.
3. The probabilities of simultaneous and independent events are multiplicative.
4. Variate values of pure chance occurrences tend to be distributed about their mean approximately in proportion to the values of the coefficients to the nth power of a suitable binomial.
5. The internal relationships of the constants of the normal curve of error are useful in indicating and measuring the characteristics of a population that approaches a normal distribution. To the degree that the population departs from normal the dependability of the inferences declines.
6. One $\sigma_x \pm$ from the mean includes 68.27 per cent of the items of a normal distribution.
7. Two $\sigma_x \pm$ from the mean includes 95.45 per cent of the items of a normal distribution.
8. Social and economic data tend to fall in more highly skewed distributions than do data of the physical and biological sciences, and, therefore, require more refined treatment.

Error

1. Error in the statistical sense is a deviation or difference. It is not a mistake.
2. The standard error is expressed in terms of the standard deviation and is applied only to variates having an approximately normal distribution.

3. The distribution of the means of a group of samples taken from the same population falls in a normal curve.

4. The means of small samples tend to be more variable and erratic than the means of large samples.

5. There is a standard error for most statistical computations, the mean, standard deviation, coefficient of correlation, regression line, etc., and *no statistical computation is complete until its standard error is computed.* The standard error of a statistic is the means of measuring its reliability and dependability.

6. The standard error of the difference between two means is the device for determining whether the difference between two means is *so great* that the two samples could not reasonably be considered as being taken from the same population. This test is frequently called the Critical Ratio.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter XXVIII. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XI. Prentice-Hall, Inc., New York, 1948.

Fisher, R. A., *Statistical Methods for Research Workers*, Chapters II, III. Oliver & Boyd, Edinburgh, 1925.

Garrett, H. E., *Statistics in Psychology and Education*, Chapter II. Longmans, Green & Company, New York, 1926.

Holzinger, K. J., *Statistical Methods for Students in Education*, Chapters XI, XII, XIII. Ginn & Company, Boston, 1928.

Mills, F. C., *Statistical Methods, Revised*, Chapter XIII. Henry Holt and Company, New York, 1938.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, Chapter X. McGraw-Hill Book Company, New York, 1940.

REVIEW QUESTIONS

1. What is a random sample? Explain fully.
2. Define *universe, population.*
3. Explain "the inertia of large numbers." Give examples.
4. Explain "the permanence or persistence of small numbers." Give examples.
5. Explain "multiplicity of forces" as used in explaining sampling. Give examples.
6. Explain "independence of forces" as used in describing samples. Give examples.

7. Explain "equality of forces" as used in describing samples. Give examples.

8. What is "bias" in sampling? Give illustrations.

9. What is a *stratified sample?* Give examples.

10. What is a *purposive sample?* Give examples.

11. What is a *stratified purposive sample?* Give examples.

12. What is a parameter? Explain fully.

13. What is the difference between a parameter and a statistic?

14. What is the meaning of *probability* as used in statistics? Give examples.

15. How often, theoretically, could one draw all four aces from a shuffled deck of 52 cards at one draw?

16. What is the normal curve of error? Of what use is it?

17. What is the relation of the binomial expansion, $(\frac{1}{2} + \frac{1}{2})^n$, to the theory of probability?

18. If ten coins are thrown 1,024 times, what is the theoretical probability of the number of times they would fall 2 heads and 8 tails?

19. What is the meaning of error as used in statistics? Give examples.

20. What is meant by the standard error of the mean? How is it computed? Of what use is it in statistics?

21. What is the standard error of the standard deviation? How is it computed? Of what use is it?

22. What is the standard error of the coefficient of correlation? How is it computed? Of what use is it?

23. What is the standard error of the difference of two means? How is it computed? Of what use is it in statistics?

24. What is T? Of what use is it?

25. What is the relation of probable error to standard error? Explain.

EXERCISES

1. $\sigma_x = 5$, $N = 65$, $\overline{X} = 40$. Write $\overline{X} \pm \sigma_{\bar{x}}$.

2. $\sigma_x = 3.5$, $N = 50$, $\overline{X} = 20$. Compute the size of the sample required to reduce the standard error of the mean to 20 ± 0.1.

3. Compute σ_σ for Exercises 1 and 2.

4. For one sample, $\sigma_x = 4$, $N = 65$, $\overline{X} = 40$. For another sample, $\sigma_x = 4.5$, $N = 81$, $\overline{X} = 44$. Compute T.

5. $Y_c = \dfrac{Ni}{2.5066\sigma} 2.71828$ at the mean. $N = 145$, $i = 2$, $\sigma_x = 6$. Construct the normal curve of error.

Part Two

Types of Business Fluctuations

CHAPTER 14

THE ANALYSIS OF TIME SERIES

Two of the most important bases for classifying data are *spatial series* and *time series*. Spatial series are based on geographic distributions such as farms, city blocks, villages, cities, counties, states, and nations. Time is not an element in such data. These data are taken at a point or period of time or without any regard for time. Examples of such geographic or spatial series are wheat yields for the same year by farms, counties or states; petroleum production in many areas for the same week or month; shoe sales for the same day in many stores; and the like. In such series time is held constant or eliminated. Most of the analyses developed in the preceding thirteen chapters are primarily for spatial series, although many of the techniques may also be used with time series.

A time series is a sequence of values or measurements which correspond to successive points of time. Wheat production for one area, whether by farm, state or nation, by years, automobile sales for successive months, the average monthly price of gasoline, and weekly freight car loadings are time series. The time units may vary from centuries or millennia to minutes or seconds as among different series, but for any given series the time units must be of equal length to obtain accurate comparisons. The most common unit for factory production and wholesale prices is the month. Stock and bond prices are quoted daily. Agricultural production

is measured by the year, but the prices of farm products are given by days, weeks, months, and years.

The same variable may supply several time series, such as (1) daily petroleum production, (2) weekly petroleum production, (3) monthly, and (4) yearly petroleum production. A time series for a shorter time period may be changed to a series of a longer period merely by adding the totals of as many short periods as there are in a longer period, as for example, changing hourly data to daily data or monthly data to yearly data.

Reasons for Measuring Time Series

We live and conduct business only through the lapse of time, therefore, the changes which come through time are of the utmost importance to all individuals, families, business firms, and social groups. Although methods of time series analyses are applicable to all time series, they were developed for and are used primarily in the management of business.

The management of a business, even a small one, is quite a complex and difficult problem. These difficulties arise largely because so many changes are taking place simultaneously. An economic time series usually involves at least four types of variations:

1. Seasonal variation
2. Secular trend
3. Cyclical fluctuations
4. Random or accidental changes

The volume of automobile production, gasoline sales, theater attendance, or college enrollment at any point of time may include all four of these movements. When new models are brought out automobile sales increase. Gasoline sales rise in the summer, but college enrollment is larger in the winter.

An upward secular trend is frequently accompanied by a downward cyclical movement as Fig. 88 indicates occurred in the automobile industry in the years of 1918, 1921, 1927, 1930 to 1934, and 1938. Although the long time growth of automobile production was still upward, the output of these specific years was down because of general depression in business.

In like manner the months of the year which ordinarily are high in business activity, such as the Thanksgiving and Christmas

seasons, may be dull or lower than usual in some years because of severe general business depression or cyclical variation. On the other hand, a downswing in the cycle in some particular industry or area may show little decline in business volume because of the steep rise of secular trend. Although the volume of business generally was lower in 1949 than in 1947 and 1948, television set production increased over 300%. During the deep general depression of 1920 and 1921, the rapidly expanding Los Angeles area suffered little or no setback. It is these varying and complex interrelations among secular, seasonal, and cyclical movements which make necessary their careful measurement and accurate analysis for successful business management. The accidental and abnormal effects of drought, floods, pestilence, or war may disrupt normal business activity at any time.

Seasonal Variation

The two principal factors causing seasonal changes are the climate or weather and customs. Since the growth of all vegetation depends upon temperature and moisture, agricultural activity is confined largely to warm weather in the temperate zones and to the rainy or post-rainy season in the torrid zone. Winter and the dry season make farming a highly seasonal business. This high irregularity of month to month agricultural production determines largely the activities of all harvesting, marketing, canning, preserving, storing, financing, and pricing of farm products. Manufacturers, bankers, and merchants who supply farmers find their business taking on the same seasonal variation which characterizes the agriculture of their area, as shown in Fig. 80.

The canning industry is most active during the harvesting season of the product which is preserved. Milk and butter production rise with the growth of green pastures. Ice cream is in greater demand in the summer; hot chocolate, heavy clothing, coal, and fuel oil, in the winter. Seasonal variations in economic activity due to climate are among the most nearly universal and well known of all business activities, but their accurate measurement is essentail to profitable business operations.

The second cause of seasonal variation is custom, education, or tradition. Such traditional days as New Year's, Easter, Fourth of July, Thanksgiving, and Christmas produce marked variations in business activity, travel, sales, gifts, freight car loadings, finance, accidents, and vacationing.

The Easter and Christmas seasons, especially, lead to large expansions of sales which must be anticipated and prepared for months in advance. The sales of Christmas trees and lights alone run into tens of millions of dollars. Some types of business have from one-fourth to one-half of their annual sales in December alone.

The successful operation of any business requires that its seasonal variations be known, measured, and exploited fully. Frequently the purchase of seasonal items is made from six months to a year in advance. Departments with opposite seasonal changes are frequently combined in the same firm to avoid dull seasons and to keep sales or production up during the entire year. The

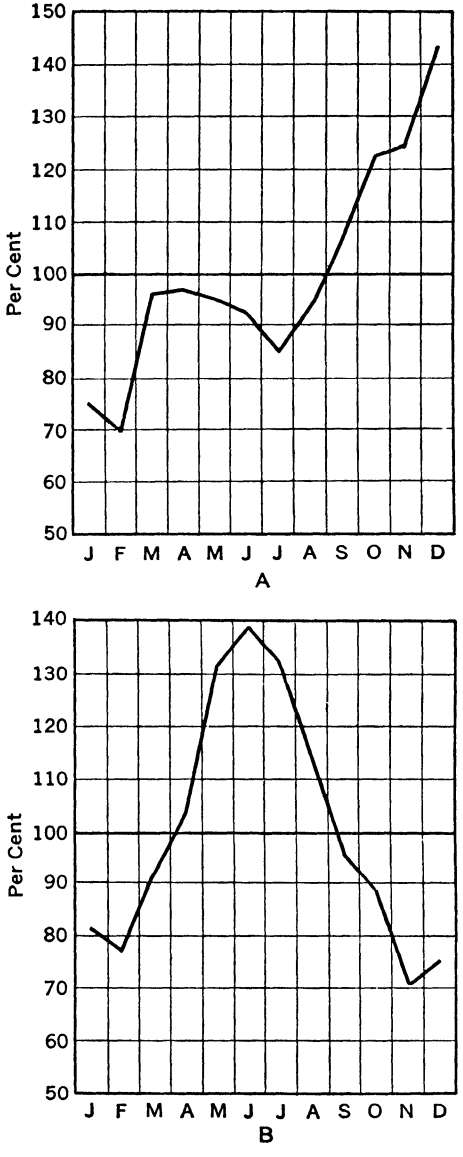


FIG. 80. (A) Seasonal sales of mail order houses. (B) Production of creamery butter. (*Survey of Current Business*)

advertising of seasonal goods just coming into demand and the storage of left over inventory for next year, or the pushing of bargain sales, "after Easter" and "after Christmas" sales form a large part of store management.

Even the proverbial abundance of "June brides" may be associated with the poetic line, "In the *spring* a young man's fancy lightly (or seriously) turns to thoughts of love." So universally are our daily lives affected by seasonal changes that we are all clearly conscious of this variation and of how to make adjustments to it.

Secular Trend

Growth and decay are universal manifestations of life and of the activities of firms, industries, and economic systems. The rates of growth and decay vary widely from one product, firm or industry to another and also during the various periods of its life. Fig. 81 showing the number of television sets produced from 1946 to 1950 inclusive, reveals how very rapid the growth of a new industry may be.

WORKSHEET NO. 67

U.S. ANNUAL AUTOMOBILE PASSENGER CAR FACTORY SALES 1900–1950 INCLUSIVE, (IN 1000) *

Year	Sales	Year	Sales	Year	Sales	Year	Sales
1900	5	1913	462	1926	3,784	1939	2,867
1901	5	1914	544	1927	2,937	1940	3,717
1902	9	1915	896	1928	3,815	1941	3,780
1903	10	1916	1,526	1929	4,587	1942	223
1904	23	1917	1,746	1930	2,785	1943	(1)
1905	24	1918	943	1931	1,973	1944	1
1906	29	1919	1,658	1932	1,135	1945	70
1907	35	1920	1,906	1933	1,574	1946	2,149
1908	58	1921	1,468	1934	2,178	1947	3,558
1909	115	1922	2,274	1935	3,252	1948	3,909
1910	170	1923	3,625	1936	3,670	1949	5,114
1911	168	1924	3,186	1937	3,916	1950	6,635
1912	313	1925	3,735	1938	2,001		

* Source: *Statistical Abstract of U.S.*, 1951.

The expansion of total air express and freight ton-miles carried, 1937–1948 inclusive, is a slightly older, but similar example of a new industry's expansion.

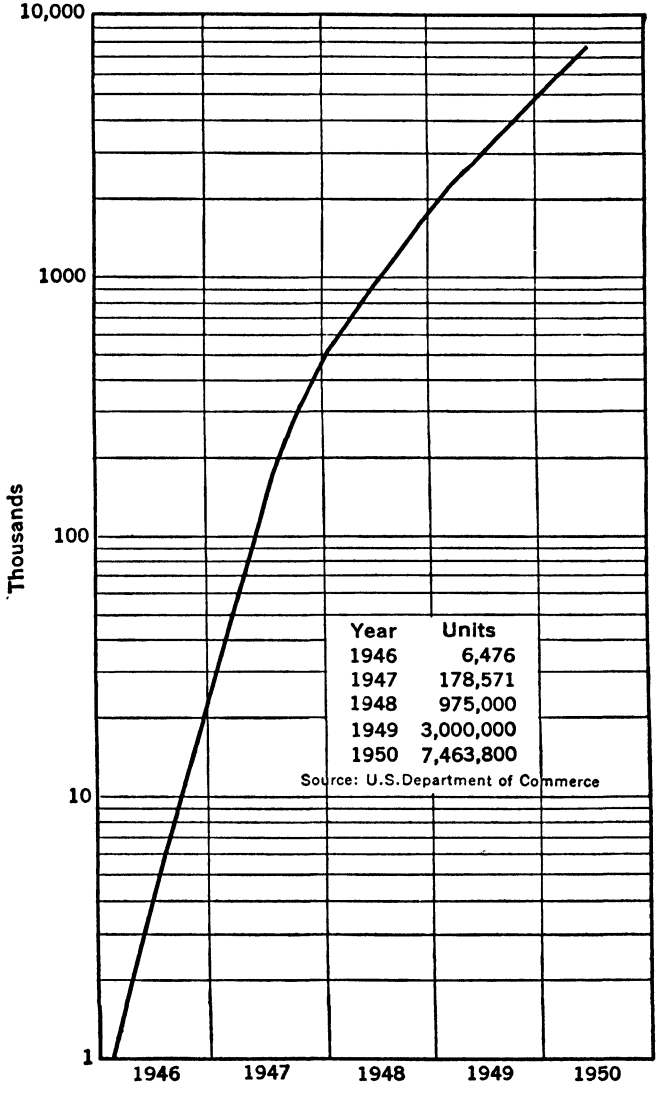


FIG. 81. Television receivers manufactured in United States, 1946–1950

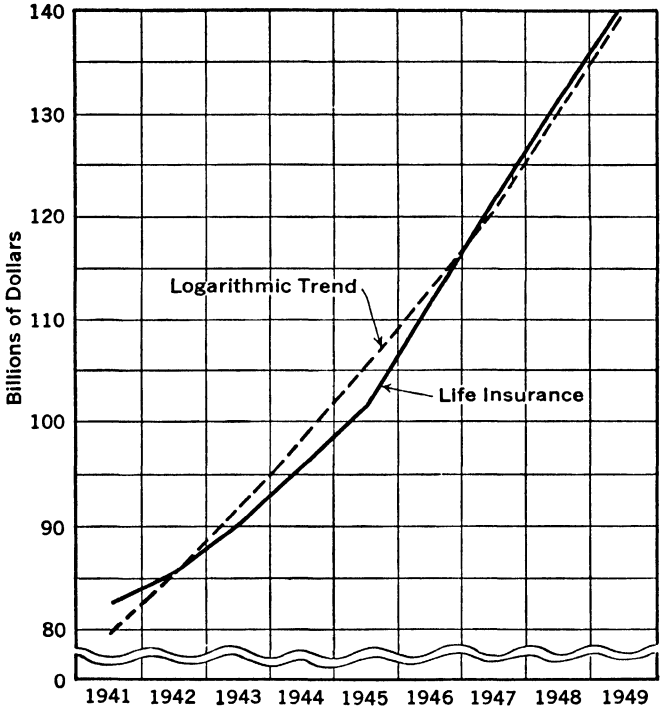


FIG. 82. Logarithmic trend fitted to ordinary life insurance in U.S., 1941–1949. (*Insurance Fact Book*, 1950, p. 6)

Automobile production, 1900–1950 is an example of the growth to maturity of a large industry. Such a long period of growth usually falls into three well defined segments, (1) the slow beginning of the new industry (1900–1911) during which it is perfecting its product and techniques of production, sales and management and gaining public favor, (2) the period of rapid expansion (1912 to 1929) when it reaches maturity, and (3) the period of market saturation (1930 to date) when most of its sales are replacements to previous users. Before 1925 most automobile sales were to new users. After 1930 most of them were replacements.

The length of time required for an industry to pass through these three periods varies greatly from one to another. It required

thirty years for the automobile industry to reach maturity. The railroads required eighty (1830 to 1910). Radios attained it in twenty years (1920 to 1940) or less. Yo-yo tops and "baby golf" made it in less time. How long will television require? The time needed usually varies closely with the size and cost of the product. The smaller the product is (1) the less the capital required to develop it, (2) the less time required to educate the public to it, and (3) the weaker the vested competitive interests posed against it.

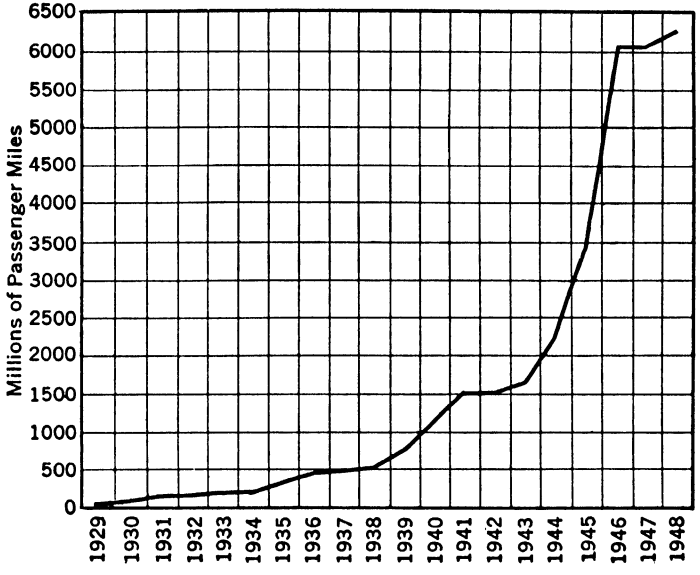


FIG. 83. Total passenger miles flown by U.S. airlines, 1929–1948. (*Moody's Industrials*, 1949, p. a.I.)

Whether the firm or industry is large or small or the time required to bring it to maturity long or short the curve of its expansion from beginning to maturity tends to be an elongated letter S, as $\int$. The Gompertz and Pearl-Reed curves presented in Chapter 10 usually give satisfactory fits for such data for the over-all period. For the initial period exponential curves are good. For short periods of from five to eight years straight lines fit quite well.

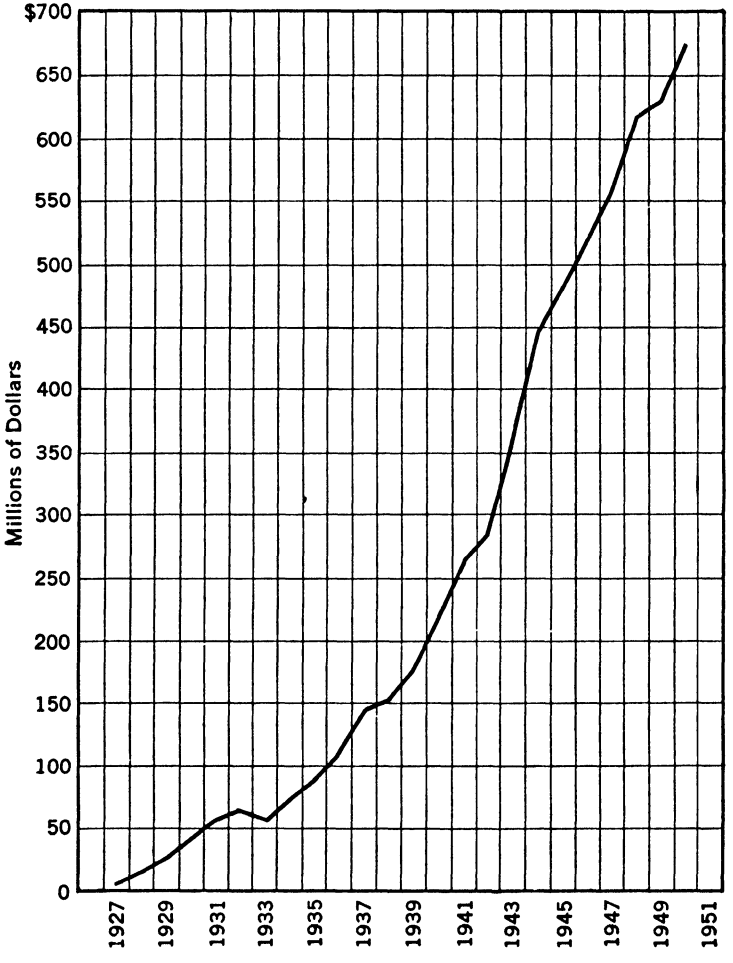


FIG. 84. Total gross billings of U.S. radio stations, 1927–1950. (*Broadcasting-Telecasting Year Book*, 1951)

WORKSHEET NO. 68

AIR PLANE PRODUCTION AND TRANSPORT IN U.S., 1933–1948

Year	Total	Production of Planes Military	Production of Planes Civilian	Passenger Miles Flown (in 1,000,000)	Express & Freight Transport in 1000 Ton Miles
1933	1,324	331	993	175	
1934	1,615	393	1,222	190	
1935	1,710	336	1,374	316	1,098
1936	3,010	858	2,152	439	1,866
1937	3,773	858	2,915	481	2,162
1938	3,623	925	2,698	516	2,182
1939	5,859	921	4,935	755	2,713
1940	12,791	6,019	6,785	1,158	3,476
1941	26,277	19,433	6,844	1,506	5,259
1942	47,836	47,836	War	1,501	11,902
1943	85,898	85,898	War	1,671	15,139
1944	96,318	96,318	War	2,212	16,991
1945	49,761	47,714	2,047	3,408	22,197
1946	36,670	1,669	35,001	6,068	38,611
1947	17,717	2,100	15,617	6,038	64,678
1948			7,302	6,230	101,376

Source: *Moody's Industrial Manual*, 1949, p. a 1.

WORKSHEET NO. 69

CHICAGO WHOLESALE PRICE OF FRESH EGGS PER MONTH, 1941–1948 INCLUSIVE (IN CENTS PER DOZEN)

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Monthly Average
1941	18.4	16.4	18.1	21.6	22.3	25.1	26.0	26.8	28.8	30.5	35.8	34.3	25.4
1942	32.9	28.9	28.3	29.3	29.9	30.4	31.6	33.7	35.1	39.0	39.0	39.0	33.1
1943	38.4	35.5	37.4	37.2	37.9	38.6	38.2	39.9	41.7	42.4	42.8	40.0	39.2
1944	35.0	33.4	32.1	31.1	30.8	33.2	34.8	33.8	36.8	38.9	42.3	41.8	35.3
1945	38.0	34.9	34.3	34.3	34.3	35.1	35.6	37.8	34.6	40.1	43.7	42.9	37.2
1946	35.6	33.1	33.2	33.3	33.6	33.2	34.0	34.6	40.6	42.0	40.6	38.8	36.1
1947	38.8	37.8	41.8	42.5	40.9	41.4	43.4	42.2	45.0	46.4	45.5	51.7	43.3
1948	44.1	43.4	43.2	42.9	41.0	41.6	41.2	44.4	44.2	45.6	49.4	48.2	44.1

Source: *Survey of Current Business* U.S. Dept. of Commerce.

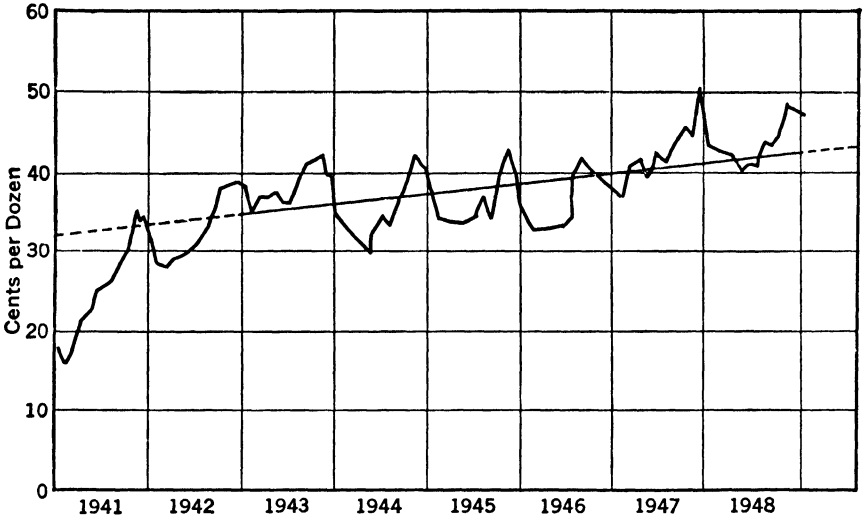


FIG. 85. Wholesale price of fresh eggs in Chicago, 1941–1948. Cents per dozen. (*Survey of Current Business*. Data in Worksheet No. 69)

WORKSHEET NO. 70

NUMBER OF HORSES, MULES, COMMERCIAL BANKS, POST OFFICES AND PRODUCTION OF ANTHRACITE COAL SINCE 1870 BY DECADES

	Horses (1000)	Mules (1000)	Commercial Banks	Post Offices	Anthracite Coal (1000 tons)
1870	8,054	1,305	1,937	28,492	
1880	10,357	1,813	3,355	42,989	
1890	15,266	2,252	8,201	62,401	41,011
1900	18,267	3,265	10,382	76,688	50,710
1910	19,220	4,101	23,095	59,580	73,623
1920	20,091	5,651	30,139	52,641	76,844
1930	13,742	5,382	24,079	49,063	46,176
1940	10,444	4,034	15,017	44,000	51,425
1950	5,775	2,250	14,121	41,000	44,000

Source: *Statistical Abstract of U.S.*

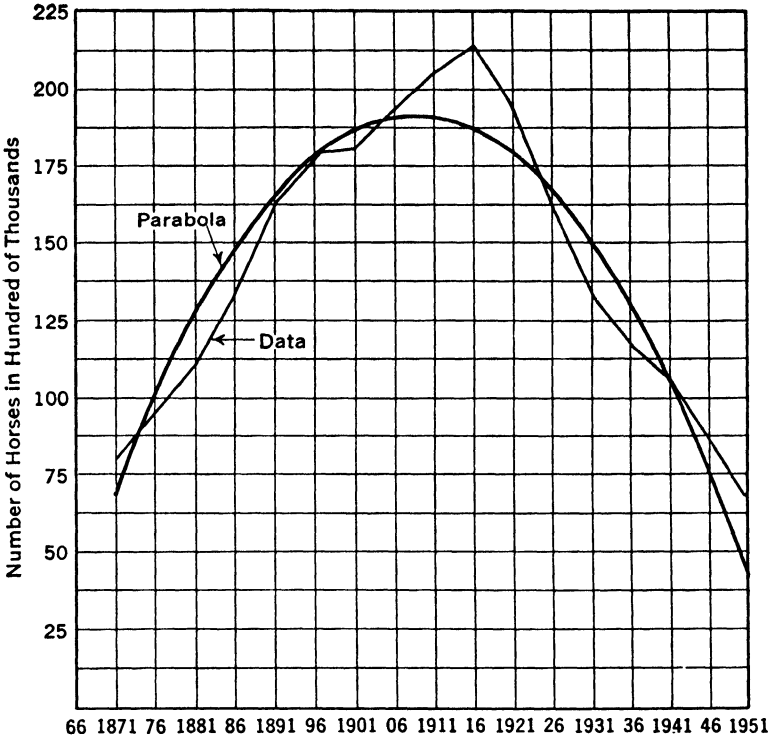


FIG. 86. Horse population of U.S., 1866–1951, and parabolic trend. (*Statistical Abstract of U.S.*, 1951)

Trends may be downward as well as upward. Illustrations of such series are the production of farm wagons, the number of horses and mules on farms, the production of buggies and other horse drawn carriages, the acreage of wheat in Iowa or Wisconsin since 1900, the acreage of cotton since 1933, the production of anthracite coal since 1910, the decline of population in decadent mining communities, and other series which the student may readily call to mind.

Such trends may be caused by soil erosion, mineral depletion, changes in transportation, substitution of new products for old, new methods of production, Congressional or legislative enactments, and many other factors. In an expanding dynamic society

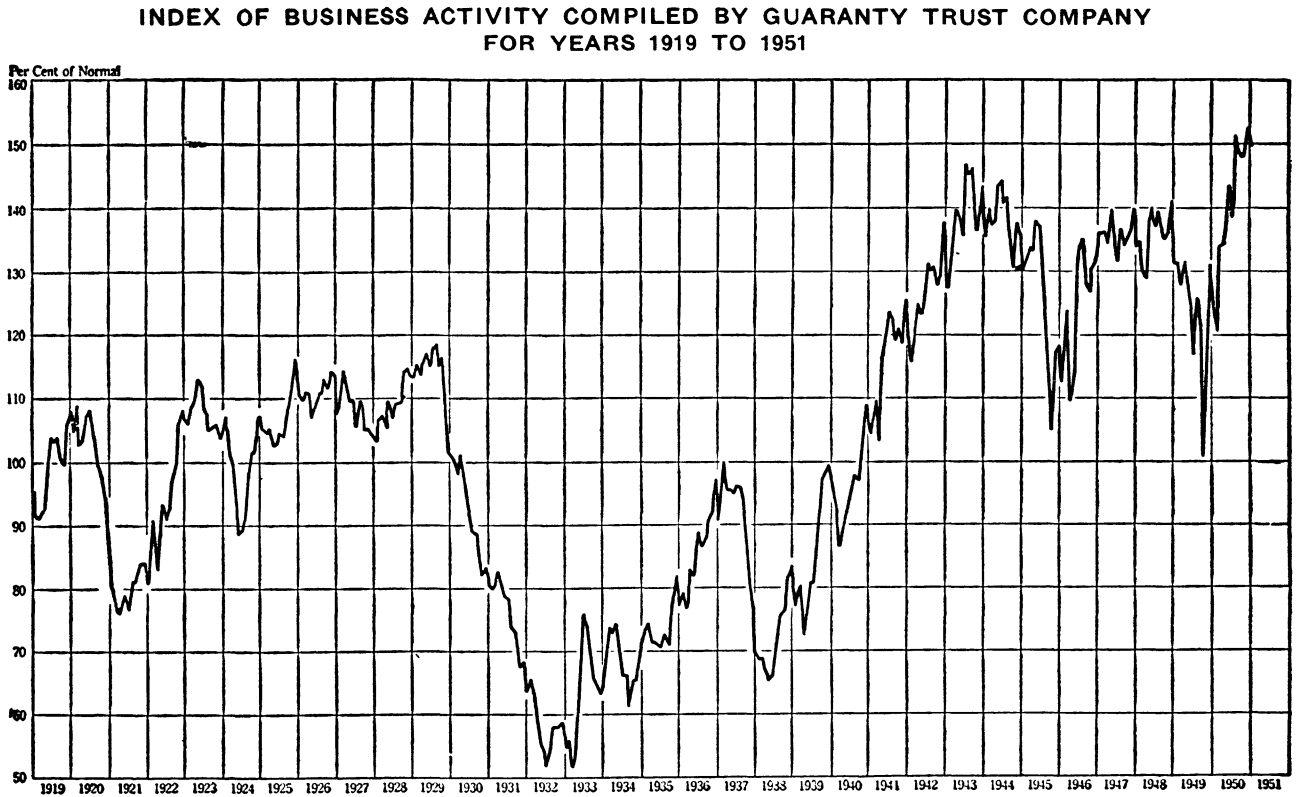


FIG. 87. Cyclical index of business activity, 1919–1951. (*The Guaranty Survey*, Guaranty Trust Co. of New York, Vol. XXX, No. II, p. 15. Used with permission)

such successions of waves of growth and decay of industrial activity are normal and inevitable.

The complete measurement of trend for some industries, firms or products from their early developments through their maximum expansion to their ultimate decline frequently takes a simple parabola or a slightly modified form of such a curve as shown in Worksheet No. 48 and Fig. 70, Chapter 10.

The decline in the number of horses and mules as well as wagons and carriages is due to the mechanization of agriculture and city transportation. They have been replaced by automobiles, trucks, busses and tractors. The decline in the number of banks is the result of establishing too many small banks prior to 1920 and their subsequent failure or consolidation.

A new invention or discovery places a new product on the market. At first few persons accept it because it is untried or expensive and has to compete with older better known products. Gradually through improvements its superiority may be demonstrated and increased efficiency may reduce its cost. Advertising and the testimony of satisfied users increase its popularity. Finally it gains wide acceptance and large sales which it holds until some newer and still better product replaces it. This succession of events takes place in the life of every product and tends to produce the successive trend movements of growth and decay manifest in the series listed above.

These forces which determine growth and decay affect individual firms as well as industries and economic systems. More than three hundred firms have been organized and produced automobiles for a few or even several years but of this large number less than thirty survive today. The origin, rise and decline of firms in every industry are a matter of general knowledge. An accurate knowledge of the trend of every firm is a necessary factor in its successful management.

Public utilities and industrial firms of large size such as producers of steel, chemicals, farm machinery, railway equipment and others have found that successful management requires the estimation of their capital expansion for from ten to twenty years in the future. Electric power and telephone companies have frequently

perfected plans based on known trends extrapolated for ten or more years to indicate building programs. By such planning they can take advantage of the best points in the business cycle to reduce costs of construction.

Trends may be measured in terms of absolute numbers or in ratios, percentages or logarithmic scales. The value of the logarithmic scale for measuring trends is that businessmen are often more interested in the *rate* at which their volume is increasing or decreasing than in the absolute amount. In comparing the progress of salesmen or the growth of branch plants or units of a chain store system, percentage changes are more significant than total figures because of the relative size of the units compared.

Business Cycles

Seasonal variation and secular trend are usually more regular in movement than cyclical fluctuations. There are, however, some cycles which over long periods have repeated themselves with almost as much regularity as does seasonal variation. The corn-hog cycle, the building cycle and the forty-two month commercial credit cycle as well as some of the short term stock market fluctuations are quite uniform.

Perhaps the reason that total cyclical fluctuations seem so erratic at times is that they are not a single movement such as seasonal variation is, but are a composite of several cycles of various lengths moving together. Since some are shorter than others they do not all reach their low or high points at the same time. Some are rising while others are falling and thus a confused picture of irregular motion results. Since statistics are concerned primarily with the measurement of cycles rather than with their full description and theory, the multiplicity of cyclical movements is introduced here only as it is related to the more accurate measurements of business cycles.

Multiplicity of Cycles

There are frequently several cycles in the same industry. The price of bread, taking both quality and size of loaf into consideration, may vary only ten to twenty-five percent during a cycle. At

the same time the price of flour may fluctuate from fifty to one hundred percent, while the price of wheat is changing from two to three hundred percent. As a general principle the nearer a good is to the stage of final consumption such as bread instead of flour or wheat, the less amplitude the cycle has, and the nearer it is to the raw material stage (wheat instead of bread) the wider is its cycle. This principle, however, is also affected by the amount of monopoly or supply control that is present in the firm or industry. These points are introduced only to indicate how difficult an accurate statistical measurement of business cycles is and how necessary it is that the statistician refine his tools and apply them with careful judgment.

The more complete measurement of cycles is a highly specialized phase of economics and statistics and is clearly beyond an elementary text in statistics. However, in Chapter 17 the two most widely accepted statistical methods of business cycle measurement are presented and applied to current time series. Figure 88 on automobile sales and Fig. 85 on egg prices indicate the apparent irregularity of cyclical movements.

The reason for the widespread efforts to measure business cycles with all its difficulties is that only by forecasting cyclical movements with sufficient accuracy can a firm or industry maximize profits or even survive in the long run. Successful controls over inventory, production, supply, prices, investments and finance all turn on the degree of accuracy with which cyclical changes are forecast. Much of the money spent by business firms on statistical services is for this purpose.

Specific Uses of Time Series Analyses

Marketing. The physical quantities of most goods and services sold each year tend to remain more nearly constant than do their prices. Total money sales fluctuate more widely than physical unit sales. This result is associated largely with changes in consumer income or purchasing power. Manufacturers and merchants, therefore, find it greatly to their advantage to measure accurately the quantity of purchasing power in the hands of consumers. This knowledge is necessary in planning production and controlling in-

ventory. If too many goods are produced or purchased they cannot be sold at present prices and loss will result. If too few are made or stocked to meet demand loss will also result through loss of sales. Most progressive firms make such estimates of the market. General Motors, General Electric, and most other large firms, make these studies with a high degree of accuracy and at large cost. Smaller firms use simpler methods but as best they can, they seek the same results. The methods used in any case consist of some or all of the techniques of time series analysis.

Advertising. Since sales are closely related to volume of advertising the decision on a sales budget largely determines within the given condition of the market, the advertising budget. The timing and volume of outlay on advertising is largely related to time series analysis.

Personnel. Since the number of factory workers, sales personnel or other employees must depend on the volume of business to be done the profitable and efficient direction of a personnel program depends largely on the accuracy with which the labor requirements of a firm are managed. These requirements vary with the time of year, the growth of the business and the swings of the business cycle. Careful time series analyses are, therefore, essential to efficient personnel management.

Investments. The creation or acquisition of capital goods all the way from the construction of a new $400,000,000 steel mill to the addition of a $1000 wash room to a garage depends on accurate estimates as to the demand, costs, and price of the output or service of the capital plant. If the life of the plant is long it is the more necessary that it be acquired at an original cost as low as possible, considering other present necessities. Careful analyses of both secular trend and cyclical fluctuations are essential to valid decisions.

In our present dynamic business world the success of most firms is attained through the continuous and careful analysis of the firm, its competitors, the industry, its markets, sources of raw materials and other supplies, personnel problems, and public relations. All these studies require the use of all or at least many of the time series techniques.

SOURCES OF TIME SERIES DATA

Every movement and activity in the world or within the knowledge of man from the range of the microscope to that of the telescope that continues through time may be considered a time series. In ordinary statistical analysis time series may be considered as limited primarily to business, social, and political activities. Our illustrations will be confined to business or economic data. The sources of such information are numerous and varied. The more prominent and useful ones are listed and described briefly in the following pages.

The Daily Newspaper

The daily paper is a large and useful source of time series data. The larger city dailies contain reports on the prices and sales of bonds and stocks of many corporations and governmental units, interest rates, bank debits, commodity prices of farm products, metals, minerals, hides, and textiles, besides exports, imports, ship loadings, car loadings, air transport, and commodity production. Every successful businessman reads carefully two sources of business data, a daily newspaper, and his own trade journal. Both are full of time series, many of which are closely related to his economic interests.

Business Records

The largest single source of time-series data is the records of business firms themselves. Accounting, sales, financial, production, transportation, personnel, cost, and nearly all other business records supply time series. Business firms make analyses of all of these records as devices of management. Budgeting requires careful study of past records for future planning. Many pricing procedures require careful study of cost conditions over periods of time. The production records of the past supply this information. Marketing surveys to measure demand are based largely on sales data of the past and on current income data as compared with previous years. A company's own records and filing cabinets contain much of this necessary information.

Government Statistics

The largest single producer of statistical information is the United States Government. Its services cover every field of social and business activity. Among its publications are:

The Federal Reserve Bulletin, issued monthly by the Board of Governors of the Federal Reserve Board and sold at a price of $3.00 a year, covers all banking and financial subjects, besides a large section of index numbers on industrial production, employment, and general business conditions. Most banks are subscribers, besides public libraries, and many business offices. It is an important source of data for all students of business.

The Survey of Current Business is issued monthly with a four-page weekly supplement by the United States Bureau of Foreign and Domestic Commerce at a cost of $3.00 a year. It covers the entire field of production and trade with monthly and annual data. It is the largest and most complete source of information in this field.

Crops and Markets, issued monthly by the United States Department of Agriculture at a price of $1.00 a year, covers the entire field of current agricultural production and prices.

The Monthly Labor Review is issued monthly by the United States Department of Labor and covers wage rates, employment, and all other labor statistics besides important index numbers.

Besides these and many other minor monthly and occasional publications which may be had for a nominal price to cover costs of publication, there are the large annual reports of the several departments of the federal government. These include *The Mineral Yearbook, Agricultural Yearbook, Statistical Abstract of the United States,* The Reports of the Treasury Department, The Interior Department, and The Reports on Domestic and Foreign Commerce, The Market Data Handbook, and The Biennial Census of Manufactures.

The Census Reports

Besides the great Decennial Census which covers every phase of population, agriculture, family, trade, commerce, and wealth

from New York and Chicago down to the smallest village, township, and farm in the nation, there is the Five-Year Census of Agriculture, The Biennial Census of Manufactures, The Census of Vital Statistics, and the Ten-Year Census of Religious Bodies. Besides these there are thousands of special bulletins and statistical analyses in all fields of economic activities. An important part of any business education is an accurate working knowledge of the sources of statistical information supplied by our federal government. It is truly a servant of the farmer and businessman.

Private Statistical Services

There are dozens of private statistical services, some of them of large magnitude, and only a few of which can be listed here. The student may soon become acquainted with others in his special field of interest by perusing the shelves of the public library or the trade journals of the business office. Listed alphabetically below are a few of the larger ones:

Babson's Reports, which include several types of services and publications for the investor and industrialist, among which is the *Babsonchart*, covering prices, production, the securities markets, and business forecasting.

Business Week (consolidated with the *Annalist*) is published by McGraw-Hill Publishing Company at a cost of $5.00 a year. It includes much data besides many analyses and forecasts of current business.

The Commercial and Financial Chronicle, published by William B. Dana Company at a cost of $15.00 a year, covers all phases of current business and corporate reports and statistics. It is one of the best of the older business publications.

Dun and Bradstreet, Inc., publish *Dun's Review* and *Dun's Statistical Review* at a cost of $5.00 a year. The Dun and Bradstreet services listing the credit standing or rating of all companies and firms of any consequence in the United States are of invaluable service to businessmen. Their index numbers and data on commodity prices are widely used.

The Moody's Investor Service publishes a number of services on

security ratings and prices and corporation profit and loss statements, balance sheets, and statistical summaries. Their services are very extensive covering railroads, public utilities, banks, and industrial companies of all kinds. The Moody publications contain a vast amount of time series data.

The Standard and Poor's Corporation publishes a large number of services including information on security prices of all types and a large number of price and production index numbers besides large tabulations of business data. Extensive and frequent analyses of the major fields of business are a large and valuable part of their publications. Almost all important time series in the fields of finance, securities, banking, prices, and production are found in their services.

The United Business Service covers in summary from all the principal forecasting services of the United States.

Besides these services listed above, many of the large banks including the Chase National Bank and The National City Bank, both of New York, and the Cleveland Trust Company, of Cleveland, Ohio, and all twelve of the separate Federal Reserve Banks publish monthly analyses of business conditions and supply a large volume of business data.

All of the hundreds of trade association publications and trade journals publish large volumes of data in the special fields covered by their respective interests.

In preparing many of these reports and services, statisticians employ the theory and methods outlined in the next four chapters of this text. Some of the data appear in their crude and unanalyzed form. Other portions are corrected for seasonal variation. Appropriate trend lines are fitted to some of them. Many of the indices appear as cycle deviations. Hundreds of complex index numbers appear among their offerings. Although many college students may never be actively engaged in creating or analyzing business statistics, all of them will be *consumers of statistics* all their lives. They will have to read and interpret business data and social data of a wide variety. Every newspaper, magazine, trade journal, and radio commercial today is full of data on business, social, and political conditions requiring some knowledge of

statistics in order for the reader or hearer to understand them. We live in a world of figures, numbers, balance sheets, profit-and-loss statements — a world of statistics and statistical analysis — and a study of time series analysis is in many ways as useful to the consumer of statistical information as it is to the analytical scientist. In the modern world we literally live by statistics.

SUMMARY

1. A spatial series is a group of measurements for a number of geographic areas for the same point in time or for the same period of time.

2. A time series is a sequence of values which correspond to successive points or periods of time. A time series requires time sequence.

3. A time series may be set up for time intervals of any prescribed length, such as for seconds, minutes, hours, days, weeks, months, years, decades or centuries, or even longer periods. Monthly and yearly measures are most common.

4. A time series is usually quite complex. Frequently several types of variation are occurring in the series at the same time.

5. The four most frequently measured and most important time variations are (1) seasonal variation, (2) secular trend, (3) cyclical fluctuations, and (4) random or accidental change.

6. Seasonal variation is change in the time series due to the changing seasons of the year. In most portions of the earth these changes are quite evident and far-reaching in their effect on man's activities.

7. Secular trend is variation in a time series due to long time growth. It measures the changes due to evolution, expansion, and decay.

8. Cyclical fluctuations are those more or less regular periods of expansion and contraction which succeed each other at intervals longer than seasonal change but shorter than secular change.

9. Time series analyses may be employed in any field of activity extending through time, but they have been more highly developed in the field of finance and business activity than elsewhere.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter XXI. Richard D. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XIV. Prentice-Hall, Inc., New York, 1948.

Mills, F. C., *Statistical Methods, Revised*, Chapter VII. Henry Holt & Company, New York, 1938.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics*, Chapter XIII. McGraw-Hill Book Company, Inc., New York, 1951.

Waugh, A. E., *Elements of Statistical Method*, Chapter VIII. McGraw-Hill Book Company, Inc., New York, 1938.

REVIEW QUESTIONS

1. Define a time series.
2. Explain in detail and give examples of the differences between a spatial series and a time series.
3. Why are time series difficult to analyze?
4. What four variations are usually found in a time series?
5. Why is a seasonal index essential to the management of a business?
6. What is the usual shape of the growth trend of an industry?
7. Is it possible that an individual basic unit or corporation might also have a similar growth trend? Explain.
8. What is the difference between an arithmetic scale and a logarithmic scale? Illustrate.
9. What are the chief characteristics of a business cycle?
10. What are erratic or accidental movements in business and how are they distinguished from cyclical movements?
11. Name and explain the content of four different United States Censuses.
12. Name and describe the nature of four private statistical services.
13. Name ten series of data found in the Federal Reserve Bulletin.
14. Name ten series of data found in the Survey of Current Business.
15. Of what advantage is it to the average businessman to understand the analysis of time series? Explain.

CHAPTER 15

SECULAR TREND

As explained in Chapter 8, the average long-time movement of a time series is known as secular trend. As distinguished from seasonal variation, which is a month-to-month change within the year, secular trend is the average long-time increase or decrease of a series caused by growth or decay. As distinguished from cycle, which is the fluctuation of a series over a few years caused by prosperity or depression, secular trend is the gradual change over many years caused by the expansion or contraction of the activity as a whole. It is evident to all of us that all human activity, including business and economic institutions and social and political movements, are subject to slow long-time processes of growth and decay. The majestic ruins of Egypt and Babylon are mute reminders of the growth, power, and decay of cultures so old that they are obscured by the mists of antiquity. Edward Gibbon's *Decline and Fall of Rome* is an eloquent reminder of the melancholy ruin of the village on the Tiber which grew in five centuries to be the Mistress of the World. The rise, wealth, and decline of the Hanseatic League, of Tyre and of Venice are examples of the same force in the business world. Indeed, this world movement of growth and decay applies not only to empires, cultures, and industries, but also to business firms, houses, and products. It is the law of life from empires and corporations, from Rome and the East India Company, to grocery stores, tailor shops, and peanut stands. No statistical analysis of a business can be complete without a measurement of its trend.

At this point it would be well for the student to review Chapter 8 which gives a full general treatment of regression. Time

series trends are only a special case under general regression. Because time units are relatively uniform and are continually repeated, it is possible to simplify and abridge time series trend formulas to save from one-half to three-fourths of the time required in the general regression computations. This is why it is preferable in statistical analysis to treat time series trend in a chapter separate from the basic principles of regression. A careful review of those principles, however, will be of great aid to the student at this point.

GRAPHIC ANALYSIS

The first step in every case of computing a secular trend is to place the data on coördinate graph paper, either of the arith-

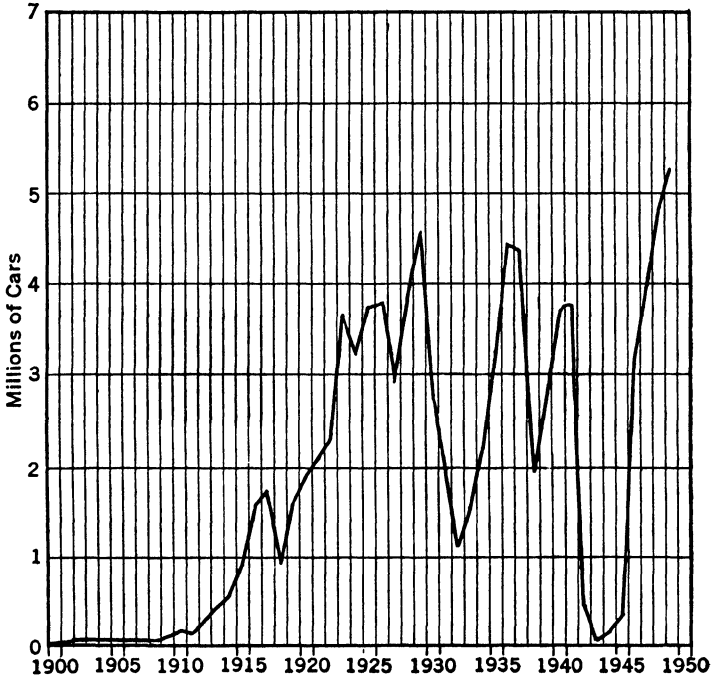


FIG. 88. Total automobile passenger car sales in U.S., 1900–1950, inclusive. Arithmetic scale

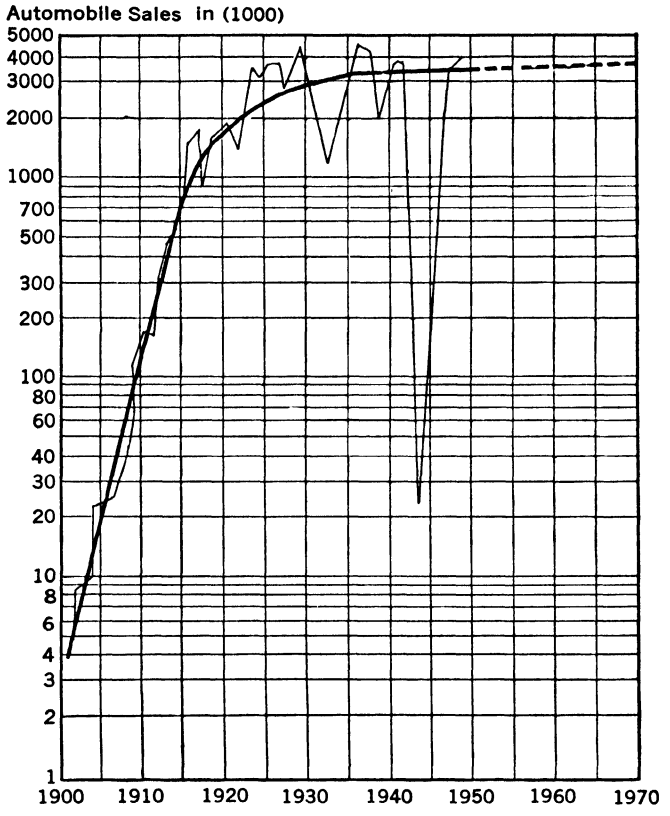


FIG. 89. Auto passenger car sales in the United States, 1910–1950. Semi-logarithmic scale

metic or semi-logarithmic type. This graph will enable the statistician to judge much more accurately and at once the correct type of trend line and formula to use. *Graph the data first.* If it is desired to measure the absolute growth of the business, the arithmetic scale is necessary as is shown in Fig. 88. If the *rate of growth,* or percentage change is desired, the semi-logarithmic scale must be used as in Fig. 89. The logarithmic graph gives the better trend. From 1900 to 1917 the growth of automobile production was at a constant rate of about 30% a year. After 1917 the rate of expansion declined to almost zero by 1941, when the industry reached its full growth with a saturated market.

In Fig. 86 the horse population of the United States is shown with a simple parabola fitted to it. The parabola fits fairly well the trend of an industry or business which has reached its maturity and has begun to decline. The number of banks in the United States from 1850 to 1940 is not accurately measured by the parabola. The period from 1890 to 1940 would fit a parabola quite well, but the earlier period is better represented by the normal growth curve.

These examples are sufficient to indicate the importance, if not the necessity, of plotting all time series data on one or both types of coordinate paper before attempting any mathematical trend computations. Freehand regression curves which were explained in detail in Chapter 8 may be used to good advantage as time series trends in many cases, but considerable experience and a high degree of skill are necessary for good results. In most cases the student should depend on mathematical lines for time series trends.

MOVING AVERAGE TRENDS

One of the simplest types of secular trend is the moving average of the data for the time periods. The method of its computation is illustrated in Worksheet No. 71. The number of time periods averaged may vary from two to any larger number, but the length of the time covered by a moving average should be uniform throughout the series. For instance, if it is decided to use a five-period moving average, the data for the first five time periods are totaled and divided by five and the resulting average is written opposite the middle year of the periods averaged. In Worksheet No. 71, the five items of data, 6, 8, 10, 9, and 7 are totaled. The sum is 40 which divided by 5 equals 8, which is written opposite the third time period. The datum, 6, for the first time period is dropped and the datum, 7, for the sixth time period is added. The total of the five items 8, 10, 9, 7, and 7 is 41, the average of which is 8.2 which is written opposite period number 4 which is the middle of these five periods. After each average the top figure in the group is dropped and the figure next below the last datum of the previous average is added to make the new total.

By this method of dropping the top figure and adding the next one below the previous group the total moves down successively through the entire number of time periods. There can never be as many averages as there are original time periods, because the average must be in the center of the group averaged. With a three-period moving average there can be no average for the first and last time periods. With a seven-period average there can be no average for the first three and the last three time periods.

The number of time periods averaged, as far as possible, should be identical with the length of the cyclical movements in the data. If this can be done, a smooth trend can be secured. In Work-

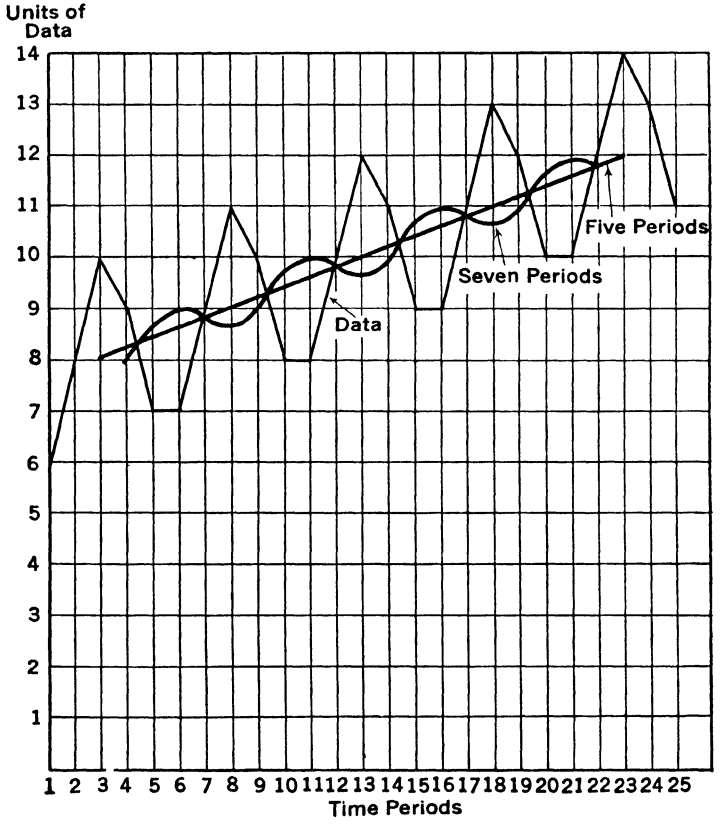


FIG. 90. Showing uniform cycles and five- and seven-period moving averages

sheet No. 71 the cycles are five time periods long and the five-period average gives a smooth line as shown in Fig. 90.

WORKSHEET NO. 71

TYPICAL MOVING AVERAGES WITH UNIFORM CYCLES

Time Periods	Original Data	2-Period Moving Average	3-Period Moving Average	5-Period Moving Average	7-Period Moving Average	10-Period Moving Average
1	6					
		7				
2	8		8			
		9				
3	10		9	8		
		9.5				
4	9		8⅔	8.2	8	
		8				
5	7		7⅔	8.4	8⅝	
		7				8.5
6	7		7⅔	8.6	9	
		8				8.7
7	9		9	8.8	8⅝	
		10				8.9
8	11		10	9.0	8¾	
		10.5				9.1
9	10		9⅔	9.2	9	
		9				9.3
10	8		8⅔	9.4	9⅛	
		8				9.5
11	8		8⅔	9.6	10	
		9				9.7
12	10		10	9.8	9⅞	
		11				9.9
13	12		11	10.0	9¾	
		11.5				10.1
14	11		10⅔	10.2	10	
		10				10.3
15	9		9⅔	10.4	10⅛	
		9				10.5
16	9		9⅔	10.6	11	
		10				10.7
17	11		11	10.8	10⅛	
		12				10.9
18	13		12	11.0	10¼	
		12.5				11.1
19	12		11⅓	11.2	11	
		11				11.3
20	10		10⅔	11.4	11⅛	
		10				11.5
21	10		10⅔	11.6	12	
		11				
22	12		12	11.8	11⅛	
		13				
23	14		13	12.0		
		13.5				
24	13		12⅓			
		12				
25	11					

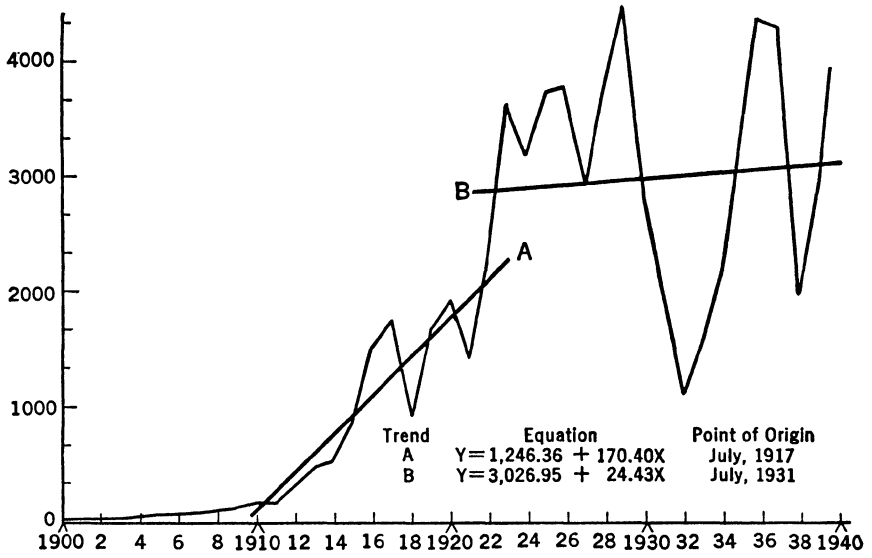


FIG. 91. Automobile production in U.S., 1900–1941, with broken trends

From the relationships revealed in Worksheet No. 71 it is evident that a moving average has two important limitations as a measure of time series trend.

1. No moving average can be as long as the period of time and series of data on which it is based. Even a 2-year moving average will always lack one year of coming up to date. A 5-year moving average will be two and one-half years short of the present time. This inherent defect of a moving average greatly limits its use as a trend measure. One of the most important services of a time series trend is its use as a means for forecasting future growth or change. To serve this purpose well, the trend should come up to the immediate present.

2. The second limitation of a moving average is that it will not give a smooth trend unless the length of the period averaged is identical with or a multiple of the business cycle. In Worksheet No. 71 the cycles are uniformly five time periods long. The student will note that only the 5-year and the 10-year moving averages give smooth trends. This is because the 5-year trend is exactly the length of the cycle and the 10-year average is an even

WORKSHEET NO. 72

MOVING AVERAGE TRENDS FOR AUTOMOBILE PRODUCTION IN U.S., 1900–1941

Years	Annual Automobile Production (000)	3-Year Moving Average	7-Year Moving Average	10-Year Moving Average
1900	5			
1901	5	6.3		
1902	9	8.0		
1903	10	14.0	15.0	
1904	23	19.0	19.3	
1905	24	25.3	26.4	31.3
1906	29	29.3	42.0	47.8
1907	35	40.7	64.9	64.1
1908	58	69.3	85.6	94.5
1909	115	114.3	111.0	139 7
1910	170	151.0	165.1	191 8
1911	168	217.0	277.6	279 0
1912	313	314.3	408.9	428 7
1913	462	443.0	591.0	599.8
1914	544	634.0	770.2	688 3
1915	896	988.7	946.2	842 6
1916	1,526	1,389.3	1,115.5	1,016.2
1917	1,746	1,405.0	1,290.9	1,143 6
1918	943	1,449.0	1,468 0	1,339 7
1919	1,658	1,502.0	1,676.4	1,656 0
1920	1,906	1,668 3	1,910.5	1,920 2
1921	1,442	1,874.0	2,212.0	2,204 1
1922	2,274	2,447.0	2,514.7	2,429.9
1923	3,625	3,028 3	2,798 0	2,548 9
1924	3,186	3,515.3	3,061.4	2,836 1
1925	3,735	3,568.3	3,333.5	3,129.0
1926	3,784	3,485 0	3,516.7	3,216.9
1927	2,936	3,511.7	3,529.1	3,270 0
1928	3,815	3,779.3	3,307.5	3,156 1
1929	4,587	3,729.0	3,020.6	2,950.9
1930	2,785	3,115 0	2,759.0	2,850 1
1931	1,973	1,964.3	2,590 8	2,801 8
1932	1,135	1,560.3	2,521.1	2,868 8
1933	1,573	1,653.3	2,561.6	3,008 5
1934	2,178	2,334.3	2,630.0	2,827.1
1935	3,252	3,291.3	2,790.8	2,661.3
1936	4,454	4,013.0	2,981.5	2,758 3
1937	4.333	3.596 0	3,245.3	2,942.6
1938	2,001	3,087.7		
1939	2,929	2,895.0		
1940	3,755	3,500.0		
1941	3,816			

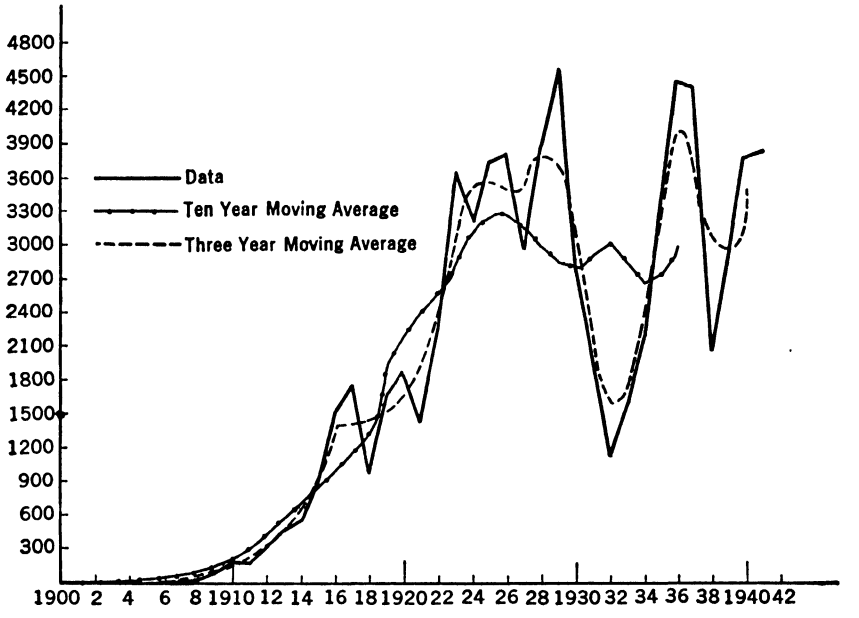


FIG. 92. Automobile production in U.S., 1900–1941, with ten-year and three-year moving averages

multiple of it. The 2-year, 3-year, and 7-year trends are so irregular that they are not usable. This limitation of a moving average is so great that it largely destroys its use as a trend measure. Business cycles are rarely of equal length. They vary from two to ten years in duration. The moving average that would fit the short cycles would make a poor trend for the long ones.

It is evident from Worksheet No. 72 and Fig. 92 that a moving average cannot be made to serve as an adequate trend for automobile production. The reason is that the cycles are unequal in length and irregular in depth. While a three- or four-year moving average might fit the data fairly well up to 1922, after that date no moving average would fit. Even a 10-year moving average is inadequate. The 10-year average is too high for the first 20 years. This is always true with data that are increasing at an increasing rate. For data with more regular cycles and a more even trend the moving average could serve quite well except for the fact that it would not extend up to the present.

LEAST SQUARES METHOD

The most satisfactory device for computing time series trends for short periods of from five to twelve years is the straight-line least squares method. For short periods of from one to three business cycles, most business trends are relatively straight. Even for longer periods a satisfactory result is often obtained by fitting a series of short straight-line trends to the data as is shown for automobile production in Fig. 91. Such trends are much easier to fit than more complex curves and often give as good or nearly as good results.

But before they are fitted the student should be convinced from a careful study of the development of the series through the entire period that there is a definite break in the long time development of the series. This point is relatively evident in the case of automobile production from 1920 to 1924. New improvements in cars and new highway construction after the First World War caused the automobile industry to reach saturation production quickly. It was an expanding industry up to 1922. After 1924 it was a mature industry. On the basis of any such definite break in trend, it is permissible to use two or more short trends instead of one long and more complex trend. The student must always be careful not to confuse definite breaks in trend because of some clearly indicated industrial long-time change with simply a large or violent cycle. If the change is purely cyclical there should be no break in the trend. Some years usually must elapse before this fact is clearly evident. For the long-time trend for a business which has reached its peak and is declining, a simple parabola is often a good fit, as is shown for the number of horses in the United States. The number of banks in the United States since 1890 is accurately measured by the same method.

The factor which makes the least squares method so easy to apply to time series is that it may be abbreviated by a short cut. If the worksheet is set up so that the time periods (X) have their point of origin (0) at their mid-point, the sum of the X's (ΣX) becomes zero. This eliminates from the normal equations all terms including ΣX, the ΣX^3, and the ΣX^5. $Na + b\Sigma X = \Sigma Y$

becomes $Na = \Sigma Y$, and a $\Sigma X + b\Sigma X^2 = \Sigma XY$ becomes $b\Sigma X^2 = \Sigma XY$. These two formulas reduce to

$$a = \frac{\Sigma Y}{N} \quad \text{and} \quad b = \frac{\Sigma XY}{\Sigma X^2}$$

for the straight line. $b\Sigma X$ and $a\Sigma X$ become zero because $\Sigma X = 0$. This greatly reduces the work.

For the parabola the normal equations become

$$b = \frac{\Sigma XY}{\Sigma X^2} \qquad \begin{aligned} Na + c\Sigma X^2 &= \Sigma Y \\ a\Sigma X^2 + c\Sigma X^4 &= \Sigma X^2 Y \end{aligned}$$

Similar short cuts apply to the cubic parabola and other longer equations. The student must remember that *these short cuts apply only* to time series. They apply only because time periods are equal and can be set up with the point of origin at the mid-point so that all odd powers of X cancel out. For all other data the basic formulas of Chapters 8 and 10 must be used.

Odd Number of Years

WORKSHEET NO. 73

SHORT CUT LEAST SQUARES LINEAR TREND OF GASOLINE IN STOCK IN U.S., 1940–1950, INCLUSIVE (IN 1,000,000 BARRELS MONTHLY MEAN FOR YEARS)

Years	x	Y	xY	x^2
1940	− 5	84.2	− 421.0	25
1941	− 4	82.0	− 328.0	16
1942	− 3	80.8	− 242.4	9
1943	− 2	69.3	− 138.6	4
1944	− 1	70.6	− 70.6	1
1945 ←———	0	77.6		0
1946	+ 1	85.4	85.4	1
1947	+ 2	84.1	168.2	4
1948	+ 3	93.5	280.5	9
1949	+ 4	106.2	424.8	16
1950	+ 5	108.7	543.5	25
	0	942.4	+ 1,502.4	110
			− 1,200.6	
			+ 301.8	

FORMULA NO. 106

$$a = \frac{\Sigma Y}{N} = \frac{942.4}{11} = 85.67$$

FORMULA NO. 107

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{301.8}{110} = 2.744$$

$$Y = a + bX$$

$$Y = 85.67 + 2.744X \quad \text{annual equation}$$

$$Y = \frac{85.67}{1} + \frac{2.744}{12}X$$

$$Y = 85.67 + .228X \quad \text{monthly equation}$$

The reason the a's in the annual and in the monthly equations have the same value (85.67) is that the original annual trend line is fitted to mean monthly data. If the original Y values had been annual totals the annual a would be divided by 12 to reduce it to a correct monthly value. Since this division was made before the trend line was computed it is not necessary to repeat it afterward.

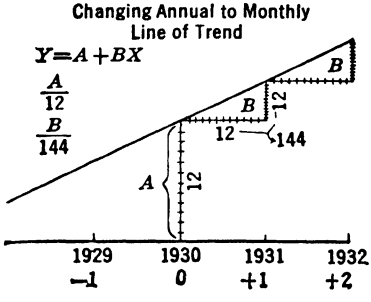


FIG. 93. Method of changing annual to monthly line of trend

The reason for dividing the b value of the annual equation (2.744) by 12 to obtain the monthly b value is that b in the annual equation measures points one year apart, from one mid-year to the next mid-year but in the monthly equation b measures the change from one month to the next month, or one-twelfth of a year. If the yearly trend line is fitted to annual totals the annual b value must be divided by 144 to reduce it to a monthly b value, as is shown in Fig. 93.

Shifting Monthly Trend to Mid-Month

As has been stressed many times in this and in earlier chapters, all measurements for time periods must be centered in the middle of the time period. The reason of this requirement is that if there is any trend in the series at all the values at the beginning and at the end of the month will not be identical. For instance, if the trend is up for later time periods, the trend value at the beginning of each month will be smaller than that of the end of the month. The first value will be too small to represent the month correctly. The last value will be too large. Only the average or middle value will be the correct measure of the month. This fact coupled with the fact that there is an even number (12) of months in a year makes the yearly trend centered at the middle of the year, between June and July, or July 1, always fall at the end of a month. When, therefore, we change the trend from years to months we must also change the monthly trend from the first to the middle of the month. In Worksheet No. 73 the annual trend is centered at July 1, 1945. When this annual trend is changed to months it must be shifted from July 1 to July 16. This is done by adding one-half of the monthly *b* to the monthly *a* as follows:

$$85.670 + .114 = 85.784$$

At July 1, 1945, where $X = 0$, $Y = 85.67$. If we substitute $\frac{1}{2}$ for X, then $Y = 85.670 + .228$ times $\frac{1}{2}$, or $85.670 + .114$. $Y = 85.784$ at July 16, which is the value written in Worksheet No. 87 opposite July 1945. Since all time values must be measured at the middle of the time period to represent accurately the time trend movement, this original shift to the middle of the month at the point of origin is necessary. Having once centered our monthly trend values at the middle of the month, as of July 16, 1945, we compute all other monthly trend values by adding or subtracting the full monthly *b* value to that of the previous or the following month. The monthly trend value for August is $Y = 85.784 + .228X$ with $X = 1$, or $Y = 85.784 + .228 \times 1 = 86.012$. For September it is:

$Y = 85.784 + .228 \times 2$ or $85.784 + .456 = 86.240$ and so for all

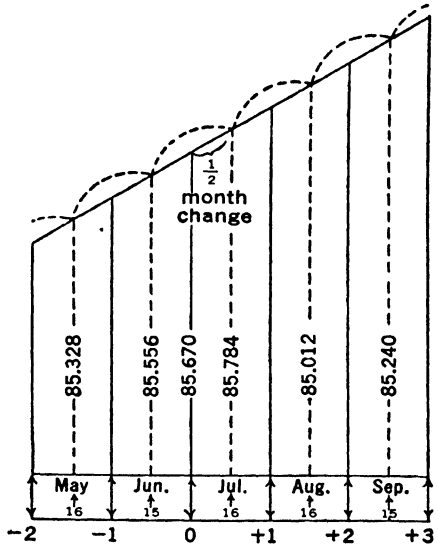


FIG. 94. Shifting a monthly trend value to middle of month

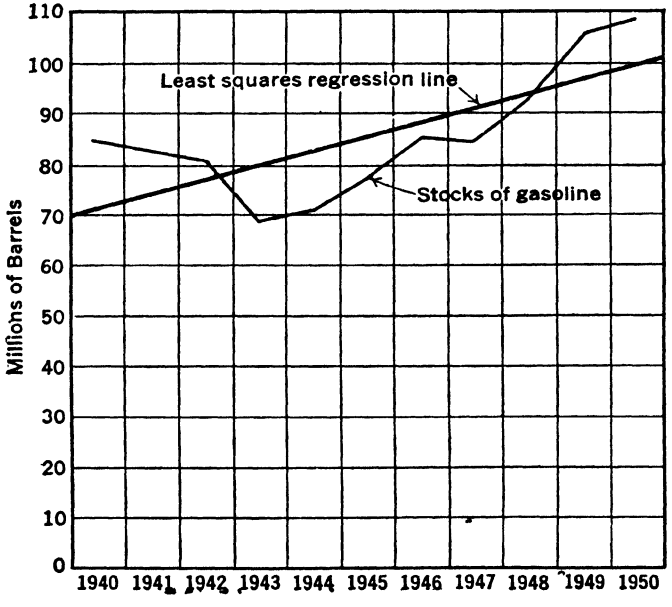


FIG. 95. Least squares regression line fitted to stocks of gasoline, 1940–1950. See Worksheet No. 73

the later months when the value of b is plus $(+ b)$. For the earlier months the b value is subtracted because the trend is declining for earlier months. If the b value is minus $(- b)$ these operations are reversed. These relationships are shown in Fig. 94.

Worksheet No. 73 contains an odd number of years (eleven), and, therefore, centered the problem on July 1, 1945, on the exact middle of the middle year. Sometimes it is impossible or undesirable to use an odd number of years. With a slight modification of the X values the computations may still be centered at the middle of the total period and the short method used.

Even Number of Years

WORKSHEET NO. 74

SHORT-CUT LEAST SQUARES LINEAR TREND OF GASOLINE IN STOCKS IN U.S., 1941–1950 INCLUSIVE (IN 1,000,000 BARRELS, MONTHLY MEAN FOR YEAR)

Years	x	Y	xY	x^2
1941	− 4.5	82.0	− 369.00	20.25
1942	− 3.5	80.8	− 282.80	12.25
1943	− 2.5	69.3	− 173.25	6.25
1944	− 1.5	70.6	− 105.90	2.25
1945	− .5	77.6	− 38.80	.25
←———	0			
1946	+ .5	85.4	+ 42.70	.25
1947	+ 1.5	84.1	+ 126.15	2.25
1948	+ 2.5	93.5	+ 233.75	6.25
1949	+ 3.5	106.2	+ 371.70	12.25
1950	+ 4.5	108.7	+ 489.15	20.25
	0	858.2	+ 1,263.45	82.50
			− 969.75	
			+ 293.70	

$$a = \frac{\Sigma Y}{N} = \frac{858.2}{10} = 85.82 \qquad Y = a + bX$$

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{293.70}{82.50} = 3.560 \qquad Y = 85.82 + 3.560X$$

This computation may be made in a slightly different form as follows:

WORKSHEET NO. 75

SHORT-CUT LEAST SQUARES LINEAR TREND OF GASOLINE IN STOCKS IN U.S., 1941–1950 INCLUSIVE (IN 1,000,000 BARRELS MONTHLY MEAN PER YEAR)

Year	x	Y	xY	x^2
1941	− 9	82.0	− 738.0	81
1942	− 7	80.8	− 565.6	49
1943	− 5	69.3	− 346.5	25
1944	− 3	70.6	− 211.8	9
1945	− 1	77.6	− 77.6	1
←———	0			
1946	+ 1	85.4	+ 85.4	1
1947	+ 3	84.1	+ 252.3	9
1948	+ 5	93.5	+ 467.5	25
1949	+ 7	106.2	+ 743.4	49
1950	+ 9	108.7	+ 978.3	81
	0	858.2	+ 2,526.9	330
			− 1,939.5	
			587.4	

$$a = \frac{\Sigma Y}{N} = \frac{858.2}{10} = 85.82$$

FORMULA NO. 108

$$b = \frac{\frac{\Sigma xY}{2}}{\frac{\Sigma x^2}{4}} = \frac{\frac{587.4}{2}}{\frac{330}{4}} = \frac{293.7}{82.5} = 3.560$$

$$Y = a + bX$$
$$= 85.82 + 3.560X$$

The advantage of this method is that it eliminates fractional values for x. In turn, however, it has the disadvantage of doubling the x, or time values which also double the xY values and quadruple the x^2's. These errors must be compensated for by dividing the xY total by 2 and the x^2 totals by 4 in solving the formula. The x values must be taken in only consecutive *odd* numbers, as

1, 3, 5, etc. The two worksheets give identical results when these necessary corrections are made.

WORKSHEET NO. 76

LINEAR LEAST SQUARES TREND FOR WHOLESALE EGG PRICES, CHICAGO, 1943–1948 INCLUSIVE (CENTS PER DOZEN, MONTHLY MEAN FOR YEAR)

Years	x	Y	xY	x^2
1943	− 2.5	39.2	− 98.00	6.25
1944	− 1.5	35.3	− 52.95	2.25
1945	− .5	37.2	− 18.60	.25
←———	0			
1946	+ .5	36.1	+ 18.05	.25
1947	+ 1.5	43.3	+ 64.95	2.25
1948	+ 2.5	44.1	+ 110.25	6.25
		235.2	+ 193.25	17.50
			− 169.55	
			23.70	

$$a = \frac{\Sigma Y}{N} = \frac{235.2}{6} = 39.20$$

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{23.70}{17.50} = 1.354$$

$$Y = a + bX$$

$Y = 39.20 + 1.354X$ annual equation
Point of origin Jan. 1, 1946

$$Y = 39.20 + \frac{1.354}{12}X$$

$Y = 39.20 + .113X$ monthly equation
Point of origin Jan. 1, 1946

$$Y = 39.20 + .113X\tfrac{1}{2} = 39.20 + .056 = 39.256$$

Jan. 16, 1946. This is the value written in the Secular Trend Column (3) in Worksheet No. 88, opposite January, 1946.

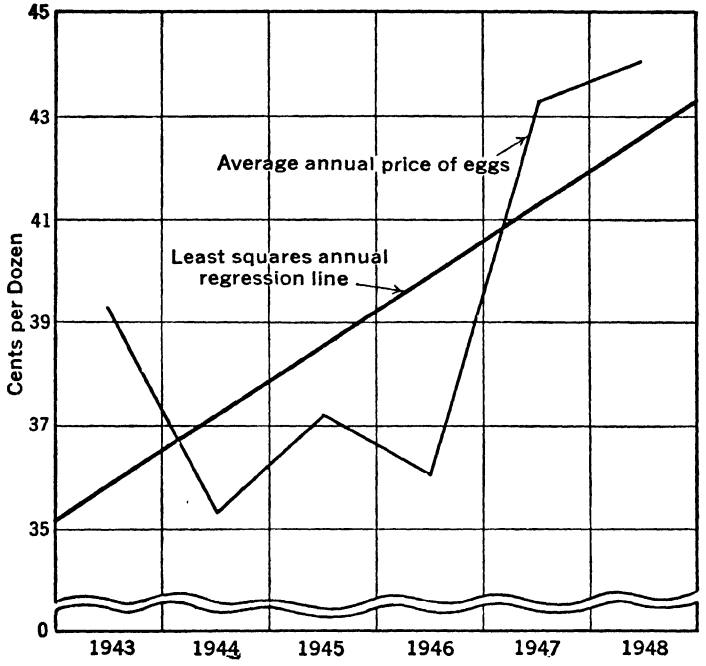


FIG. 96. Least squares regression line fitted to annual average price of eggs wholesale at Chicago, 1943–1948. See Worksheet No. 76

Ratio Trends or Log-lines

Whenever the original data plotted on a semi-logarithmic scale fall in a straight line, they are best represented by a ratio or logarithmic trend. At this point the student would do well to review the exponential, Gompertz and Pearl-Reed curves developed in Chapter 10. Computed by the short method the normal equations for time series for an exponential curve may be reduced to:

FORMULA NO. 109

$$\log b = \frac{\Sigma x (\log Y)}{\Sigma x^2}$$

FORMULA NO. 110

$$\log a = \frac{\Sigma \log Y}{N}$$

for the equation $\log Y = \log a + (\log b)X$

WORKSHEET NO. 77

LOGARITHMIC TREND FITTED TO ORDINARY LIFE INSURANCE IN FORCE IN U.S., 1941–1949 (IN BILLIONS OF DOLLARS)

Year	x	Y	log Y	x log Y	x^2	Trend in Logarithms	Trend in Natural Numbers
1941	− 4	82.6	1.91698	− 7.66792	16	1.899398	79.3
1942	− 3	85.1	1.92993	− 5.78979	9	1.930132	85.1
1943	− 2	89.8	1.95328	− 3.90656	4	1.960866	91.4
1944	− 1	95.4	1.97955	− 1.97955	1	1.991600	98.1
1945	0	101.9	2.00817			2.022334	105.3
1946	+ 1	113.4	2.05461	+ 2.05461	1	2.053068	113.0
1947	+ 2	123.0	2.08991	+ 4.17982	4	2.083802	121.3
1948	+ 3	132.1	2.12090	+ 6.36270	9	2.114536	130.2
1949	+ 4	140.5	2.14768	+ 8.59072	16	2.145270	139.7
			18.20101	+ 21.18785	60		
				− 19.34382			
				+ 1.84403			

$$\log a = \frac{18.20101}{9} = 2.022334 \qquad \log b = \frac{1.84403}{60} = .030734$$

$$\log Y = \log a + (\log b)X$$
$$= 2.022334 + .030734X$$

EXTRAPOLATION

Trend lines may be extended or projected beyond the limits of the data on which they are based. For instance, a trend line based on data from 1940 to 1950 inclusive might be projected along the same curve for 1952 or even later years. Such a practice is useful to the extent that the extended trend correctly forecasts the future movement of the series. The further the trend is extrapolated, the less likely is it a correct measure of the future development of the business. The further the trend extends into the past the more likely is it correctly to measure the future, but this is true only if the same conditions hold. The trend of growth of Chicago can be projected more accurately

than that of a new mining town. The size and nature of the business are important factors in any such extrapolation.

SUMMARY

1. Secular trend is that variation in a time series that is due to long-time expansion or contraction, growth or decay.

2. A secular trend line may be obtained by plotting the data on co-ordinate graph paper and then drawing through these plotted points a freehand line so that the area of the plotted points above the line is as nearly equal as possible to the area of the plotted points below the line.

3. A secular trend line may be obtained by dividing the data into approximately two equal parts, computing the means for each part and drawing a straight line through these plotted averages on the coordinate scale on which the data have been plotted.

4. A secular trend line may be obtained by means of a moving average. To secure a dependable trend by this method the number of time periods for which each average is taken should be as nearly as possible the length of the cyclical movements in the series. If the cycle is four years long, the moving averages should be four years long. The number of items in each average should be uniform throughout the period covered.

5. The least squares method of securing a trend line is a mathematical method which locates the trend line so that the sum of the squared deviation of the items of data from the line are a minimum.

6. A monthly least squares trend line can be secured with a minimum of time and effort by fitting the line to annual totals or averages and then shifting the line from the annual values to monthly values. Such a procedure saves about nine-tenths of the work otherwise required.

7. The standard error of estimate for time series data is not as dependable or valid as for random samples of spatial data, because its valid use is based on the relationships of the normal curve and random sampling neither of which can hold fully in a time series. The items of a time series are not free random items but are largely interdependent. The values of present periods are partially controlled and determined by the conditions of past periods.

8. Extrapolation is the projection of a trend line beyond the limits of the actual data employed in the problem. Extrapolation is not dependable in most cases when extended far beyond the data. a. The longer the period of past data, the longer the future period for which extrapolation is useful. b. The larger the area included in the measurements, the safer is the extrapolation. Extending the trend of growth for Chicago or New York would likely result in less error than extrapolating the trend of a village or new mining camp.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics,* Chapters XXI, XXII. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics,* Chapters, XV, XVI. Prentice-Hall, Inc., New York, 1939.

Davies, G. R., and Crowder, W. F., *Methods of Statistical Analysis in the Social Sciences,* Chapter VI. John Wiley & Sons, New York, 1933.

Ezekiel, N., *Methods of Correlation Analysis,* Chapters V, VI. John Wiley & Sons, New York, 1941.

Mills, F. C., *Statistical Methods, Revised,* Chapter VII. Henry Holt & Company, New York, 1938.

Richardson, C. H., *An Introduction to Statistical Analysis,* Chapter VI. Harcourt, Brace & Company, New York, 1943.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics,* Chapter XIV. McGraw-Hill Book Company, New York, 1951.

REVIEW QUESTIONS

1. Define secular trend and differentiate it from other time series fluctuations.
2. What is the difference between the arithmetic and logarithmic scales?
3. What are the limitations and weaknesses of a moving average trend?
4. Under what conditions may two or more short straight-line trends be used on a single time series?
5. What condition and device permit the shortening of the normal least squares equations for time series? Explain fully.
6. Write the original normal equations and the shortened time series normal equations for a parabola.
7. How may an annual trend line be changed into a monthly trend line?
8. Why is the annual total b divided by 144 to reduce it to a monthly b?
9. What change is made in the worksheet when an even number of years is used instead of an odd number?
10. What type of data is preferable when annual trend lines are fitted to prices?
11. How may an annual parabola be changed to a monthly parabola?
12. What should be the characteristic of data for which a Log-line is used?
13. What is meant by extrapolating a trend line? Under what conditions is it desirable?
14. Name three principal uses of trend lines and explain each use.

EXERCISES [1]

PRODUCTION OF FLUID MILK IN U.S. 1944 TO 1950 (IN 100,000,000 OF LBS.)

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual Total
1944	86	86	97	101	118	124	115	103	92	90	83	85	1180
1945	88	84	99	106	124	129	122	110	96	90	82	83	1214
1946	85	82	97	104	122	126	119	108	94	90	83	85	1195
1947	89	84	98	104	121	128	121	106	92	87	79	80	1189
1948	83	81	92	99	107	122	115	105	91	87	80	82	1184
1949	87	83	96	102	109	123	115	105	94	90	85	86	1175
1950	90	87	100	106	120	125	118	106	94	91	84	85	1206

Source: *Survey of Current Business.*

SALES OF ELECTRIC POWER TO RESIDENTIAL USERS 1944 TO 1950 (IN 100,000,000 OF KILOWATT HOURS)

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual Total
1944	29	28	27	26	25	24	24	24	25	25	27	29	313
1945	32	31	29	27	27	27	26	26	27	28	30	33	343
1946	37	35	33	31	30	30	29	29	30	31	34	37	386
1947	41	40	37	36	34	34	33	33	35	36	39	43	441
1948	48	46	44	42	39	38	38	38	40	42	45	50	510
1949	54	53	50	48	45	44	44	44	46	47	50	56	581
1950	63	60	58	55	52	51	50	50	53	55	58	66	671

Source: *Survey of Current Business.*

TOTAL RETAIL SALES IN U.S. 1944–1950 (IN $100,000,000)

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual Total
1944	49	48	56	55	58	57	55	58	60	62	62	74	694
1945	54	51	64	56	62	63	60	63	64	71	72	85	765
1946	67	64	76	79	83	80	91	91	87	94	96	108	1006
1947	83	78	93	95	100	95	98	98	103	109	107	126	1182
1948	97	89	107	107	108	109	106	107	110	115	110	132	1298
1949	93	89	105	111	108	108	102	106	110	111	109	128	1280
1950	95	93	111	111	117	120	123	127	125	121	116	145	1404

Source: *Survey of Current Business.*

[1] The teacher or student may find scores of other time series data in the *Survey of Current Business* and in the *Federal Reserve Bulletin.*

CHAPTER 16

SEASONAL VARIATION

As defined in the previous chapter, seasonal variation is the month to month change in a time series due to the time of year. Since the principal factors causing seasonal change are relatively uniform and permanent, this movement may be considered as entirely normal. Summer and winter, spring and autumn, the rainy season and the dry season, planting and reaping and their effects on customs and social activities result in about the same month-to-month change, year after year. More ice is sold in the summer, and more coal for heating homes in the winter. The price of strawberries and wheat are lowest in the month of principal harvest and are higher in months when the supply is less. Seasonal variation is the simplest form of time series change, and is the easiest to observe and measure.

Before the various methods of measuring seasonal changes are presented, some preliminary considerations for accurate measurement must be considered. The first point is the imperfections of the calendar as a device for measuring business time. The months are of unequal length. February is ordinarily only 89.3% as long as January or March. If total business for February is less than its two companion months, it by no means indicates a seasonal decline unless it falls to a figure less than 89% of the January total. The volume of business per day in this short month might be as large or larger than the daily volume of January and the monthly total still be less than that of the previous month. In certain lines of business, freight car loadings and banking, February is still further handicapped because it contains two common holidays, Washington's birthday and Lincoln's

birthday. For all businesses affected by Sundays and legal holidays, there are really only 22 business days in February as compared with 26 or 27 in both January and March. Any seasonal computation to be accurate must take into consideration differences in the length of months.

Not only are the months of different length and unequally affected by holidays, but they also vary from year to year in the number of Sundays and holidays per month. Easter is a shifting holiday of special significance for business. It sometimes falls in March but more often in April. If Easter is in March and that same year March has five Sundays, it will have but 26 business days while April that same year will have 26 business days. If the next year Easter is in April and this year April has five Sundays, it will have only 25 business days while March has 27, or a difference of 4%.

A third point to be predetermined is the units in which the seasonal index is computed. Should it be measured as a weekly or monthly change? Ordinarily the amount of seasonal change from one week to the next is so small that it is almost imperceptible. In most parts of the temperate zone the see-saw, or uneven fluctuations of the weather, would make the week an undesirable unit for seasonal measure. The average temperature of April is always higher than that of March for the same locality, but the second week in April may be much colder than the third week of March. The longer period of the month averages out this uneven fluctuation of weather to a large degree with its consequent effects on business. It is usually agreed, therefore, that the month is the correct time unit for measuring seasonal variation.

AMPLITUDE AND UNIFORMITY

By amplitude is meant the range of seasonal variation from the lowest to the highest month. In some series the amplitude is quite small, as for instance the sale of table salt. Approximately the same amount is consumed in each month of the year. If we take the average monthly consumption as 100, we find that the lowest month is only a little below 100 while the highest month is

very little above 100. In some other series, as for instance the canning of tomatoes or peas, the highest month may be three or four hundred percent of the annual average while the lowest month is zero. Most agricultural and retail merchandising series have a seasonal variation of wide amplitude. Series representing the consumption of daily necessities, such as bread, milk, tobaccos, shoes, and the like, have a narrow seasonal amplitude.

A point of large importance in the computation of seasonal variation is the regularity of the seasonal movement from year to year. After the inequalities in the length of months within one year and between years have been eliminated by proper adjustments, the question remains, Do the highs and lows of succeeding years fall at approximately the same points of time? If they do, a uniform seasonal variation is easily computed. If the low and high points shift about from year to year, more than one seasonal index may be necessary. This point is illustrated by the seasonal variation of prices in certain agricultural products, especially cotton. The seasonal price movement of a small crop after a large crop is often the opposite of that of a large crop after a large crop. Other combinations of this variation are frequent. In computing a seasonal index the student should note whether the movement is regular from year to year or whether it is highly variable.

Three Methods of Computing Seasonal Variation [1]

In this text the three following methods of computing seasonal variation are presented:

1. Monthly Medians of Original Data
2. Percentage Deviations from Least Squares Trend
3. Percentage Deviations from 12-Months'-Moving Average

[1] Two other methods of computing seasonal indexes which are not widely used are (1) The Monthly Percentages of Each Year's Data which does not give a result significantly superior to Method No. 1, above, and (2) The Link Relative Method which is long and difficult and yet inferior to Method No. 3 above. Both may be found in the older texts but are omitted from this Revised Edition because of their small and declining use.

The first method is the easiest to compute and will give quite an accurate method of the true seasonal variation if there is little trend in the series, but if the trend is steep, either up or down, the seasonal index tends to be tilted in the direction of the trend and exaggerates the values for the beginning and last months of the year. Since it is made from the original data with all their complex movements of secular trend, cyclical fluctuations, and accidental factors still in them the result is not a *pure* seasonal measure. But in cases in which the (1) trend is small and (2) with as many as six to ten years of homogeneous data available it gives a remarkably good result. It requires but a few minutes for computation.

First Method

Monthly Medians of Original Data

Steps in Computation:

1. Arrange the original data for each month in an array. This step facilitates the computation but is not absolutely necessary when only a few years are used.
2. Select the median value for each month.
3. Total the twelve monthly medians. For No. 1. Index in Worksheet No. 78 based on seven years, 1944–1950, Inclusive, of the Gasoline Stocks data the total of the 12 medians is 1,033.7.
4. Divide the total of the twelve medians by 12 to get the monthly mean of the medians which is used as the base of 100 for the index. For No. 1. Index in Worksheet No. 78 this computation is $1{,}033.7 \div 12 = 86.14$.
5. Reduce the twelve monthly medians to percentages by dividing each one by the monthly mean, as

Jan.	Feb.	Mar.
$\frac{93.3}{86.14} = 108.3$	$\frac{96.3}{86.14} = 111.7$	$\frac{97.0}{86.14} = 112.6$

and so on for all the months. The resulting quotients *are* the *Index* and should equal 1200, which in this problem they do.

WORKSHEET NO. 78

SEASONAL VARIATION INDEX, GASOLINE IN STOCKS — METHOD NO. 1
(IN 1,000,000 BARRELS END OF MONTH)

	Jan.	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*
1944	69.4	72.9	75.3	76.6	74.5	70.2	68.9	66.5	64.9	65.9	68.1	73.6
1945	78.9	85.5	85.7	79.7	77.2	74.1	74.5	74.3	65.5	68.0	78.1	89.4
1946	94.1	96.3	95.2	90.4	85.8	83.7	79.4	78.8	78.8	77.6	80.0	84.5
1947	90.3	95.0	97.0	92.7	86.7	81.2	77.1	77.2	75.9	74.7	78.7	83.1
1948	93.3	102.2	103.4	101.3	99.6	96.2	90.3	87.2	82.3	84.0	87.3	95.4
1949	108.5	117.5	118.8	117.0	113.2	106.0	103.9	97.7	94.4	96.2	97.2	103.6
1950	116.6	124.2	124.9	119.6	112.9	106.0	102.8	99.4	97.9	97.8	98.4	104.3
	Original Data Arranged in Arrays for Each Month											
	69.4	72.9	75.3	76.6	74.5	70.2	68.9	66.5	64.9	65.9	68.1	73.6
	78.9	85.5	85.7	79.7	77.2	74.1	74.5	74.3	65.5	68.0	78.1	83.1
	90.3	95.0	95.2	90.4	85.8	81.2	77.1	77.2	75.9	74.7	78.7	84.5
	93.3	96.3	97.0	92.7	86.7	83.7	79.4	78.8	78.8	77.6	80.0	89.4
	94.1	102.2	103.4	101.3	99.6	96.2	90.3	87.2	92.3	84.0	87.3	95.4
	108.5	117.5	118.8	117.0	112.9	106.0	102.8	97.7	94.4	96.2	97.2	103.6
	116.6	124.2	124.9	119.6	113.2	106.0	103.9	99.4	97.9	97.8	98.4	104.3
Medians	93.3	96.3	97.0	92.7	86.7	83.7	79.4	78.8	78.8	77.6	80.0	89.8
Index	108.3	111.7	112.6	107.6	100.6	97.2	92.2	91.5	91.5	90.1	92.9	103.4

Steps in Computation of Second Method

1. Divide the original monthly data for each month by the least squares trend value for that month and express the quotient as a percentage. The monthly percentages show the relation of data to trend and remove the influence of trend from the data, which corrects the inherent error of the First Method.
2. Arrange these percentages for each month in an array.
3. Locate the median percentage values for each month.
4. Total the twelve median percentage values and divide by 12 to secure the mean median percentage value which will be the base of 100 for the Index. (1,205.2 ÷ 12 = 100.433).
5. Divide each of the twelve median percentage values by this mean, as

$$\text{Jan.}\quad \frac{108.0}{100.433} = 107.5 \qquad \text{Feb.}\quad \frac{110.2}{100.433} = 109.7 \qquad \text{Mar.}\quad \frac{111.1}{100.433} = 110.6$$

and so on for all the twelve months. The result is the seasonal Index which should total 1,200. In this problem it does.

Second Method

Percentage Deviations From Least Squares Trend

WORKSHEET NO. 79

SEASONAL VARIATION INDEX GASOLINE IN STOCKS, BASED ON PERCENTAGE DEVIATIONS OF ORIGINAL DATA FROM LEAST SQUARES TREND (IN 1,000,000 BARRELS AT END OF MONTH)

		Jan.	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*
	Data	94.1	96.3	95.2	90.4	85.8	83.7	79.4	78.8	78.8	77.6	80.0	84.5
1946	Trend	87.152	87.380	87.608	87.836	88.064	88.292	88.520	88.748	88.976	89.204	89.432	89.660
	%	108.0	110.2	108.7	102.9	97.4	94.8	89.7	88.8	88.6	87.0	89.4	94.2
	Data	90.3	95.0	97.0	92.7	86.7	81.2	77.1	77.2	75.9	74.7	78.7	83.1
1947	Trend	89.888	90.116	90.344	90.572	90.800	91.028	91.256	91.484	91.712	91.940	92.168	92.396
	%	100.5	105.4	107.4	102.3	95.5	89.2	84.5	84.4	82.8	81.2	85.4	89.9
	Data	93.3	102.2	103.4	101.3	996.6	96.2	90.3	87.2	82.3	84.0	87.3	95.4
1948	Trend	92.624	92.852	93.080	93.308	93.536	93.764	93.992	94.220	94.448	94.676	94.904	95.132
	%	100.7	110.1	111.1	108.6	106.5	102.6	96.1	92.5	87.1	88.7	92.0	100.3
	Data	108.5	117.5	118.8	117.0	113.2	106.0	103.9	97 7	94.4	96.2	97.2	103.6
1949	Trend	95.360	95.588	95.816	96.044	96.272	96.500	96.728	96.956	97.184	97.412	97.640	97.868
	%	113.8	122.9	124.0	121.8	117.6	109.8	107.4	100.8	97.1	98.8	99.5	105.9
	Data	116.6	124.2	124.9	119.6	112.9	106 0	102.8	99.4	97.9	97.8	98.4	104.3
1950	Trend	98.096	98.324	98.552	98.780	99.008	99.236	99.464	99.692	99.920	100.148	100.376	100.604
	%	118.9	126.3	126.7	121.1	114.0	106.8	103.3	99.7	98.0	97.6	98.0	103.7
						Percentages Arranged in an Array							
		100.5	105.4	107.4	102.0	95.5	89.2	84.5	84.4	82.8	81.2	85.4	89.9
		100.7	110.1	108.7	102.9	97.4	94.8	89.7	88.8	87.1	87.0	89.4	94.2
		108.0	110.2	111.1	108.6	106.5	102.6	96.1	92.5	88.6	88 7	92.0	100.3
		113.8	122.9	124.0	121.1	114.0	106.8	103.3	99.7	97.1	97.6	98.0	103.7
		118.9	126.3	126.7	121.8	117.6	109.8	107.4	100.8	98.0	98.8	99.5	105.9
	% Median	108.0	110.2	111.1	108.6	106.5	102 6	96.1	92.5	88.6	88.7	92.0	100.3
	Index	107.5	109.7	110.6	108.1	106.0	102.2	95.7	92.1	88.2	88.3	91.6	100.0

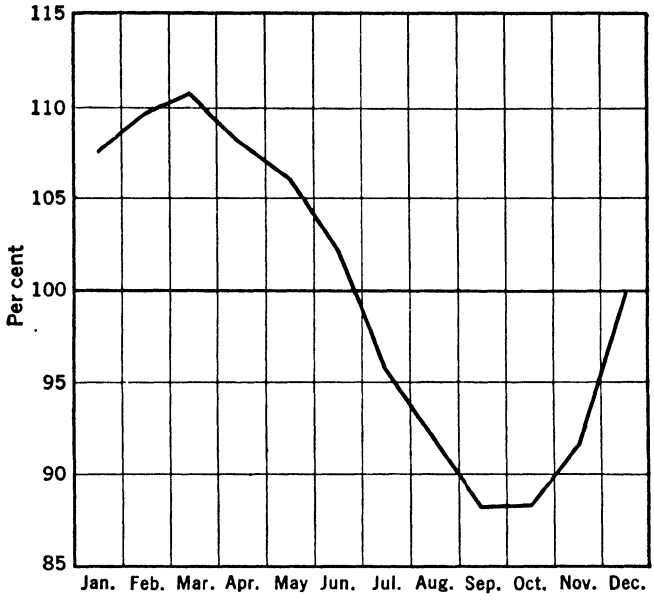


FIG. 97. Seasonal index of gasoline stocks, deviation from linear trend method. (Method No. 2)

Third Method

Moving Average Method of Computing Seasonal Index

In the first method the seasonal index was computed from the raw data with all the secular trend, cyclical fluctuations, and irregular factors in them. Needless to say such a device frequently gives a less accurate measure than is required. The second method removed the trend to a large extent but left the cyclical fluctuations and irregular movements still in the data, which sometimes skewed the seasonal index. The moving average method, however, removes all the non-seasonal elements to the highest degree and leaves a purer seasonal index than any other device. The moving average is a highly flexible device swinging up or down as the data move, whether the cause of their movement is secular trend, cyclical fluctuations, or irregular factors. It leaves a pure seasonal variation, cleaned of all extraneous movements.

A 12-months' moving average is a succession of means of a series

of twelve months' data obtained by dropping the first month of the preceding group and adding the same month of the next year. Thus the means move forward one month with each succeeding average, as in Worksheet No. 80. In the preceding worksheet, 12 items are totaled at one time and the total, 1,009.6, written at the middle of the year, between June and July, or at July 1st. The value for January, 1947, or 90.3, is then dropped from the total and the value for January, 1948, or 93.3, is added to obtain the total 1012.6, which is written between July and August, or at August 1st. These successive totals appear in column 3 of Worksheet No. 80.

WORKSHEET NO. 80

COMPUTATION OF 12-MONTHS' MOVING AVERAGE BY LONG METHOD FOR GASOLINE STOCKS AT END OF MONTH, 1947–1950, SURVEY OF CURRENT BUSINESS

(1)		(2) Original Data	(3) 12-Months' Moving Total	(4) Two 12-Months' Moving Totals Centered	(5) 12-Months' Moving Average Centered (4) ÷ 24	(6) Percentage Deviation of Data from 12-Months' Moving Average (2) ÷ (5)
1947	Jan.	90.3				
	Feb.	95.0				
	Mar.	97.0				
	Apr.	92.7				
	May	86.7				
	June	81.2				
			1,009.6			
	July	77.1		2,022.2	84.3	91.5
			1,012.6			
	Aug.	77.2		2,032.5	84.5	91.4
			1,019.9			
	Sept.	75.9		2,046.2	85.3	89.0
			1,026.3			
	Oct.	74.7		2,071.2	86.3	86.6
			1,034.9			
	Nov.	78.7		2,082.7	86.8	90.7
			1,047.8			
	Dec.	83.1		2,110.6	87.9	94.5
			1,062.8			
1948	Jan.	93.3		2,138.8	89.1	104.7
			1,076.0			
	Feb.	102.3		2,162.0	90.1	113.5
			1,086.0			
	Mar.	103.4		2,188.4	91.2	113.3
			1,102.4			
	Apr.	101.3		2,224.1	92.7	109.3
			1,121.7			
	May	99.6		2,252.0	93.8	106.2
			1,130.3			
	June	96.2		2,272.9	94.7	101.6
			1,142.6			

(1)		(2) Original Data	(3) 12-Months' Moving Total	(4) Two 12-Months' Moving Totals Centered	(5) 12-Months' Moving Average Centered (4) ÷ 24	(6) Percentage Deviation of Data from 12-Months' Moving Average (2) ÷ (5)
	July	90.3		2,300.4	95.8	94.3
			1,157.8			
	Aug.	87.2		2,330.8	97.1	89.8
			1,173.0			
	Sept.	92.3		2,361.4	98.4	93.8
			1,188.4			
	Oct.	94.0		2,392.5	99.7	94.3
			1,204 1			
	Nov.	87.3		2,421.8	100.9	86.5
			1,217.7			
	Dec.	95.4		2,445.2	101.9	93.6
			1,227.5			
1949	Jan.	108.5		2,468.6	102.9	105.4
			1,241.1			
	Feb.	117.5		2,492.7	103.9	113.1
			1,251.6			
	Mar.	118.8		2,505.3	104.4	113.8
			1,253.7			
	Apr.	117.0		2,509.6	104.6	111.8
			1,255.9			
	May	113.2		2,521.7	105.1	107.7
			1,265.8			
	June	106.0		2,539.8	105.8	100.2
			1,274.0			
	July	103.9		2,556.1	106.5	97.5
			1,282.1			
	Aug.	97.7		2,570.9	107.1	91.2
			1,288.8			
	Sept.	94.4		2,583.7	107.7	87.6
			1,294.9			
	Oct.	96.2		2,592.4	108.0	89.1
			1,297.5			
	Nov.	97.2		2,584.7	107.7	90.2
			1,297 2			
	Dec.	103.6		2,594.4	108.1	95.8
			1,297 2			
1950	Jan.	116.6		2,593.3	108.1	107.9
			1,296.1			
	Feb.	124.2		2,593.9	108.1	114.9
			1,297.8			
	Mar.	124.9		2,599.1	108.3	115.0
			1,301.3			
	Apr.	119.6		2,604.2	108.5	110.2
			1,302.9			
	May	112.9		2,607.0	108.6	104.0
			1,304.1			
	June	106.0		2,608.9	108.7	97.5
			1,304.8			
	July	102.8				
	Aug.	99.4				
	Sept.	97.9				
	Oct.	97.8				
	Nov.	98.4				
	Dec.	104.3				

The purpose of column 4 is to center these totals and, therefore, the moving average, at the middle of the months. Since the original totals in column 3 do not fall at the middle of the month but between the months, the moving averages based on them do not represent the values of the months accurately. In column 4 the successive totals of column 3 are totaled, which centers their combined value at the middle of the month. When the value, 1,009.6 at July 1 is combined with the value 1,012.6 at July 31, or Aug. 1, and averaged it centers the average at Mid-July. And so do the succeeding 24 months' totals for the successive months.

The true moving average in column 5 is obtained by dividing the 24 months' totals of column 4 by 24. *The moving averages themselves are not the seasonal index*, but are only the base values, or line, from which the original monthly data are computed as percentage deviations. The moving average value of each month is divided into the original data value for that month and the result expressed as a percentage. For July 1947 for the Gasoline Stocks the computations are: $77.1 \div 84.3 = 91.5\%$. For August the computations are: $77.2 \div 84.5 = 91.4\%$, and so on for all succeeding months in the series. The seasonal index is computed in Worksheet No. 81 from the medians of these monthly percentage values. The procedure of reducing these monthly medians to the seasonal index is the same as that employed in the two previous methods. Total the twelve medians. Divide this total by 12 to obtain their mean. Divide this mean into each one of the original medians. The result is the index.

Seasonal Indexes Corrected for Variation in Length of Months

It will be recalled that earlier in this chapter, reference was made to the effects of the inequality of the lengths of the months on many seasonal indexes. February ordinarily has only 28 days. April, June, September, and November have 30, and all the others 31. In none of the indexes computed so far in this chapter has any adjustment been made for the varying length of month. In seasonal indexes of prices length of month has little effect, but in quantity data the seasonal is invalidated from 3% to 10% by this

WORKSHEET NO. 81.

Computation of Seasonal Index of Gasoline Stocks at End of Month, 1947–1950. From Percentage Deviations from 12-Months' Moving Averages Centered

	Jan.	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*	
	104.7	113.1	113.3	109.3	104.0	97.5	91.5	89.8	87.6	86.6	86.5	93.6	
	105.4	113.5	113.8	110.2	106.2	100.2	94.3	91.2	89.0	89.1	90.2	94.5	
	107.9	114.9	115.0	111.8	107.7	101.6	97.5	91.4	93.8	94.3	90.7	95.8	
Percentage Medians	105.4	113.5	113.8	110.2	106.2	100.2	94.3	91.2	89.0	89.1	90.2	94.5	$= \frac{1,197.6}{12} = 99.8$
Index	105.6	113.7	114.0	110.4	106.4	100.4	94.5	91.4	89.2	89.3	90.4	94.7	
Index Corrected for Difference in Length of Month													
Percentage Medians	105.4	113.5	113.8	110.2	106.2	100.2	94.3	91.2	89.0	89.1	90.2	94.5	
Corrected Medians	105.4	125.6	113.8	113.8	106.2	103.5	94.3	91.2	92.0	89.1	93.2	94.5	$= \frac{1,222.6}{12} = 101.883$
Adjusted Index	103.4	123.3	111.6	111.6	104.0	101.4	92.6	89.5	90.3	88.0	91.5	92.8	

error. It is usually impossible to produce or sell as large a volume of goods in February as in January and March although the totals for each separate day for the three months were identical because January and March are 10.7% longer than February.

WORKSHEET NO. 82

COMPARISON OF FOUR INDEXES OF SEASONAL VARIATION FOR GASOLINE STOCKS

Type of Index	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
First	108.3	111.7	112.6	107.6	100.6	97.2	92.2	91.5	91.5	90.1	92.9	103.8
Second	107.5	109.7	110.6	108.1	106.0	102.2	97.7	92.1	88.2	88.3	91.6	100.0
Third	105.6	113.7	114.0	110.4	106.4	100.4	94.5	91.4	89.2	89.3	90.4	94.7
Corrected	103.4	123.3	111.6	111.6	104.0	101.4	92.6	89.5	90.3	88.0	91.5	92.8

In Worksheet No. 81 the correction is made by bringing down the median percentages of Worksheet No. 81 and then multiplying the medians of Jan., Mar., May, July, Aug., Oct., and December by 100%, or leaving them unchanged. The medians of April, June, Sept., and November are multiplied by 103.3, and that of

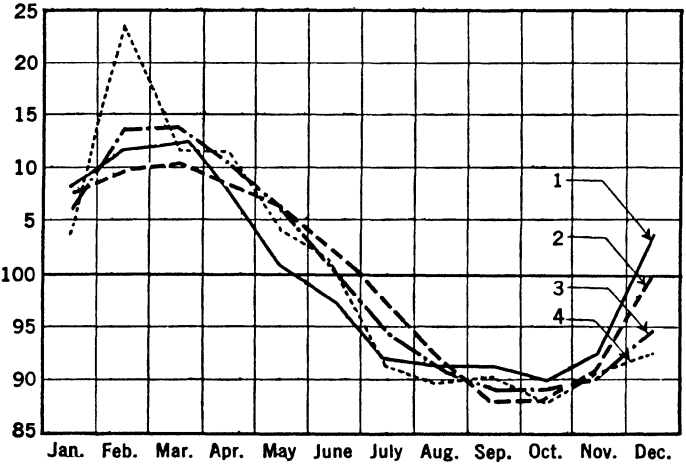


FIG. 98. Comparison of four seasonal indexes of gasoline stocks by Methods 1, 2, and 3, and adjusted index for variation in lengths of months. (From Worksheet No. 82)

February by 110.7. These medians now adjusted for variation in length of month are reduced to a seasonal index by the same method employed in the earlier indexes. Total the twelve adjusted medians. Obtain their mean. Divide each adjusted median by this mean. The result is the adjusted index.

This type of correction is only roughly accurate. Because of the variation in the number of Sunday and legal holidays in the various months a more accurate correction is necessary to reduce the seasonal index to *business* days. This more accurate correction can be made by adjusting the original monthly medians to the actual number of open business days or hours per month. Rarely would such a high degree of accuracy be of practical value in the management of a business.

Comparison of Four Indexes

WORKSHEET NO. 83

SHORT METHOD FOR COMPUTING 12-MONTHS' MOVING AVERAGE FROM A CENTERED 24-MONTHS' MOVING TOTAL

LONG METHOD				SHORT METHOD		
12-Months' Total Centered between June and July 1		12-Months' Total Centered between July and August 2		13-Months' Weighted Moving Total Centered on Mid-July 1 + 2		
Jan.	90.3			Jan.	90.3	
Feb.	95.0	Feb.	95.0	Feb.	95.0	+ 95.0
Mar.	97.0	Mar.	97.0	Mar.	97.0	+ 97.0
April	92.7	Apr.	92.7	Apr.	92.7	+ 92.7
May	86.7	May	86.7	May	86.7	+ 86.7
June	81.2	June	81.2	June	81.2	+ 81.2
July	77.1	July	77.1	July	77.1	+ 77.1
Aug.	77.2	Aug.	77.2	Aug.	77.2	+ 77.2
Sept.	75.9	Sept.	75.9	Sept.	75.9	+ 75.9
Oct.	74.7	Oct.	74.7	Oct.	74.7	+ 74.7
Nov.	78.7	Nov.	78.7	Nov.	78.7	+ 78.7
Dec.	83.1	Dec.	83.1	Dec.	83.1	+ 83.1
		Jan.	93.3	Jan.		+ 93.3
	1,009.6		1,012.6			2,022.2

Short Method of Computing a Moving Average Seasonal Index

The short-cut method eliminates column 3 used in Worksheet No. 80 by running 24 months in the moving total as is indicated in the Short Method in Worksheet No. 84. To do this 24 months are carried continuously. For the second total the first January value, 90.3 and the first February value, 95.0 are both subtracted from the first total of 2,022.2 and the January and February values of the next year, 93.3 and 102.3 respectively, are added making the second total 2,032.5. For each succeeding total the first two items of the previous total are subtracted and the values for the same months of the next year added. This process reduces the work considerably but is quite difficult for persons not highly accurate on calculating machines. The long method is suggested for beginners.

WORKSHEET NO. 84

SHORT METHOD OF COMPUTING MOVING AVERAGE FOR SEASONAL INDEX OF WHOLESALE EGG PRICES, CHICAGO, 1943–1948

1 Years and Months	2 Original Data	3 13-Months' Wtd. Moving Total Centered	4 12-Months' Moving Average Centered	5 Percentages Ratio of Data to Moving Averages
1943				
Jan.	38.4			
Feb.	35.5			
Mar.	37.4			
Apr.	37.2			
May	37.9			
June	38.6			
July	38.2	936.6	39.0	97.9
Aug.	39.9	931.1	38.8	102.8
Sept.	41.7	923.7	38.5	107.3
Oct.	42.4	912.3	38.0	111.6
Nov.	42.8	899.1	37.5	114.1
Dec.	40.0	886.6	36.9	108.4
1944				
Jan.	35.0	877.8	36.6	95.6
Feb.	33.4	867.9	36.2	92.3

WORKSHEET NO. 84 (continued)

1 Years and Months	2 Original Data	3 13-Months' Wtd. Moving Total Centered	4 12-Months' Moving Average Centered	5 Percentages Ratio of Data to Moving Averages
Mar.	32.1	856.9	35.7	89.9
Apr.	31.1	848.5	35.4	87.8
May	30.8	845.4	35.2	87.5
June	33.2	847.6	35.3	94.0
July	34.8	852.4	35.5	98.0
Aug.	33.8	856.9	35.7	94.7
Sept.	36.8	860.6	35.9	102.5
Oct.	38.9	866.0	36.1	107.7
Nov.	43.2	872.7	36.4	118.7
Dec.	41.8	878.0	36.6	114.2
1945				
Jan.	38.0	880.6	36.7	103.5
Feb.	34.9	886.4	36.9	94.6
Mar.	34.3	889.2	37.1	92.4
Apr.	34.3	888.2	37.0	92.7
May	34.3	889.9	37.1	92.5
June	35.0	891.5	37.1	94.3
July	35.6	890.2	37.1	96.0
Aug.	37.8	886.0	36.9	102.4
Sept.	34.6	883.1	36.8	94.0
Oct.	40.1	881.0	36.7	109.3
Nov.	43.7	879.3	36.6	119.4
Dec.	42.9	876.8	36.5	117.5
1946				
Jan.	35.6	873.4	36.4	97.8
Feb.	33.1	868.6	36.0	91.9
Mar.	33.2	871.4	36.3	91.5
Apr.	33.3	879.3	36.6	91.0
May	33.6	878.1	36.6	91.8
June	33.2	870.9	36.3	91.4
July	34.0	870.0	36.2	93.9
Aug.	34.6	873.8	36.4	95.0
Sept.	40.6	887.1	37.0	109.7
Oct.	42.0	904.8	37.7	111.4
Nov.	40.6	921.3	38.4	105.7
Dec.	38.8	936.8	39.0	99.5

WORKSHEET NO. 84 (continued)

1 Years and Months	2 Original Data	3 13-Months' Wtd. Moving Total Centered	4 12-Months' Moving Average Centered	5 Percentages Ratio of Data to Moving Averages
1947				
Jan.	38.8	954.4	39.8	97.5
Feb.	37.8	971.4	40.5	93.3
Mar.	41.8	983.4	41.0	102.0
Apr.	42.5	992.2	41.3	102.9
May	40.9	1001.5	41.7	98.1
June	41.4	1019.3	42.5	97.4
July	43.4	1037.5	43.2	100.5
Aug.	42.2	1048.4	43.8	96.3
Sept.	45.0	1055.4	44.0	102.3
Oct.	46.4	1057.2	44.1	105.2
Nov.	45.5	1057.7	44.1	103.2
Dec.	51.7	1058.0	44.1	117.2
1948				
Jan.	44.1	1056.0	44.0	100.2
Feb.	43.4	1056.0	44.0	98.6
Mar.	43.2	1057.4	44.1	98.0
Apr.	42.9	1055.8	44.0	97.5
May	41.0	1058.9	44.1	93.0
June	41.6	1059.3	44.2	94.1
July	41.2			
Aug.	44.4			
Sept.	44.2			
Oct.	45.6			
Nov.	49.4			
Dec.	48.2			

The seasonal index for Wholesale Egg Prices, Chicago, 1943–1948, Inclusive is computed by the short method for moving averages and Worksheet No. 84 as compared with Worksheet No. 80, indicates the advantage of this shorter computation. The percentage deviations of the original data from the moving average in column 5 of Worksheet No. 84 are transferred to Worksheet No. 85 in array order and the seasonal index computed from these medians by the usual method. Since this index is for monthly

WORKSHEET NO. 85

COMPUTATION OF SEASONAL INDEX FOR WHOLESALE EGG PRICES, CHICAGO 1943–1948, FROM PERCENTAGE DEVIATIONS FROM 12-MONTHS' CENTERED MOVING AVERAGE IN WORKSHEET NO. 84

	Jan.	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*	
	95.6	91.9	89.9	87.8	87.5	91.4	93.9	94.7	94.0	105.2	103.2	99.5	
	97.5	92.3	91.5	91.0	91.8	94.0	96.0	95.0	102.3	107.7	105.7	108.4	
	97.8	93.3	92.4	92.7	92.5	94.1	97.9	96.3	102.5	109.3	114.1	114.2	
	100.2	94.6	98.6	97.5	93.0	94.3	98.0	102.4	107.3	111.4	118.7	117.2	
	103.5	98.6	102.0	102.9	98.1	97.4	100.5	102.8	109.7	111.6	119.4	117.5	
Percentage Medians	97.8	93.3	92.4	92.7	92.5	94 1	97.9	96.3	102.5	109.3	114.1	114.2	= 1197.1 ÷ 12 = 99.758
Index	98.0	93.5	92.6	92.9	92.7	94.3	98.2	96.5	102.8	109.6	114.4	114.5	

prices, it is not adjusted for variation in length of months. Monthly prices are averages of daily quotations, or are taken as of some specific day in the month, such as the middle. Such data are not subject to the same degree of error from variation in length of months as are quantity data, such as total monthly sales or production, which are in part a function of time.

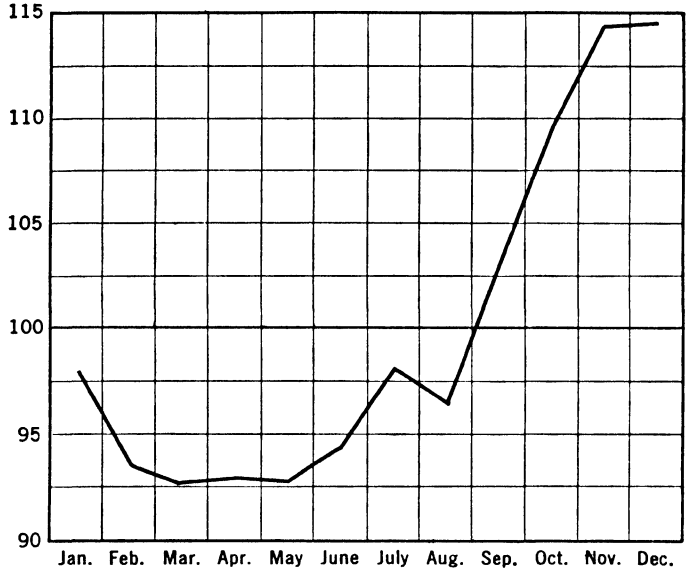


FIG. 99. Seasonal index of prices of fresh eggs in Chicago by Method 3

CHANGING SEASONAL VARIATION

Seasonal activity in business frequently changes. An index which measures present conditions accurately may not do so after a few years. The opening of new industries or the exploitation of new resources may so change the total economic activity of a community that a new seasonal distribution of business may develop. Even new methods of selling, packing, delivery, or of credit extension may either accentuate or reduce the amplitude of seasonal variation or shift the high or low points of the year to earlier or later months.

Statistical methods have been developed which aid in measuring such changes, but in the main they are more complicated and expensive than the computation of a new seasonal index. Alert businessmen are usually quite conscious of any considerable change in the flow of business through the year, and as a safeguard for accurate budgeting should recompute their seasonal indexes every two or three years, or even annually if marked changes seem to be evident. This is the most economical and effective device for accurate annual budgeting.

SUMMARY

1. Seasonal variation is usually computed as a month to month change expressed as a percentage with the average for the years as the base of 100%.

2. The simplest method of computing a seasonal index is to reduce the twelve totals for the twelve months for the original data to percentages of their mean as a base of 100%. Such a result will include a large error if the secular trend is quite steep or the cycle quite violent. Otherwise the results obtained by this simple method are dependable.

3. More exact results are obtained from each of the three methods: (1) deviations from the least squares trend, (2) link relative method, or (3) deviations from 12-months' moving average. The latter method is usually superior to the other two because it largely eliminates the influence of both secular trend and cyclical fluctuation from the seasonal index.

4. A seasonal index to be dependable should be computed from data which are as uniform as possible. If there has been an organic or permanent change in the nature of the seasonal variation in the recent past, it is still better to use a few years which truly represent the present condition than to include a longer period of past years which are no longer representative of the variation.

5. The seasonal index is easily understood and widely used by businessmen in preparing their monthly budgets, inventories, and production activities.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter XXIII. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapters XVII, XVIII. Prentice-Hall, Inc., New York, 1948.

Mills, F. C., *Statistical Methods, Revised*, Chapter VIII. Henry Holt and Company, New York, 1938.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics*, Chapter XV. McGraw-Hill Book Company, New York, 1938.

REVIEW QUESTIONS

1. What are the principal characteristics of seasonal variation?
2. What are the chief causes of seasonal variation?
3. In what terms is a seasonal index expressed? Why?
4. What are the strong and weak points of Monthly Medians of Original Data Method of computing a seasonal index? Under what circumstances should it be used?
5. What is the advantage of the Second Method over the First Method of computing seasonal variation? Is the difference large or small? Why?
6. In what respects does the Third Method, or Deviation from Least Squares Trend Method, differ from the other four methods of computing seasonal variation? In what respects is it superior or inferior?
7. What are the principal characteristics of the Link Relative Method of computing seasonal? In what respects is it undesirable? When should it be used? Does it eliminate cycle and trend?
8. Compare the Moving Average Method of computing seasonal variation with all the other methods as to (1) ease of computation, (2) accuracy, (3) reasons for accuracy.
9. Compare Worksheets No. 80 and No. 83 as to (1) ease of computation, and (2) accuracy. Under what conditions should each be used?
10. Of what practical value are seasonal indices in business management?

CHAPTER 17

CYCLICAL FLUCTUATIONS

A business cycle consists of those relatively irregular fluctuations of a business above and below its normal or long-time average activity. The higher portions of a cycle are ordinarily called prosperity and the lower portions depression. There is no such thing as *the* business cycle. There are actually hundreds or even thousands of separate cycles. Each industry has its own peculiar cycle. Each corporation or even each separate plant in every industry tends to vary somewhat in its cyclical movements from other plants and firms in the same industry because of different conditions of location, personnel, product, management, credit, and other factors. Plants or industries located in the large cities, on the seacoast, or dependent on export markets may have quite a different cycle from inland, domestic, or local firms or plants in the same field. The cycles of stock and bond prices, interest rates, money in circulation, retail prices, wholesale prices, raw material prices, finished goods prices, and other related factors may begin and end at different times, have different lengths and amplitudes, and affect the management of the business in very different ways. For purposes of practical management, therefore, the use of business cycle information requires a detailed study of the specific cycles occurring in a particular business or phase of business. What is sometimes called *the* business cycle is only a generalized or average picture of large or total business movements. It is a broad generalization based on the average of many separate cycles and as such has the limitations of all broad generalizations in that it obscures or eliminates many useful and necessary details for any specific industry or particular business firm. Local

and regional or national conditions vary so widely that a particular business may be very prosperous in California or Texas while it is depressed in Illinois or Massachusetts or vice versa. In the study and in the application of the measurements of business cycles these many and wide variations should be kept in mind.

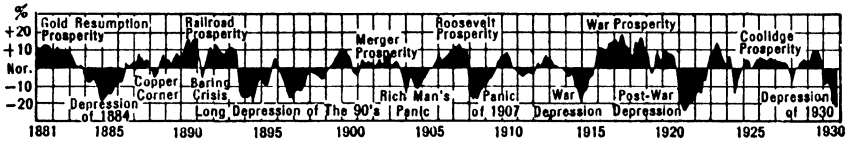


FIG. 100. Half a century of American business, showing variations in cycles. (Used with permission of the Cleveland Trust Company *Business Bulletin*)

Business cycles for single plants or industries as well as the average or generalized cycles vary greatly in duration and amplitude. The cycles of short-lived consumers' goods such as bread, cigarettes, shoes, etc., have small amplitude. The production and prices of such goods do not fluctuate widely. Raw materials and capital goods, on the other hand, such as wheat, copper, iron, cotton, etc., have cycles of wide amplitude. The fluctuations of successive cycles in the same series vary greatly, ranging from ten to forty percent. The price of a loaf of bread of the same weight rarely fluctuates more than from seven to twelve cents a loaf. Wheat, on the other hand, varies from \$0.25 to \$2.25 a bushel or nine times the lowest value. Copper fluctuates from \$0.06 to \$0.25 a pound while copper wire and alarm clocks change in price but a small fraction of the price variation of the raw material.

As indicated in Fig. 100, the length of cycles varies as much or more than their amplitude. The shortest cycles are about two years long while the longest ones are from seven to ten years in duration. Their average length is about three and one-half years. This wide variation in the length and amplitude of the cycle makes its measurement difficult and often inaccurate. Sometimes a depression lasts but a few months while others endure for years.

RESIDUAL METHOD

The traditional and still the most common method of measuring the cycle is to treat it as a residual. The trend and seasonal variation of a time series are removed by direct statistical attack through the use of averages and dispersions based on the original data. After these are taken out, the remainder is called the cycle. This method leaves a more or less *impure* cycle, containing all irregular movements and whatever of trend or seasonal change those computations failed to remove. This method is theoretically quite simple. It is like picking the beans and beads out of a box in which beans, beads, and buttons were originally mixed. The beans are picked out and put in one pile. That is like removing the trend. Then the beads are picked out and put into another pile. That is like taking out the seasonal variation. Then all that is left is called buttons, but it may include all the clasps, hooks and eyes, tacks, and dust and threads that have accumulated in the box over a long time. The residual method can never give a pure cycle. It is a conglomerate, a mixture including all movements except trend and seasonal variation, and if they have been incorrectly computed, some of their amplitude remains in the cyclical residual. This method is easy to compute and if carefully done will give results that are sufficiently accurate for most practical purposes.

The problem is somewhat like reducing a chemical compound. The elements must be separated, but it is a question of which one to eliminate first. The four major elements in a time series are Trend (T), Seasonal (S), Cycle (C), and Irregular Movements (I). The compound which must be dissolved is, therefore,

$$(T \times S \times C \times I),$$

if we consider the combined elements as bound together in a *multiplicative* or *ratio* structure. If their relationship is *additive*, the formula would be $(T + S + C + I)$. The ratio structure is frequently considered the correct one, but analyses are often made on the additive relation or on a combination of the two. The most commonly used formulas for the solution are:

1. $(C \times I) = \dfrac{T \times C \times S \times I}{T \times S}$

2. $(C \times I) = \dfrac{S \times C \times I}{S}$, and $\dfrac{T \times C \times S \times I}{T} = S \times C \times I$

3. $(C \times I) = \dfrac{T \times C \times I}{T}$, and $\dfrac{T \times S \times C \times I}{S} = T \times C \times I$

4. $(C + I) = \dfrac{T(S + C + I)}{T} - S$

The details of the first and fourth methods are illustrated in the worksheet on page 467.

The first method of computing cycle in Worksheet No. 86 is based on the assumption that the relationships among trend, seasonal, cycle, and irregular changes are all multiplicative or ratio connections. This means that if trend (T) and seasonal (S) are increasing, the increase is not their sum but their product. It is not $T + S$ but TS. It may be illustrated with arithmetic numbers; if $T = 120$, and $S = 95\%$, then $TS = 120 \times .95 = 114$. Since seasonal variation is computed as a percentage, this relationship is perfectly logical. If the original data are divided by T, on the assumption that the relationship is multiplicative, the result is a ratio, or percentage. Suppose that $\dfrac{\text{Data}}{T} = 110\%$. If $S = .95$, then $\frac{110}{95} = 115.8 - 100 = 15.8$. But if the relationship is considered to be additive, $110 - 95 = 15$. In the first case the result is 15.8. By the second method it is 15. Which is correct? There is no positive or categorical answer. From what is known, the relationships seem to vary with different series. Sufficient research has not been done to prove which method is preferable in every case or even in many cases. The most that can be done is to evaluate the logical relationships which exist among trend, seasonal, and cycle in a given series and use the method that seems to meet the logical relationships in each particular series. If it seems that the cycle expands in proportion as the trend and seasonal rise, and contracts as the trend and seasonal fall, the ratio method should be used. But if there is no ratio increase in the cycle as the trend and cycle increase, their relation is only a

WORKSHEET NO. 86

TWO ALTERNATIVE METHODS OF COMPUTING THE CYCLICAL-IRREGULAR MOVEMENTS IN THE WHOLESALE PRICE OF FRESH EGGS IN CHICAGO, 1940

Method No. 1

Year and Month	Original Data $T\times S\times C\times I$	Trend Values T	Seasonal Index S	Normal Values $T\times S$	Cyclical-Irregular Percentages $C\times I=\frac{T\times S\times C\times I}{T\times S}$	Cycle $(C\times I)-100$
1941						
Jan.	18.4	22.246	103.8	23.09	79.7	− 20.3
Feb.	16.7	22.274	84.4	18.80	88.8	− 11.2
Mar.	17.8	22.302	87.0	19.40	91.7	− 8.3
Apr.	21.6	22.330	87.4	19.52	110.7	+ 10.7
May	22.3	22.358	83.4	18.65	119.6	+ 19.6
June	25.4	22.386	79.4	17.77	142.9	+ 42.9
July	26.1	22.414	89.4	20.04	130.2	+ 30.2
Aug.	27.7	22.442	97.3	21.84	126.8	+ 26.8
Sept.	29.0	22.470	106.3	23.89	121.4	+ 21.4
Oct.	31.0	22.498	118.3	26.62	116.5	+ 16.5
Nov.	36.0	22.526	136.3	30.70	117.3	+ 17.3
Dec.	34.5	22.554	127.0	28.64	120.5	+ 20.5

Method No. 4

Year and Month	Original Data $T(S+C+I)$	Trend Values T	Percent of Trend $(S+C+I)=\frac{T(S+C+I)}{T}$	Seasonal Index Percent S	Cyclical-Irregular Percentages $(S+C+I)-S=C+I$
1941					
Jan.	18.4	22.246	82.71	103.8	− 21.1
Feb.	16.7	22.274	74.97	84.4	− 9.4
Mar.	17.8	22.302	79.81	87.0	− 7.2
Apr.	21.6	22.330	96.73	87.4	+ 9.3
May	22.3	22.358	99.74	83.4	+ 16.3
June	25.4	22.386	113.46	79.4	+ 44.0
July	26.1	22.414	116.44	89.4	+ 27.0
Aug.	27.7	22.442	123.43	97.3	+ 26.1
Sept.	29.0	22.470	129.06	106.3	+ 22.7
Oct.	31.0	22.498	137.79	118.3	+ 19.5
Nov.	36.0	22.526	159.81	136.3	+ 23.5
Dec.	34.5	22.554	152.97	127.0	+ 25.9

matter of the sum of the parts. This means that $T + S + C + I$ = Data, instead of $T \times S \times C \times I$ = Data.

Method No. 1 in effect says, "This trend which we *do* have times this seasonal which we *do* have gives us a clear measure of the business in its normal activities. The product of TS divided into the data gives us the actual cyclical fluctuation." The whole approach is realistic and positive. For this reason Method No. 1 is usually to be preferred in computing the cycle. Method No. 4 is based on the idea that the cycle does not vary in size with variation in the amplitude of trend and seasonal. In other words, the cycle is only an *added* fluctuation which may be *subtracted* out. In some series this seems to be the case. It has the advantage of being less laborious to compute. It requires only six columns instead of seven. It substitutes one division and a simple subtraction for either one multiplication, one division and one subtraction, or two divisions and one subtraction. It seems probable that in many cases the cycle does not expand in ratio to the amplitude of seasonal. In other words, its relation to seasonal is additive instead of multiplicative. Method No. 1 is used in Worksheet No. 87 and Method No. 4 in Worksheet No. 88.

We began our time series analysis to discover the trend, seasonal variation, and cycle of the original data. Worksheet No. 87 completes this task. In column 2 are the original data by months. In column 3 is the monthly trend. In column 4 is the seasonal index. In column 7 is the cycle in percentage deviations from normal. The analysis is complete at the end of column 7 for most purposes. The reason for computing column 9 is to express the cycle in terms of the standard deviation of the percentage figures. The two cycles in columns 7 and 9 are identical in all particulars, shape, quantities, and monthly signs, except that the first is in percentages of normal and the second is in standard deviations of those percentages. The purpose of computing the last cycle is to make it more readily comparable with other cycles. Two or more cycles stated in percentages are not easily compared because of the wide range of the percentages in one cycle as compared with another. For instance, the percentage variations of the cycle of bread prices might possibly range

WORKSHEET NO. 87

COMPLETE ANALYSES OF THE TREND SEASONAL VARIATION AND CYCLICAL FLUCTUATIONS OF GASOLINE STOCKS, END OF MONTH, 1945–1950, SURVEY OF CURRENT BUSINESS

1 Year and month	2 Original Data	3 Monthly Secular Trend	4 Seasonal Index	5 Normal 3 × 4	6 Percentage Cycle 2 ÷ 5	7 Cycle Percent	8 Squared Percentage Cycle	9 Cycle σ
1945								
Jan.	78.9	84.416	107.5	90.75	86.9	− 13.1	171.61	− 1.54
Feb.	85.5	84.664	109.7	92.85	92.1	− 7.9	62.41	− .93
Mar.	85.7	84.872	110.6	93.87	91.3	− 8.7	75.69	− 1.02
Apr.	79.7	85.100	108.1	91.99	86.6	− 13.4	179.56	− 1.58
May	77.2	85.328	106.0	90.45	85.4	− 14.6	213.16	− 1.72
June	74.1	85.556	102.2	87.44	84.7	− 15.3	234.09	− 1.80
→July←	74.5	85.784	95.7	82.10	90.7	− 9.3	86.49	− 1.09
Aug.	74.3	86.012	92.1	79.22	93.8	− 6.2	38.44	− .73
Sept.	65.5	86.240	88.2	76.06	86.1	− 13.9	193.21	− 1.64
Oct.	68.0	86.468	88.3	76.35	89.1	− 10.9	118.81	− 1.28
Nov.	78.1	86.696	91.6	79.41	98.3	− 1.7	2.89	− .20
Dec.	89.4	86.924	100.0	86.92	102.9	+ 2.9	8.41	+ .34
1946								
Jan.	94.1	87.152	107.5	93.69	100.4	+ .4	.16	+ .05
Feb.	96.3	87.380	109.7	95.89	100.5	+ .5	.25	+ .06
Mar.	95.2	87.608	110.6	96.89	98.3	− 1.7	2.89	− .20
Apr.	90.4	87.836	108.1	94.95	95.2	− 4.8	23.04	− .56
May	85.8	88.064	106.0	93.35	91.9	− 8.1	65.61	− .95
June	83.7	88.292	102.2	90.23	92.8	− 7.2	51.84	− 85
July	79.4	88.520	95.7	84.71	93.7	− 6.3	39.69	− .74
Aug.	78.8	88.748	92.1	81.74	96.4	− 3.6	12.96	− .42
Sept.	78.8	88.976	88.2	78.48	100.4	+ .4	.16	+ .05
Oct.	77.6	89.204	88.3	78.77	98.5	− 1.5	2.25	− .18
Nov.	80.0	89.432	91.6	81.92	97.7	− 2.3	5.29	− .27
Dec.	84.5	89.660	100.0	89.66	94.2	− 5.8	33.64	− .68
1947								
Jan.	90.3	89.888	107.5	96.63	93.4	− 6.6	43.56	− .78
Feb.	95.0	90.116	109.7	98.86	96.1	− 3.9	15.21	− .46
Mar.	97.0	90.344	110.6	99.92	97.1	− 2.9	8.41	− .34
Apr.	92.7	90.572	108.1	97.91	94.7	− 5.3	28.09	− .62
May	86.7	90.800	106.0	96.25	90.1	− 9.9	98.01	− 1.16
June	81.2	91.028	102.2	93.03	87.3	− 12.7	161.29	− 1.49
July	77.1	91.256	95.7	87.33	88.3	− 11.8	139.24	− 1.39
Aug.	77.2	91.484	92.1	84.26	91.6	− 8.4	70.56	− .99
Sept.	75.9	91.712	88.2	80.89	93.8	− 6.2	38.44	− .73
Oct.	74.7	91.940	88.3	81.18	92.0	− 8.0	64.00	− .94
Nov.	78.7	92.168	91.6	84.43	93.2	− 6.8	46.24	− .80
Dec.	83.1	92.396	100.0	92.40	89.9	− 10.1	102.01	− 1.19

In Worksheet No. 87 the relationship between monthly secular trend, column 3, and the seasonal index, column 4, is multiplicative instead of additive. They are multiplied to obtain column 5.

WORKSHEET NO. 87 (continued)

1 Year and month	2 Original Data	3 Monthly Secular Trend	4 Seasonal Index	5 Normal 3 × 4	6 Percentage Cycle 2 ÷ 5	7 Cycle Percent	8 Squared Percentage Cycle	9 Cycle σ
1948								
Jan	93.3	92.624	107.5	99.57	93.7	− 6.3	39.69	− .74
Feb.	102.3	92.852	109.7	101.86	100.4	+ .4	.16	+ .05
Mar.	103.4	93.080	110.6	102.95	100.4	+ .4	.16	+ .05
Apr.	101.3	93.308	108.1	100.87	100.4	+ .4	.16	+ .05
May	99.6	93.536	106.0	99.15	100.5	+ .5	.25	+ .06
June	96.2	93.764	102.2	95.83	100.4	+ .4	.16	+ .05
July	90.3	93.992	95.7	89.95	100.4	+ .4	.16	+ .05
Aug.	87.2	94.220	92.1	86.78	100.5	+ .5	.25	+ .06
Sept.	92.3	94.448	88.2	83.30	110.8	+ 10.8	116.64	+ 1.27
Oct.	94.0	94.676	88.3	83.60	112.4	+ 12.4	153.76	+ 1.46
Nov.	87.3	94.904	91.6	86.93	100.4	+ .4	.16	+ .05
Dec.	95.4	95.132	100.0	95.13	100.3	+ .3	.09	+ .03
1949								
Jan.	108.5	95.360	107.5	102.51	105.8	+ 5.8	33.64	+ .68
Feb.	117.5	95.588	109.7	104.86	112.1	+ 12.1	146.41	+ 1.42
Mar.	118.8	95.816	110.6	105.97	112.1	+ 12.1	146.41	+ 1.42
Apr.	117.0	96.044	108.1	103.82	112.7	+ 12.7	161.29	+ 1.49
May	113.2	96.272	106.0	102.05	110.9	+ 10.9	118.81	+ 1.28
June	106.0	96.500	102.2	98.62	107.5	+ 7.5	56.25	+ .88
July	103.9	96.728	95.7	92.57	112.2	+ 12.2	148.84	+ 1.44
Aug.	97.7	96.956	92.1	89.30	109.4	+ 9.4	88.36	+ 1.11
Sept.	94.4	97.184	88.2	85.72	110.1	+ 10.1	102.01	+ 1.19
Oct.	96.2	97.412	88.3	86.02	111.8	+ 11.8	139.24	+ 1.39
Nov.	97.2	97.640	91.6	89.44	108.7	+ 8.7	75.69	+ 1.02
Dec.	103.6	97.868	100.0	97.87	105.9	+ 5.9	34.81	+ .69
1950								
Jan.	116.6	98.096	107.5	105.45	110.6	+ 10.6	112.36	+ 1.25
Feb.	124.2	98.324	109.7	107.86	115.1	+ 15.1	228.01	+ 1.78
Mar.	124.9	98.552	110.6	109.00	114.6	+ 14.6	213.16	+ 1.72
Apr.	119.6	98.780	108.1	106.78	112.0	+ 12.0	144.00	+ 1.41
May	112.9	99.008	106.0	104.95	107.6	+ 7.6	57.76	+ .89
June	106.0	99.236	102.2	101.42	104.5	+ 4.5	20.25	+ .53
July	102.8	99.464	95.7	95.19	108.0	+ 8.0	64.00	+ .94
Aug.	99.4	99.692	92.1	91.82	108.3	+ 8.3	68.89	+ .98
Sept.	97.9	99.920	88.2	88.13	111.1	+ 11.1	123.21	+ 1.31
Oct.	97.8	100.148	88.3	88.43	110.6	+ 10.6	112.36	+ 1.25
Nov.	98.4	100.376	91.6	91.94	107.0	+ 7.0	49.00	+ .82
Dec.	104.3	100.604	100.0	100.60	103.7	+ 3.7	13.69	+ .44

$$\sigma = \sqrt{\frac{5{,}207.70}{72}} = \sqrt{72.3292} = 8.5$$

Sum of Col. 8 = 5,207.70

The multiplicative relationship between trend and cycle means that as trend rises the absolute fluctuations of seasonal variation tend to increase in amplitude by the same amount. Plotting the original data of all the years on a seasonal graph will aid in discovering this relationship.

WORKSHEET NO. 88

COMPLETE ANALYSIS OF THE TREND SEASONAL VARIATION AND CYCLICAL FLUCTUATIONS OF WHOLESALE PRICE OF EGGS IN CHICAGO, 1944–1948, SURVEY OF CURRENT BUSINESS

1		2	3	4	5	6	7	8
Year	Month	Original Data	Secular Trend	Actual Trend	Seasonal Index	Cycle in Percentage	Squared Percentage Cycle	Cycle in σ
1944	Jan.	35.0	36.544	95.8	98.0	− 2.2	4.84	− .34
	Feb.	33.4	36.657	91.1	93.5	− 2.4	5.76	− .37
	Mar.	32.1	36.770	87.3	92.6	− 5.3	28.09	− .82
	Apr.	31.1	36.883	84.3	92.9	− 8.6	73.96	− 1.33
	May	30.8	36.996	83.2	92.7	− 9.5	90.25	− 1.47
	June	33.2	37.109	89.5	94.3	− 5.8	33.64	− .90
	July	34.8	37.222	93.5	98.2	− 4.7	22.09	− .73
	Aug.	33.8	37.335	90.5	96.5	− 6.0	36.00	− .93
	Sept.	36.8	37.448	98.3	102.8	− 4.5	20.25	− .70
	Oct.	38.9	37.561	103.6	109.6	− 6.0	36.00	− .93
	Nov.	42.3	37.674	112.3	114.4	− 2.1	4.41	− .32
	Dec.	41.8	37.787	110.6	114.5	− 3.9	15.21	− .60
1945	Jan.	38.0	37.900	100.3	98.0	+ 2.3	5.29	+ .36
	Feb.	34.9	38.013	91.8	93.5	− 1.7	2.89	− .26
	Mar.	34.3	38.126	90.0	92.6	− 2.6	6.76	− .40
	Apr.	34.3	38.239	89.7	92.9	− 3.2	10.24	− .49
	May	34.3	38.352	89.4	92.7	− 3.3	10.89	− .51
	June	35.1	38.465	91.3	94.3	− 3.0	9.00	− .46
	July	35.6	38.578	92.3	98.2	− 5.9	34.81	− .91
	Aug.	37.8	38.691	97.7	96.5	+ 1.2	1.44	+ .19
	Sept.	34.6	38.804	89.2	102.8	− 13.6	193.21	− 2.10
	Oct.	40.1	38 917	103.0	109.6	− 4.6	21.16	− .71
	Nov.	43.7	39.030	112.0	114.4	− 2.4	5.76	− .37
	Dec.	42.9	39.143	109.6	114 5	− 4.9	24.01	− .76
1946	Jan.	→35.6←	39.256	90.7	98.0	− 7.3	53.29	− 1.13
	Feb.	33.1	39.369	84.1	93.5	− 9.4	88.36	− 1.45
	Mar.	33.2	39.482	84.1	92.6	− 8.5	72.25	− 1.31
	Apr.	33.3	39.595	84.1	92.9	− 8.8	77.44	− 1.36
	May	33.6	39.708	84.6	92.7	− 8.1	65.61	− 1.25
	June	33.2	39.821	83.4	94.3	− 10.9	118.81	− 1.68
	July	34.0	39.934	85.1	98.2	− 13.1	171.61	− 2.02
	Aug.	34.6	40.047	86.4	96.5	− 10.1	102.01	− 1.56
	Sept.	40.6	40.160	101.1	102.8	− 1.7	2.89	− .26
	Oct.	42.0	40.273	104.3	109.6	− 5.3	28.09	− .82
	Nov.	40.6	40.386	100.5	114.4	− 3.9	15.21	− .60
	Dec.	38.8	40.499	95.8	114.5	− 8.7	75.69	− 1.34

In Worksheet No. 88 the relationship between secular trend, column 3, and seasonal index, column 5 is additive. They are separated by subtraction instead of by multiplication or division as is done in Worksheet No. 87. This means that seasonal variation remains constant although trend rises.

WORKSHEET NO. 88 (continued)

1 Year	Month	2 Original Data	3 Secular Trend	4 Actual Trend	5 Seasonal Index	6 Cycle in Percentage	7 Squared Percentage Cycle	8 Cycle in σ
1947	Jan.	38.8	40.612	95.5	98.0	− 2.5	6.25	− .38
	Feb.	37.8	40.725	92.8	93.5	− 0.7	.49	− .11
	Mar.	41.8	40.838	102.4	92.6	+ 9.8	96.04	+ 1.51
	Apr.	42.5	40.951	103.8	92.9	+ 10.9	118.81	+ 1.68
	May	40.9	41.064	99.6	92.7	+ 6.9	47.61	+ 1.07
	June	41.4	41.177	100.5	94.3	+ 6.2	38.44	+ .96
	July	43.4	41.290	105.1	98.2	+ 6.9	47.61	+ 1.07
	Aug.	42.2	41.403	101.9	96.5	+ 5.4	29.16	+ .83
	Sept.	45.0	41.516	108.4	102.8	+ 5.6	31.36	+ .87
	Oct.	46.4	41.629	111.5	109.6	+ 1.9	3.61	+ .29
	Nov.	45.5	41.742	109.0	114.4	− 5.4	29.16	− .83
	Dec.	51.7	41.855	123.5	114.5	+ 9.0	81.00	+ 1.39
1948	Jan.	44.1	41.968	105.1	98.0	+ 7.1	50.41	+ 1.10
	Feb.	43.4	42.081	103.1	93.5	+ 9.6	92.16	+ 1.48
	Mar.	43.2	42.194	102.4	92.6	+ 9.8	96.04	+ 1.51
	Apr.	42.9	42.307	101.4	92.9	+ 8 5	72.25	+ 1.31
	May	41.0	42.420	96.7	92.7	+ 4.0	16.00	+ .62
	June	41.6	42.533	97.8	94.3	+ 3.5	12.25	+ .54
	July	41.2	42.646	96.6	98.2	− 1.6	2.56	− .25
	Aug.	44.4	42.759	103.8	96.5	+ 7.3	53.29	+ 1.13
	Sept.	44.2	42.872	103.1	102.8	+ 0.3	.09	+ .05
	Oct.	45.6	42.985	106.1	109.6	− 3.5	12.25	− .54
	Nov.	49.4	43.098	114.6	114.4	+ .2	.04	+ .03
	Dec.	8.2	43.211	111.5	114.5	− 3.0	9.00	− .46
							2,513.10	

$$\sigma = \sqrt{\frac{\overline{2X^2}}{N}} = \sqrt{\frac{2,513.10}{60}} = \sqrt{41.885} = 6.47$$

from 10% above normal to 15% below normal, while wheat price cycles would certainly range from 100% or more above normal to 60 to 80% below normal. The wheat cycle would be so deep and the bread cycle so shallow that close comparisons between the two would be difficult. It would be somewhat like comparing a pig and an elephant. Although these two animals are much alike in shape, they differ so much in size that most persons might see little similarity between them. When, however, two or more cycles are reduced to their own standard deviations, they tend to have about the same amplitude. This makes comparison for forecasting purposes easier.

The methods followed in Worksheet No. 88 for securing trend, seasonal variation, and cycle are the shortest and easiest available and give fairly good results. The trend is from Worksheet No. 73

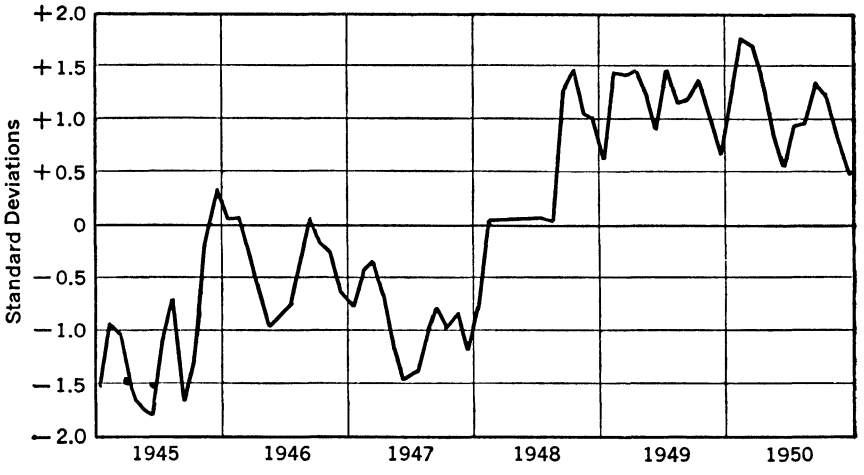


FIG. 101. Cyclical fluctuations of gasoline stocks, 1945–1950, in standard deviations, 1950. (Worksheet No. 87, column 9)

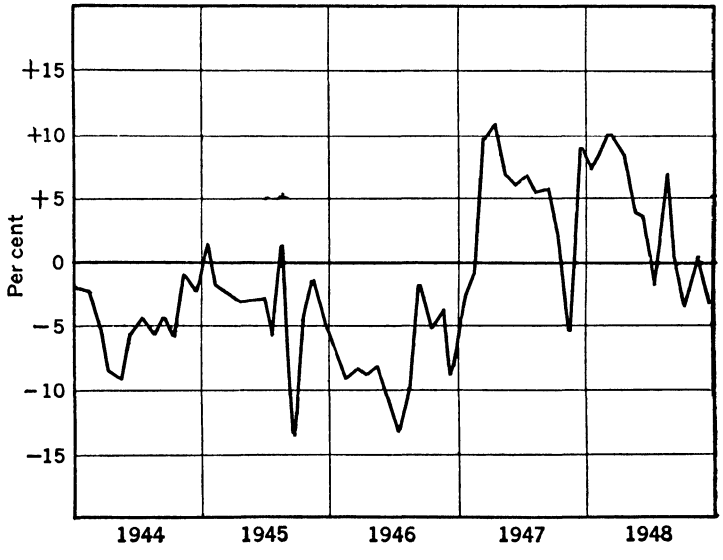


FIG. 102. Cyclical fluctuations of wholesale prices of eggs at Chicago, 1944–1948, in percentages. (Worksheet No. 88, column 6)

(p. 431). The seasonal variation is from Worksheet No. 79 (p. 448). Since the seasonal index in the Third Method is based on the percentage figures of column 4 in Worksheet No. 88 there is almost no additional labor required to compute the seasonal variation by this method. By means of the three Worksheets, No. 73, No. 79, and No. 88 time series analysis is made relatively short and simple.

DIRECT METHOD OF COMPUTING CYCLE

As the previous sections of this chapter indicate, the Residual Method of computing cycles leaves much to be desired. Since all errors in computing trend and seasonal variation under the residual method affect and distort the cycle, it is preferable to be able to make a direct statistical analysis of the cycle by means of averages and dispersions based on the original data. This direct measurement of the cycle has been developed by Wesley C. Mitchell and Arthur F. Burns and was first published in *Bulletin 57*, July 1, 1935, by the *National Bureau of Economic Research*. The larger volume, "Measuring Business Cycles," by the same authors in 1946 refines their theory and methods.

Essential Steps

The essential steps in the direct analysis of business cycles are:

A. *Securing the Cycles*

1. Run a 12-months' moving average through the original data to eliminate seasonal variation. See Worksheet No. 89.

2. Plot the de-seasonalized data on coordinate graph paper as in Fig. 103.

3. Divide the plotted data line into specific cycles by marking the low points and the high points as indicated in Fig. 103 by the x's.

B. *Measuring Length of Cycles*

4. For each cycle, record in a comparative table the date of (1) Revival, (2) Peak, and (3) Trough as indicated in Worksheet No. 90.

5. For each cycle in the worksheet the length of (4) the period

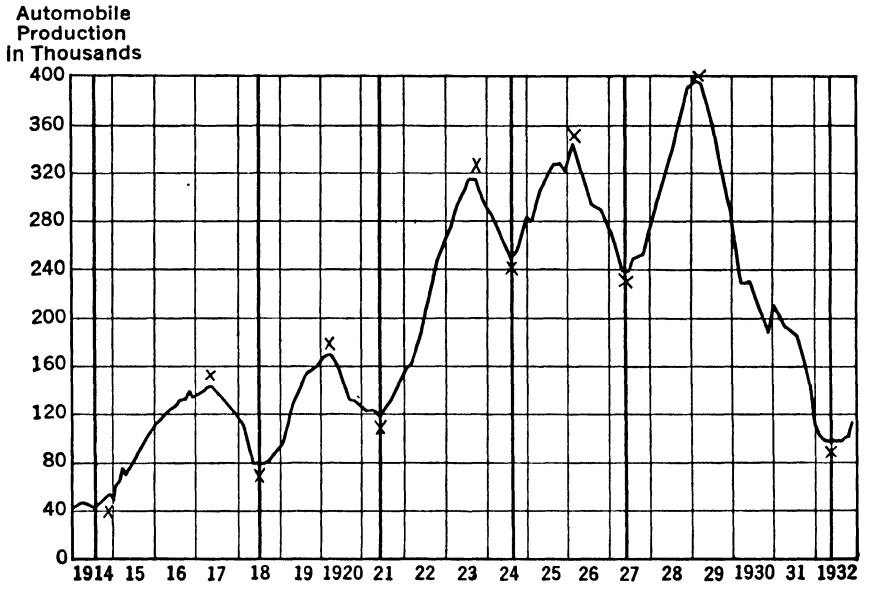


FIG. 103. Data for automobile production in U.S., 1914–1932, corrected for seasonal variation by centered 12-months' moving average, with cycles marked. (From Worksheet No. 89)

of expansion, and (5) the length of the period of contraction in months is recorded, together with (6) the total length of each cycle.

6. The periods (7) expansion, and (8) contraction are stated as percentages of the total length of the cycle in months.

7. The average of each of the columns indicated in steps No. 5 and No. 6 and the average deviation of the items from the mean are computed.

The computations in steps Nos. 4–7 give us the average length of the cycles and their deviation from that average.

C. *Measuring the Amplitude of Cycles*

8. In Worksheet No. 91 the amplitude of the cycles is measured by considering the average of the original de-seasonalized data for each cycle as 100.

9. With the base in No. 8 as 100, compute the percentage level of (2) the initial trough; (3) the peak; (4) the terminal trough;

(5) the amount of rise from (2) to (3); (6) the amount of fall from (3) to (4); and (7) the total amount of rise *and* fall for the entire cycle, together with the amount of change per month for the rise, the fall and the total rise and fall. This computation gives a double measure of the amplitude of each cycle (1) by totals, (2) by months.

10. The average is computed for each of the points in No. 9, and the average deviation of the cycles from their mean.

D. *Computing Secular Trend of Cycles*

11. Since the direct analysis of business cycles is based on the original data, it necessarily includes the secular trend, which must be removed to obtain a pure cycle. This correction is made by computing the percent of change in trend from one cycle to the next (1) from the preceding phase of the same cycle and (2) from the same phase of the preceding cycle. This is shown in Worksheet No. 92.

E. *Cycle Patterns, or Contours*

12. Each cycle is divided at nine points and the amount of change between each pair of points measured. This reveals the

WORKSHEET NO. 89

DATA OF AUTOMOBILE PRODUCTION IN U.S., 1914–1932, IN 1,000, CORRECTED FOR SEASONAL VARIATION BY CENTERED 12-MONTH MOVING AVERAGE

Years	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1914	41	43	45	47	47	45	44	43	46	48	51	54
1915	48	59	63	74	70	75	81	87	91	96	102	106
1916	110	112	115	118	122	125	128	130	131	132	138	134
1917	136	138	139	142	143	140	137	134	130	127	123	122
1918	117	112	102	92	81	79	79	79	80	81	83	87
1919	91	98	107	118	131	138	145	150	156	156	158	162
1920	166	168	169	166	162	159	150	141	133	132	129	127
1921	126	125	124	124	122	120	123	128	131	137	142	149
1922	153	160	163	170	179	190	202	214	229	240	250	257
1923	263	269	279	289	296	302	307	314	316	316	310	300
1924	295	290	287	281	274	266	260	252	250	254	262	274
1925	283	281	281	292	303	311	317	322	328	328	329	327
1926	324	338	345	336	327	315	309	299	295	292	291	287
1927	280	271	261	252	242	329	240	248	250	251	252	258
1928	267	277	288	301	311	318	330	339	351	366	377	386
1929	393	396	397	396	387	377	368	357	341	327	315	301
1930	284	262	247	229	229	231	223	215	208	200	193	187
1931	208	206	200	196	191	189	186	185	173	160	151	141
1932	109	102	99	97	96	95	96	96	96	99	100	110

cycle contour as to whether it rises rapidly and falls slowly or rises slowly and falls rapidly or is irregular in movement. This contour is measured in Worksheet No. 93.

13. Several other characteristics of the cycles are measurable by the direct method, such as lead or lag with the general business cycle, month-to-month trend, and other details.

WORKSHEET NO. 90

LENGTH OF SPECIFIC CYCLES OF AUTOMOBILE PRODUCTION IN U.S., 1914–1932

Dates of Specific Cycles			Duration of Cyclical Movements				
			Expansion Months	Contraction Months	Full Cycle Months	Percentages	
Revival	Peak	Trough				Expansion	Contraction
1	2	3	4	5	6	7	8
1. Aug., 1914,	May, 1917,	July, 1918	34	14	48	71	29
2. Aug., 1918,	Mar., 1920,	June, 1921	20	15	35	57	43
3. July, 1921,	Sept., 1923,	Sept., 1924	27	12	39	69	31
4. Oct., 1924,	Mar., 1926,	June, 1927	18	15	33	55	45
5. July, 1927,	Mar., 1929,	June, 1932	21	39	60	35	65
Arithmetic Mean			24	19	43	57	43
Mean Deviation			5	8	9	10	10

LENGTH OF SPECIFIC CYCLES

After the monthly de-seasonalized data in Worksheet No. 89 are plotted on Fig. 103 the lowest and highest points in each cycle are marked with an *X* and a heavy vertical line is drawn through the lowest points to distinguish the separate cycles. The revival month and year, the peak month and year, and the terminal trough month and year of each cycle are written in columns 1, 2 and 3 of Worksheet No. 90 and the number of months in the expansion leg of the cycle, or from the initial revival to peak are written in column 4, the number of months in the contraction leg, or from peak to trough, are written in column 5, and the total length of each cycle in months is written in column 6. With the total length of the cycle in months as 100%, the expansion and the contraction periods are reduced to percentages of the total duration of the cycle. The mean and mean deviation are computed for each of the five columns. For the five cycles of automobile production the mean expansion period is 24 months with

a mean deviation of 5 months. The mean contraction period is 19 months with a mean deviation of 8 months. The average length of the five cycles is 43 months with a mean deviation of 9 months. Fifty-seven percent of the average automobile cycle is expansion while 43 percent is contraction. The mean deviation is 10 percent. Although a set of means based on a larger number of cycles would give more dependable results, this direct statistical attack on the automobile production cycle gives us quite a clear picture of what happens in this industry in the variations from prosperity to depression. Automobile production cycles vary from 33 to 60 months and average about three and one-half years in length.

WORKSHEET NO. 91

AMPLITUDE OF SPECIFIC CYCLES OF AUTOMOBILE PRODUCTION IN U.S., 1914-1932

	Dates of Specific Cycles			Standing			Total Movement		
	Revival	I Peak	Trough	At Initial Trough 2	At Peak 3	At Terminal Trough 4	Rise 5	Fall 6	Rise and Fall 7
1	Aug., 1914,	May, 1917,	July, 1918	42	137	77	95	60	155
2.	Aug., 1918,	Mar., 1920,	June, 1921	61	130	94	69	36	105
3	July, 1921,	Sept., 1923,	Sept., 1924	53	133	107	80	26	106
4.	Oct., 1924,	Mar., 1926,	June, 1927	86	114	82	28	32	60
5.	July, 1927,	Mar., 1929,	June, 1932	95	155	38	60	117	177
	Arithmetic Mean			67	134	80	66	54	120
	Mean Deviation			18	10	18	18	27	36

AMPLITUDE OF SPECIFIC CYCLES

The amplitude of cycles is based on the same principle as seasonal variation. The average of the entire cycle is considered 100% and the low and high points are stated as percentages of this average. The average monthly automobile production from August, 1914, to July, 1918, inclusive, as given in Worksheet No. 89 is taken as 100%. The average production of the three months centered on August, 1914, or July, August, and September, is taken as the initial trough and is expressed as a percentage of the cycle average. In this case the figure is 42 in Worksheet No. 91. The value for the peak is the average centered on May, 1917, or the average production of April, May, and June, 1917.

This figure is expressed as a percentage of the cycle average. In this case it is 137. The value of the terminal trough is the monthly average of the three months centered on July, 1918, or June, July, and August, 1918, expressed as a percentage of the average of the cycle. In this case the figure is 77. The figures for the other cycles are obtained by the same method. The average of each cycle is the base of 100 for that cycle.

SECULAR MOVEMENT IN CYCLES

Since the direct statistical attack on business cycles begins with the original data which necessarily include whatever trend there is in the business, a part of the direct analysis is the measurement of the trend. The theory underlying this method is that if there were no trend, that is, if the trend were perfectly level, the average difference between the level of cycles would be zero. On the average they would all be on the same level. If, however, there is an upward trend, each succeeding cycle will tend to be on a higher level. If the trend is downward, each succeeding cycle will tend to be on a lower level. This change in trend from cycle to cycle may be measured in several ways. Two methods are introduced here. The first method is to measure the change in the average standing of each succeeding *phase* of each cycle as related to the next preceding phase. This measurement is made in columns 5 and 6 in Worksheet No. 92.

The Average Monthly Standing of the cycles in columns 2, 3, and 4 in Worksheet No. 92 is obtained as follows:

1. The original de-seasonalized values of the data in Worksheet No. 74 are averaged for each expansion period and each contraction period of each cycle. The 34 months from August, 1914, to May, 1917, are averaged for the first expansion period of the first cycle. This is 100, which is written in column 2 in Worksheet No. 92. The data values for the contraction period, June, 1917, to July, 1918, or 14 months, are averaged and are 113, which is written in column 3 in Worksheet No. 92. The average for the total cycle is 103 and is written in column 4.

2. The Average Monthly Standing values for all the other

WORKSHEET NO. 92

MEASURES OF SECULAR MOVEMENT IN AUTOMOBILE PRODUCTION CYCLES IN U.S., 1914–1932

Dates of Specific Cycles			Average Monthly Standing			Percent Change from Average of Preceding Phase of Cycle		Percent Change from Average of Same Phase of Preceding Cycle		
Revival	1 Peak	Trough	Expansion 2	Contraction 3	Full Cycle 4	Expansion 5	Contraction 6	Expansion 7	Contraction 8	Full Cycle 9
1. Aug., 1914,	May, 1917,	July, 1918	100	113	103		+ 13			
2. Aug., 1918,	Mar., 1920,	June, 1921	131	130	130	+ 16	− 1	+ 31	+ 15	+ 27
3. July, 1921,	Sept., 1923,	Sept., 1924	217	282	237	+ 67	+ 30	+ 66	+ 130	+ 82
4. Oct., 1924,	Mar., 1926,	June, 1927	306	285	297	+ 8	− 7	+ 40	+ 1	+ 26
5. July, 1927,	Mar., 1929,	June, 1932	314	222	254	+ 10	− 40	+ 3	− 22	− 14
Arithmetic Mean						+ 20	− 1	+ 28	+ 25	+ 24
Mean Deviation						15	18	16	28	15

cycles are obtained by the same method. To compute these values the cycle dates and lengths in columns 1, 2, and 3 of Worksheet No. 90 and the monthly data in Worksheet No. 89 are necessary.

The + 13 in column 6 of Worksheet No. 92 means that the average height of the cycle in its contraction period, which is 113, is 13% above the preceding expansion level in column 2.

The + 16 in column 5, second line, means that the expansion period of the second cycle is 16% above the contraction level of the first cycle. The level rose from 113 to 131 which is 16% up.

This second cycle fell from 131 to 130 or − 1% in its second phase. The next cycle, however, in its expansion period rose to 217 which is 67% over the 130 of the second phase of the preceding cycle. All the other figures in columns 5 and 6 are obtained by the same methods.

Since the average of the four expansion periods is up 20% and the contraction periods on the average are down only − 1%, the analysis shows that there is an upward trend of approximately 20% per cycle.

Columns 7, 8, and 9 measure the trend for each phase of each cycle as related to the same phase of the preceding cycle. The expansion period of the second cycle is 31% above the expansion period of the first cycle. The contraction period of the second cycle is 15% above the contraction period of the first cycle. The average upward trend from expansion phase to the next expansion phase is + 28%. For the contraction phase it is + 25%. For the cycles as a whole the upward trend averages 24%. This rapid upward trend is clearly indicated by Fig. 104.

This direct measurement of the trend from cycle to cycle does not remove the trend. It merely measures it.

The last characteristic of the cycle which we shall measure in this direct analysis is the Cycle Pattern or Contour as indicated in Worksheet No. 93.

1. The three principal measuring points of each cycle, (1) the initial trough, (2) the peak, and (3) the terminal trough as indicated in columns 2, 3, and 4, of Worksheet No. 91 are written in columns 2, 6, and 10 of Worksheet No. 93. (This method is

WORKSHEET NO. 93

SPECIFIC CYCLE PATTERNS, OR CONTOUR OF AUTOMOBILE PRODUCTION IN U.S., 1914–1932

Dates of Specific Cycles I Revival	Peak	Trough	I 3 Months Centered on Initial Revival 2	II 1st Quarter 3	III 2nd Quarter 4	IV 3rd Quarter 5	V 3 Months Centered on Recession 6	VI 1st Quarter 7	VII 2nd Quarter 8	VIII 3rd Quarter 9	IX 3 Months Centered on Terminal Revival 10
1. Aug., 1914,	May, 1917,	July, 1918	42	72	103	126	137	130	120	90	77
2. Aug., 1918,	Mar., 1920,	June, 1921	61	67	100	120	130	115	100	96	94
3. July, 1921,	Sept., 1923,	Sept., 1924	53	64	90	118	133	126	121	112	107
4. Oct., 1924,	Mar., 1926,	June, 1927	86	96	105	110	114	104	98	88	82
5. July, 1927,	Mar., 1929,	June, 1932	95	100	122	138	155	112	76	73	38
Arithmetic Mean			67	80	104	122	136	117	103	92	80
Mean Deviation			18	15	8	5	10	8	14	10	18

slightly different from that used by Mitchell, but is easier for the beginning student and is sufficiently accurate.)

2. The expansion phase of each cycle is divided into four parts In cycle one the height of the cycle is measured every 8 months. In cycle two the height of the cycle is measured each 5 months, and so on with the expansion phase of each cycle. The height of the cycle at the end of the first quarter of the expansion phase is written in column 3 of Worksheet No. 93. The height of the cycle at the middle of the expansion phase is written in column 4, and the value at the end of the third quarter in column 5.

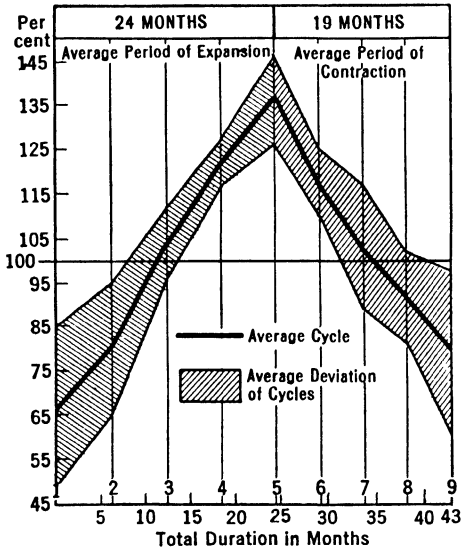


FIG. 104. Average automobile production cycle and deviations. (From Worksheet No. 93)

3. The contraction is also divided into four quarters. For the first cycle with a contraction period of 14 months, the measure of height is taken every three and one-half months. For the second cycle with a 15-month contraction period, the level is measured every 4 months. The value for the first quarter is written in column 7, the second quarter in column 8 and the third quarter in column 9.

4. This method splits each entire cycle into eight time segments with a measure of cycle level for each segment. From these measures the contour of the cycle can be plotted as in Fig. 104.

Automobile production cycles rise most rapidly during the middle half of the expansion period and fall most rapidly during the first half of the contraction period. Their variation from average is more marked at the beginning and end of the cycle and narrower during the middle two-thirds of the cycle's duration.

The details of the methods presented above differ in some minor

points from Mitchell's analysis, but in the main follow the principles of his direct statistical attack on the cycles. In Worksheet No. 93 we have measured the height of the cycle at the end of each quarter of expansion and at the end of each quarter of contraction, while Mitchell measures it at the middle of the thirds. We have also left out the measurement of Reference Cycles and lead and lag as matters which can be studied more profitably by students who are specializing in this field and are taking advanced courses in business cycles. Our purpose is to show the basic principles of the direct analysis of the cycle as compared with the older and more widely used residual method.

SUMMARY

1. Cyclical fluctuations in business and social activities are usually not uniform in either duration or amplitude.

2. There may be two or more series of cycles of varying length and amplitude running concurrently.

3. Cycles may be measured by the *residual* method which consists of computing out the portions of the total time series activity which are due to secular trend and seasonal variation and then calling all that remains *cyclical fluctuation.* This method is often quite inaccurate because it includes with the cycle all the accidental changes besides those portions of the trend and seasonal change which may not have been removed.

3. The *direct* method attacks the cycle directly with the prime statistical devices of averages, deviations, ratios, and percentages, just as the seasonal variation and secular are attacked by direct methods.

4. The relationship between cycles and seasonal variation may be either an additive or multiplicative relationship.

5. The relationship between cycles and secular trend may be either an additive or multiplicative relationship.

6. The similarity of the fluctuations of two or more cycles may be measured by computing the coefficient of correlation between them for several different amounts of lag. First, the identical months of the two cycles may be correlated. Second, one cycle may be dropped back one month, then two months, and so on for as long a lag as is desired. The length of the lag that gives the maximum correlation is the one that shows the full similarity of the cycles.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapters XXII, XXIV. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XIX. Prentice-Hall, Inc., New York, 1948.

Mills, F. C., *Statistical Methods, Revised*, Chapter VIII. Henry Holt & Company, New York, 1938.

Mitchell, W. C., *Business Cycles and Their Causes*, Part III. University of California Press, Berkeley, California, 1942.

Mitchell, W. C., and Burns, A. F., *The National Bureau's Measure of Cyclical Behavior*, Bulletin 57. National Bureau of Economic Research, 1819 Broadway, New York, 1935.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapter VII. John Wiley & Sons, New York, 1942.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics*, Chapter XVI. McGraw-Hill Book Company, New York, 1951.

REVIEW QUESTIONS

1. Differentiate a business cycle from irregular and accidental variations.
2. What is meant by "the business cycle" and "a business cycle"? How many business cycles are there? Explain in detail.
3. What types of cycles fluctuate most widely? less widely? Why?
4. What is the "residual method" of business cycle analysis? What are its strong points and its weak points?
5. Are the relationships among trend, seasonal variation, and cycle additive or multiplicative? What is meant by *additive* and *multiplicative* in relation to business cycles?
6. What are the differences between the two methods of computing the cycle in Worksheet No. 86? What are the advantages and disadvantages of each?
7. What are the reasons for thinking that business cycles expand and contract in proportion to the steepness of the trend?
8. What is meant by the phrase, "direct analysis of the business cycle"?
9. Name and explain the three steps in securing the cycles for the direct analysis.
10. Why is a 12-month moving average run through the data before the cycles are measured?

11. How is the duration or length of the cycles measured in the direct method? Explain fully.

12. How is the amplitude of the cycles measured in the direct method?

13. How is the secular trend of cycles measured in the direct method?

14. How is the Pattern, or Contour, of business cycles measured in the direct method? Explain fully.

15. What are the advantages and disadvantages of the direct method over the residual method of measuring business cycles? Explain fully.

EXERCISES

1. Compute the business cycle by the residual method for the following data:

1. Price of hogs on the Chicago Market
2. Production of pig iron
3. Exports from the United States
4. Freight car loadings
5. Petroleum production

The above data may be found in the *Survey of Current Business*, Standard and Poor's, and in Moody's.

2. Compute the measurement of the business cycle for each of the following series of data by the direct method, using at least three complete cycles:

1. Production of pig iron
2. Production of bituminous coal
3. Freight car loadings
4. Factory employment
5. Total imports into the United States

The above data may be found in the *Survey of Current Business* Standard and Poor's, and in Moody's.

CHAPTER 18

INDEX NUMBERS

Index numbers are a specialized type of average. They measure the central tendency of a group of time series or spatial series. It is evident that one cannot judge the direction of the general movement of prices by one price series. For instance, the price of strawberries may rise at the same time that the prices of other fruits are falling. The change in strawberry prices may be caused by the weather, which may be unfavorable, or by the seasonal variation in prices. Likewise, the price of one particular stock or bond may fall while most other securities show rising prices. To measure the general movement of prices an average of several time series is necessary.

In measuring the volume of production the same principle holds. Fewer buttons may be produced at the same time that other production is rising. The decrease in buttons may be caused by an increase in zippers and not by depression. Prices and production may decline in some areas and increase in other localities or markets. In any case, a correct measure of the level of either price or production for either time periods or geographic subdivisions or areas can be obtained only by averaging the prices or quantities of a representative sample of the population to be measured. The principles of sampling, probability, error, averages, and variation apply to all index numbers.

It is a matter of common knowledge that economic activity is not uniform. Periods of prosperity and depression, of low prices and high prices, of idle factories and full employment chase each other through the years like sunshine and cloudy weather. Prices and production and employment and profits and rentals and

interest rates rarely remain at the same level for more than a few months at a time. Often they change from week to week. It is evident that an informed and intelligent management of business requires accurate and up-to-date information on these economic changes. Index numbers are the answer to a large part of this need. In fact, it is probable that the average person consumes or uses more index numbers than any other form of statistics. The business sections of all newspapers and all trade journals are full of them. They are the signs and guideposts along the business highway that indicate to the businessman how he should drive, or manage, his affairs.

PRICE AND PRODUCTION RELATIVES

The data of a single series when reduced to percentages are usually referred to as relatives. By contrast the average of two or more series is called an index. The production of wheat by years for any area, when the production figures are changed to percentages would be wheat production relatives. When the prices of wheat per month or year are changed to percentages they are called wheat price relatives. Some statisticians do not make this distinction between relatives and index numbers. One series cannot be averaged; several series can be combined only by an average. One series is much less likely to be a good measure of an economic activity or field than a fair sample of series. Because of these permanent and fundamental differences, we wish to distinguish clearly between a relative based on one series and an index based on several series. We shall reserve the term index number for the averaged series only.

The usefulness of reducing large production figures to simple percentages is quite evident. The ordinary person finds it very difficult to compare accurately and quickly figures running into hundreds of millions. The percentage figures are small and easy to understand and to compare.

In computing production relatives or price relatives, one period or area is selected as a base equal to 100, and the production or price of that period is divided into the prices or production of the

other periods or areas. The results are the percentages which we call relatives. This method is quite excellent for comparing changes in two or more series which differ a great deal in size, since it reduces their variations to a comparable base.

PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

The statistician must make seven decisions in the construction of an index number. They are:

WORKSHEET NO. 94

Year	Corn Production in U.S. in Bushels	Corn Production Relatives
1941	2,651,889,000	100.0
1942	3,068,562,000	115.7
1943	2,965,980,000	111.8
1944	3,088,110,000	116.4
1945	2,880,933,000	108.6
1946	3,249,950,000	122.6
1947	2,383,970,000	89.9
1948	3,650,548,000	137.7

Source: *Statistical Abstract of U.S.*, 1950

WORKSHEET NO. 95

Year	Corn Prices per Bushel (cents)	Corn Price Relatives
1941	75.1	100.0
1942	91.7	122.1
1943	112.0	149.1
1944	109.0	145.1
1945	127.0	169.1
1946	156.0	207.7
1947	216.0	287.6
1948	136.0	181.1

Source: *Statistical Abstract of U.S.*, 1950

1. Deciding what to measure;
2. Determining the periods or areas to be measured;
3. Selecting the series for the index;
4. Selecting the base;
5. Determining the type of average;
6. Selecting the weights;
7. Testing the reliability of the index.

These decisions must usually be made in the order in which they are here listed. However, they are not entirely independent of each other and a decision on one of them necessarily affects some or all of the others. The making of an accurate index number, therefore, requires a careful preliminary study of the field to be measured in all its phases.

Deciding What to Measure

Most economic activities are so varied and complex that it is very difficult or even impossible to measure them as a whole with any large degree of accuracy or meaning. Our analysis of trend, seasonal variation, and cycle illustrates this point. If we are to make an index that is dependable, should we eliminate trend and measure the combined seasonal and cycle, or shall we also eliminate seasonal and create an index measuring only the cycle? We may do either or both. If we do not eliminate the trend, we may find our index a mixture of opposite or contradictory trends. If we do not eliminate seasonal variation, we may find that our data are selected in such a way that they overemphasize the month-to-month changes of business. The question, shall we measure trend, seasonal variation, cycle, or two of them or all, remains and must be decided before we can begin to build our index. Index numbers may be made for any specific market, field of production, type of employment, or other business activity, but in any case the exact field that is to be measured must be predetermined. In other words, we must have a definite and delimited goal for our index.

Determining Periods or Areas to be Measured

Index numbers may be made by decades, such as 1800, 1810, 1820, etc., or by years, as 1930, 1931, 1932, etc., or by months or

weeks, or even days, or any other convenient time period. They may be made by states, counties, cities, townships, precincts, city blocks, or by nations or continents, or any other convenient geographical subdivision or area. In any case, the time period or area to be measured must be decided upon as a preliminary to collecting the data.

Index numbers for agricultural production are usually made by years; those for farm and industrial prices are made by months; industrial production and employment are usually measured by months although there are some weekly indexes in these fields; stock and bond quotations are given by days, weeks, and months. Population changes are often measured by decades.

Selecting the Series for the Index

After deciding upon the particular phase of business or other activity which is to be measured and the time periods or areas at or in which the measurement is to be made, the selection of the data is the next step. The data chosen necessarily depend on the preliminary decisions. It may be that sufficient data are not available to make the type of index desired. It may be desired to make a weekly index, but most of the data may be in monthly figures. Or a monthly index may be needed but the data are available only in quarterly statements. Such necessities may compel an alteration of the index.

Index numbers are usually based on purposive sampling instead of pure random selection. The sample should be of sufficient size to give dependable results. Since an index number is a specialized type of measure, a controlled stratified sample (see Chapter 13) is usually better. The appropriateness of including any particular series in an index may be checked by correlating it with other known characteristics of the variable to be measured. Professor C. E. Lively and R. B. Almack of Ohio State University in their study, *A Method of Determining Rural Social Sub-Areas with Applications to Ohio, Part I and Part II*,[1] employ and illustrate these methods.

[1] Ohio State University, Columbus, Ohio.

Selecting the Base

The base period or area of an index is the unit that is chosen to equal 100%. The values of the base period or area are divided into the values of the other periods or areas. The base value may be the price or quantity of a single time period, day, month, year, etc., or of a single unit area, as a county, state, city, etc., or it may be the average of several time periods or of many geographic subdivisions. In any case, it is better to have a base that is as near a middle or median value as possible. The highest or lowest value in the series is less desirable as a base than a middle one. In agricultural production an average of three to five or even more years is better than a single year because of the erratic fluctuation of farm production caused by irregular weather changes such as droughts, floods, frosts, and the like. In measuring population changes by counties in a state or region, the county mean of the state or region is a better base than the value for any one county. The same principle applies to all spatial indexes.

Determining the Type of Average

Any kind of average may be used in making index numbers, but the arithmetic mean and the geometric mean are more appropriate and more widely used than any other measures. The mode may be dismissed as too erratic. It will be recalled that the median is usually employed in indexes of seasonal variation. It may also be used for other types of indexes but with less reason than in time series in which one erratic month might seriously modify the entire index. If a sample of sufficient size of relatively homogeneous data can be obtained, the arithmetic mean is a dependable method for production and quantity indexes. The geometric mean is better for price changes and percentages. (See Chapter 6 on Averages.)

There are hundreds of formulas for computing index numbers. A few of these cover all important possibilities and are used almost universally. The others are only unimportant variations of the basic relations. The most widely used types of index numbers are the *aggregative* and the *average of price relatives.* Our computations are limited to these methods.

Selecting the Weights

All index numbers are weighted. If no conscious effort is made to assign to each series a special weight, they are all weighted equally with a weight of one. Such weighting, or lack of intelligent weighting, usually gives an inaccurate or unrepresentative index. The purpose of weighting is to make the index truly representative of the population it is to measure. A good illustration of the need of such weighting would be an index of food costs which placed meat, vegetables, salt, and pepper on an equal basis. In such a case if the price of pepper rose \$0.10 a can while bacon fell \$0.02 a pound, the index would show a rise in the cost of living, while actually there would be a decline because one would use many pounds of bacon for one can of pepper. Such an index would be incorrectly weighted. *An aggregative index should be weighted by the quantity used, bought, sold, or produced.* Quantity weights are usually easy to obtain and make an aggregative index fit the economic facts.

An average of price relatives index should be weighted with values. In computing price relatives the relative size of the values is removed and all are reduced to the common base of 100%. To make such an index realistic, these value differentials must be restored by including them in the weights. These two points the student must remember, (1) *Aggregative indexes have quantity weights;* (2) *Price relative indexes have value weights.* You will recall from your economics that Value = Price × Quantity; $V = PQ$.

INDEX NUMBER FORMULAS

At this point under the discussion of methods of averaging and weighting indexes, it is necessary to introduce the basic formulas.

Meaning of symbols:

I = Index V = value

p_0 = price of base year

p_1 = price of given year (any year except the base year)

q_0 = quantity of base year

q_1 = quantity of given year

FORMULA NO. 111

$$I = \frac{\Sigma p_1}{\Sigma p_0},$$

for an unweighted aggregative index of prices.

FORMULA NO. 112

$$I = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0},$$

for a weighted aggregative index of prices.

FORMULA NO. 113

$$I = \frac{\Sigma\left(\frac{p_1}{p_0}\right)}{N},$$

for an unweighted average of price relatives.

FORMULA NO. 114

$$I = \frac{\Sigma\left[\left(\frac{p_1}{p_0}\right) p_0 q_0\right]}{\Sigma p_0 q_0} = \frac{\Sigma\left[\left(\frac{p_1}{p_0}\right) V\right]}{\Sigma V},$$

for a weighted average of price relatives weighted with values.

FORMULA NO. 115

$$\text{Log } I = \frac{\Sigma\left[p_0 q_0 \text{ Log}\left(\frac{p_1}{p_0}\right)\right]}{\Sigma p_1 q_1},$$

for the weighted geometric average of price relatives index, using logarithms.

FORMULA NO. 116

$$I = \frac{\Sigma(q_1 p_0)}{\Sigma(q_0 p_0)},$$

for quantity indexes. Quantity is weighted by price.

TESTING THE RELIABILITY OF INDEX NUMBERS

As with all other statistics, or computations from samples, the question arises as to how dependable or representative the index is of the population it is supposed to measure. In Fig. 105 is shown the dispersion of the price relatives of the nine foods included in the Index of Cost of Foods, 1940, 1945, 1950. In the base period, 1940, all the relatives are, of course, 100. In 1945 all nine have risen in price, but five of them, beef, hams, milk, flour, and sugar, have all risen at about the same rate, while the four others, eggs, lard, potatoes, and coffee have risen at a much higher rate. This sample is, of course, too small to give a smooth curve or normal distribution of all food price changes. One hundred or even thirty or forty items would show a smoother frequency distribution.

In 1950 five of the nine items, potatoes, flour, milk, lard, and eggs form a central core within a close range, but two other foods, beef and hams, rise far above this central group and coffee because of abnormal monopoly conditions rises three times as high as the others. Only sugar falls below the core group. This somewhat erratic variation in price changes in this small sample indicates that to secure a good index a larger number of series should be used. Such a larger sample would tend to give a more normal distribution of price changes and to reduce the weight of such an erratic item as coffee in 1950.

An index number is only a specialized type of average and is subject to the same limitations as any other mean. Since the dependability of a mean rests on the representativeness of the sample from which it is computed, the standard error of the mean is as necessary a computation for index numbers as for any other statistic. If the sample is inadequate or unrepresentative of the population, any index based on it will be unreliable. In the making of index numbers, no other check is so important as the standard error of the mean.

The standard error of the mean of the 1945 price relatives is 10.5. Their mean is 163.1. The mean plus and minus the standard error is 163.1 ± 10.5 or $152.6 - 173.6$. But for the 1950 price relatives,

the standard error of the mean is 55.1. This large number is due to the wide change in monopoly coffee prices. Their mean is 280.7. The mean plus and minus the standard error is 280.7 ± 55.1 or 225.6 − 335.8, which is a very wide range. If coffee is eliminated from the sample, the mean for 1950 is 227.4 and its standard error is 16.2. This gives a range of 227.4 ± 16.2 or 211.2 − 243.6 which is not far from the 1945 range. These computations indicate how important an adequate random sample is in the making of index numbers. No other type of check is comparable to this one in importance. Graphic analysis indicates the probable results.

UNWEIGHTED AGGREGATIVE INDEXES

The aggregative index is made by adding the values for each time period, selecting a base, and dividing the base aggregate into the totals for each of the other time periods. In the so-called unweighted aggregative, each item has a weight of one regardless of its importance. The greatest defect of this index is that commodities sold in large units, such as hundredweights or tons, have a large dollar value per unit, and, therefore, exert an altogether disproportionate weight in the index. Milk, flour and potatoes sold

WORKSHEET NO. 96

UNWEIGHTED AGGREGATE INDEX OF COST OF FOOD

Items	Prices 1940	Prices 1945	Prices 1950
Eggs	.188	.372	.420
Beef	.170	.210	.475
Smoked Hams	.176	.258	.524
Lard	.069	.146	.158
Milk	2.210	3.260	4.572
Flour	2.200	3.180	5.212
Potatoes	1.782	3.158	3.384
Sugar	.050	.065	.078
Coffee	.072	.136	.509
Totals	6.917	10.785	15.332
Index	100.0	155.9	221.6

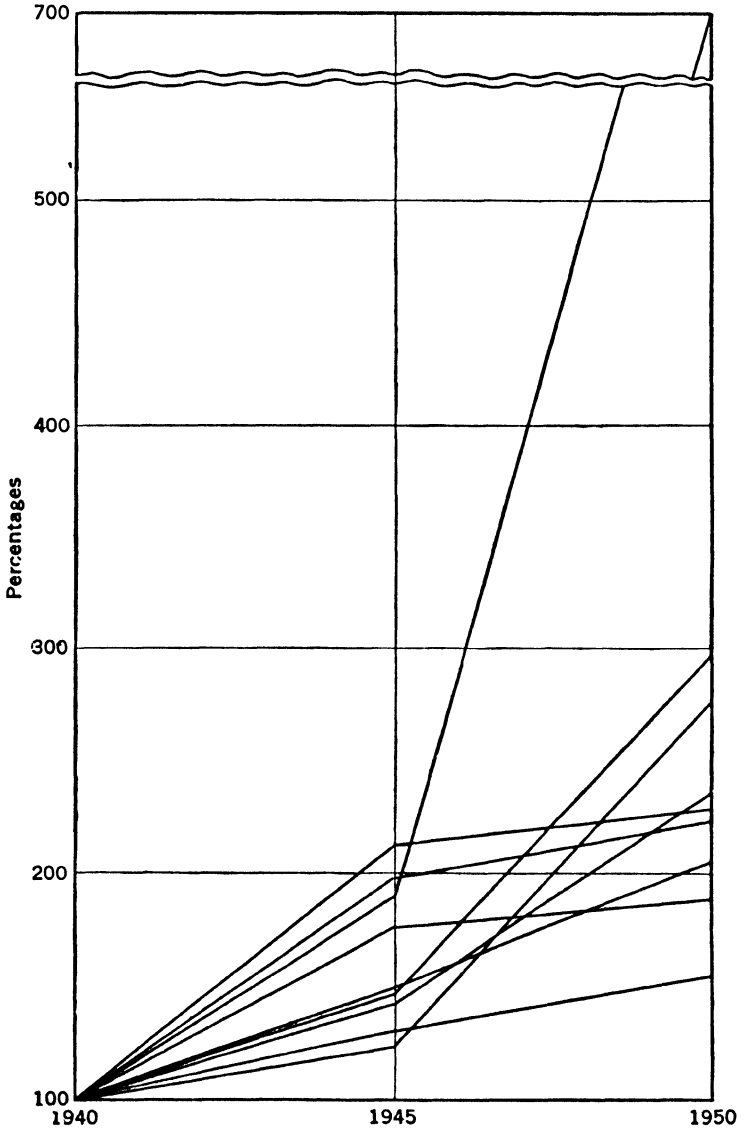


FIG. 105. Percentage changes in the prices of nine foods, 1940, 1945, and 1950. From Worksheet No. 99

Fig. 105 indicates that this sample is quite homogeneous, with the exception of coffee, which is monopoly controlled.

WORKSHEET NO. 97

Weighted Aggregative Index of Cost of Food

Items	1940			1945			1950		
	Price	Weight (in Billions)	Weighted Price	Price	Weight (in Billions)	Weighted Price	Price	Weight (in Billions)	Weighted Price
Eggs	$.188	4.66	.87608	$.372	4.66	1.73352	$.420	4.66	1.95720
Beef	.170	7.39	1.25630	.210	7.39	1.55190	.475	7.39	3.51025
Smoked Hams	.176	4.86	.85568	.258	4.86	1.25388	.524	4.86	2.54664
Lard	.069	1.76	.12144	.146	1.76	.25696	.158	1.76	.27808
Milk	2.210	1.19	2.62990	3.260	1.19	3.87940	4.572	1.19	5.44068
Flour	2.200	.26	.57200	3.180	.26	.82860	5.212	.26	1.35512
Potatoes	1.782	.26	.46332	3.158	.26	.82108	3.384	.26	.87984
Sugar	.050	13.04	.65200	.065	13.04	.84760	.078	13.04	1.01712
Coffee	.072	1.96	.14112	.136	1.96	.26656	.509	1.96	.99764
			7.56784			11.43950			17.98257
			100			151.2			237.6

by the 100 pounds dominate this index completely, each one of them having more than twice as much weight as all the other items combined. If these three foods were priced by the pound as the others are, there would be less bias in this index. The size of the unit in which the price is quoted is a significant factor and may be changed to either larger or smaller units.

WEIGHTED AGGREGATIVE INDEX

The statement that an index is unweighted merely means that all series in it are given equal weights, or are weighted by one unit. Such illogical weighting distorts the index. In the actual transactions of the market or in the production of goods, all commodities do not have equal weights. Steel is produced and sold in much larger quantities than lead or tin. The production of corn is three times that of wheat and many times that of barley or rice. Unless these actual differences are included in the diverse weights used in the making of index numbers, the results will be misleading. The differences between Worksheets No. 96 and No. 97 indicate something of the differences in results.

QUANTITY WEIGHTS

Since aggregative indexes should be weighted with quantities, the selection of accurate weights is an important part of index construction. Worksheet No. 98 indicates how weights may be

WORKSHEET NO. 98

THE ACTUAL AVERAGE QUANTITY WEIGHTS

Item	Weights	In Billions
Eggs, dozen	4,663,000,000	4.66
Beef, lbs.	7,391,522,400	7.39
Smoked Hams, lbs.	4,857,391,200	4.86
Lard, lbs.	1,764,696,000	1.76
Milk, 100 lbs.	1,192,104,000	1.19
Flour, 100 lbs.	262,312,800	.26
Potatoes, 100 lbs.	256,116,000	.26
Sugar, lbs.	13,036,286,400	13.04
Coffee, lbs.	1,960,012,800	1.96

Source: *Survey of Current Business*, 1950.

chosen. The quantities used in Worksheet No. 98 are computed from data in *The Survey of Current Business*, 1950. Since it is the relative size of the quantities that is essential to correct weighting, the large numerical quantities may be reduced to more manageable size by (1) either reducing them to percentages, or (2) dropping all but the significant left-hand digits, which is done in this case. The weights are reduced to units of billions to the second decimal place which is equivalent to tens of millions in whole numbers.

Value weights are used in a weighted average of price relatives index. Value is price times quantity. The quantities used may be those of the base period or of some other period. They may be an average of quantities extending over several periods. In agricultural production it is especially necessary to use the average quantity of the crop produced over a period of years, because of the fluctuations of annual yields. The value weights may be used as absolute figures in dollars.

Often a preferable method is to reduce the absolute dollar values to percentages as is done in Worksheet No. 101. This greatly facilitates the calculations if the index covers a large number of items and time periods.

When the weights and prices of the base year of an aggregative index are used as value weights in computing a weighted average of price relatives index, the two indexes are identical. Both are widely used in the construction of business indexes and both give fairly good results.

AVERAGE OF PRICE RELATIVES

A second method of computing index numbers is the average of price relatives. By this method the actual data of each series are changed to price relatives by dividing all the prices of a series by a base period price. These price relatives of the several series are then averaged to make the index.

In the average of price relatives method a base period is selected and the data for each item are changed to price relatives with the value of the base period as 100. The relatives for each time period are totaled and these totals divided by the number of items. This

unweighted index is not identical with the unweighted aggregative. In using the relatives of the data the absolute difference in the size of the units in which prices are quoted is eliminated. Although the unit of price for cotton lint is pounds and the unit for cottonseed is tons, and one price is in cents per pound and the other price is in dollars per ton, the relatives for the base year are identical, 100 for each item. The sizes of the relatives for the other time periods are determined only by the relative amount of change in the price of each item between the base period and the given period and not by the size of the price unit. While this device avoids one of the faults of the unweighted aggregative index, it still does not give a good index unless the relatives are weighted properly. Only a logically or scientifically weighted index can give an accurate measure of the population it is designed to represent.

In Worksheet No. 99, the prices of the commodities in 1940 are selected as the base to equal 100, and the relatives for the other time periods, 1945 and 1950, are obtained by dividing the price of each commodity for 1940 into the price of that commodity for

WORKSHEET NO. 99

UNWEIGHTED AVERAGE OF PRICE RELATIVES
INDEX OF COST OF FOOD

Item	1940		1945		1950	
	Price	Price Relative	Price	Price Relative	Price	Price Relative
Eggs	.188	100.0	.372	197.9	.420	223.4
Beef	.170	100.0	.210	123.5	.475	279.4
Hams	.176	100.0	.258	146.6	.524	297.7
Lard	.069	100.0	.146	211.6	.158	229.0
Milk	2.210	100.0	3.260	147.5	4.572	206.9
Flour	2.200	100.0	3.180	144.5	5.212	236.9
Potatoes	1.782	100.0	3.158	177.2	3.384	189.9
Sugar	.050	100.0	.065	130.0	.078	156.0
Coffee	.072	100.0	.136	188.9	.509	706.9
Totals		900.0		1467.7		2526.1
Index		100.0		163.1		280.7

each of the other time periods. The relative of 197.9 for eggs in 1945 is obtained by dividing .372, the price of eggs in 1945 by .188, the price of eggs in 1940, and in like manner for all other time periods, and for all commodities.

WORKSHEET NO. 100

VALUE WEIGHTS

FOR WEIGHTED PRICE RELATIVES INDEXES OF FOOD PRICES

Commodity	Value Weights [1] (billions of dollars)	Value Weights percentages
Eggs	.87608	11.6
Beef	1.25630	16.6
Hams	.85568	11.3
Lard	.12144	1.6
Milk	2.62990	34.7
Flour	.57200	7.6
Potatoes	.46332	6.1
Sugar	.65200	8.6
Coffee	.14112	1.9
	7.56784	100.0

It is not necessary to use the value weights of the base period as is done in Worksheet No. 101. In fact there are many cases in which it is better to use other weights because they are more truly representative of the population and periods than the values of the base period. Often it is better to use an average of two, three or more time periods than any one of them alone. Value weights as well as quantity weights should be as representative of the entire period covered by the index as possible.

[1] Taken from Column 3 (Weighted Price) in the base year, 1940 of Worksheet No. 97.

These are the values (Value = Price × Quantity) of these commodities in the base year, 1940.

WORKSHEET NO. 101

WEIGHTED ARITHMETIC MEAN OF PRICE RELATIVES INDEX OF COST OF FOOD

	1940				1945				1950			
	Data Prices	Price Relatives	Value Weights	Weighted Relatives	Data Prices	Price Relatives	Value Weights	Weighted Relatives	Data Prices	Price Relatives	Value Weights	Weighted Relatives
Eggs	.188	100.0	11.6	1160.0	.372	197.9	11.6	2295.64	.420	223.4	11.6	2591.44
Beef	.170	100.0	16.6	1660.0	.210	123.5	16.6	2050.10	.475	279.4	16.6	4638.04
Hams	.176	100.0	11.3	1130.0	.258	146.6	11.3	1656.58	.524	297.7	11.3	3364.01
Lard	.069	100.0	1.6	160.0	.146	211.6	1.6	338.56	.158	229.0	1.6	366.40
Milk	2.210	100.0	34.7	3470.0	3.260	147.5	34.7	5118.25	4.572	206.9	34.7	7179.43
Flour	2.200	100.0	7.6	760.0	3.180	144.5	7.6	1098.20	5.212	236.9	7.6	1800.44
Potatoes	1.782	100.0	6.1	610.0	3.156	177.2	6.1	1080.92	3.384	189.9	6.1	1158.39
Sugar	.050	100.0	8.6	860.0	.065	130.0	8.6	1118.00	.078	156.0	8.6	1341.60
Coffee	.072	100.0	1.9	190.0	.136	188.9	1.9	358.91	.509	706.9	1.9	1343.11
			100.0	10,000.0			100.0	15,115.16			100.0	23,782.86
				100.0				151.2				237.8

WORKSHEET NO. 102

WEIGHTED GEOMETRIC MEAN OF PRICE RELATIVES INDEX OF COSTS OF FOOD

Items	1940 Relative	1940 Log of Relatives	1940 Value Weights	1940 Weighted Logs	1945 Relative	1945 Log of Relatives	1945 Value Weights	1945 Weighted Logs	1950 Relative	1950 Log of Relatives	1950 Value Weights	1950 Weighted Logs
Eggs	100.0	2.00000	11.6	23.20000	197.9	2.29654	11.6	26.639864	223.4	2.34908	11.6	27.249328
Beef	100.0	2.00000	16.6	33.20000	123.5	2.09167	16.6	34.721722	279.4	2.44623	16.6	40.607418
Hams	100.0	2.00000	11.3	22.60000	146.6	2.16613	11.3	24.477269	297.7	2.47334	11.3	27.948742
Lard	100.0	2.00000	1.6	3.20000	211.6	2.32552	1.6	3.720832	229.0	2.35984	1.6	3.775744
Milk	100.0	2.00000	34.7	69.40000	147.5	2.16879	34.7	75.257013	206.9	2.31576	34.7	80.356872
Flour	100.0	2.00000	7.6	15.20000	144.5	2.15987	7.6	16.415012	236.9	2.37457	7.6	18.046732
Potatoes	100.0	2.00000	6.1	12.20000	177.2	2.24846	6.1	13.715606	189.8	2.27852	6.1	13.898972
Sugar	100.0	2.00000	8.6	17.20000	130.0	2.11394	8.6	18.179884	156.0	2.19312	8.6	18.860832
Coffee	100.0	2.00000	1.9	3.80000	188.9	2.27623	1.9	4.324837	706.9	2.84936	1.9	5.413784
			100	200.0000			100.0	217.452039			100.0	236.158424
				2.00000				2.174533				2.361584
				100.0				149.5				229.9

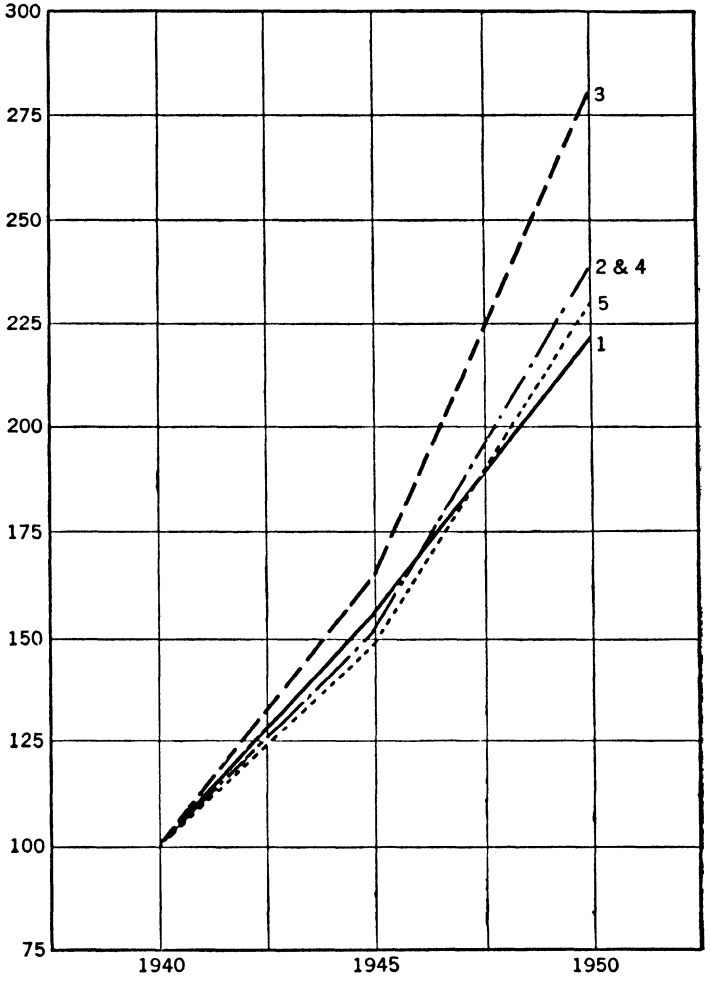


FIG. 106. Comparison of five index numbers. 1. Unweighted aggregative, 2. Weighted aggregative, 3. Unweighted average of price relatives, 4. Weighted average of price relatives, 5. Weighted geometric average of price relatives

Use of Geometric Mean in Making Index Numbers

The only difference between the methods of the arithmetic mean and the geometric mean is that in using the geometric mean, instead of averaging the price relatives, one weights and averages the logarithms of the price relatives. This method reduces the effect of the larger relatives in the result. Since a price may never fall more than 100% but may rise many thousand percent, if a few of the prices included in the index have risen several hundred percent, the arithmetic mean will give these few very high prices such a large effect in the index that it will be too high to give a true picture of the general price movement. The geometric mean minimizes the effect of the few very high prices so as to give a more accurate measure of the general movement. Review the Geometric Mean in Chapter 6.

Changing Basis of Index Numbers

The base of any index number may be shifted from the original base period to any other period for which the index has been calculated. This change of base is computed by dividing the value for each period into the completed index by the value of the period which is desired as the new base. The process may be illustrated with the values of the geometric average of price relatives as computed in Worksheet No. 103 below.

WORKSHEET NO. 103

METHOD OF SHIFTING THE BASES OF INDEX NUMBERS

Year	Present Index Base 1940	New Index Base 1945	New Index Base 1950
1940	100.0	66.9	43.5
1945	149.5	100.0	65.2
1950	229.9	153.8	100.0

This is a very useful device in the comparison and manipulation of index numbers. It often occurs that one wishes to compare two or more index numbers which have been originally computed

with different bases. One index may have 1913 as a base, another may have 1926 and another 1930 or some other period. By the method indicated above, all these indexes may be shifted to the same base, say 1936. This shift makes possible their comparison without recalculating them from the beginning.

QUANTITY INDEX

If one wished to measure fluctuations in the mineral production of California, the period to be covered should be decided and a sample of mineral production series selected. Although the sample is small, the method is illustrated in Worksheet No. 104. It is decided to measure the fluctuations in California's mineral production in 1930, 1935, and 1940. The production of copper, gold, silver, and petroleum is selected for the sample. The units are for copper 1,000 lbs., for gold and silver 1,000 of ozs., and for oil 1,000 bbls. At once the student faces the problem of how to add such diverse units as pounds, ounces, and barrels. Some common denominator must be found. Value or dollars, or quantity of worth, is the most available measure. Price, or quantity in

WORKSHEET NO. 104

COMPUTATION OF MINERAL PRODUCTION INDEX FOR CALIFORNIA, 1930, 1935, 1940

Mineral	1930			1935			1940		
	Production	Price Weights	Value	Production	Price Weights	Value	Production	Price Weights	Value
Copper (1,000 lbs.)	570,897	$ .09	$ 51,381	278,519	$ .09	$ 25,067	574,533	$ .09	$ 51,708
Gold (1,000 ozs.)	450	35.00	15,750	870	35.00	30,450	1,444	35.00	50,540
Silver (1,000 ozs.)	1,434	0.775	1,111	1,065	0.775	825	2,225	0.775	1,724
Oil (1,000 bbls.)	227,392	1.40	318,260	207,832	1.40	290,965	223,881	1.40	313,433
Totals			$386,502			$347,307			$417,405
Index			100.0			89.9			108.0

(Source, *Statistical Abstract of U.S.* 1941. The depression year of 1935 fell to 89.9 from a base of 100.0 in 1930 but the war year of 1940 rose to 108.0. Since the price weights are held constant, this index measures quantity production variations. The variations of the several items are not uniform. Copper and silver production fall a great deal from 1930 to 1935, but gold rises and oil falls only a little. The average of all of them is a better indicator of production change for the state than any single series.)

dollars, weights each commodity according to its relative value in the market. These price weights may be chosen as of some particular time or may be the average of a period of time. In this case they are for the year of 1940. The quantity of each mineral produced in each period is multiplied by the price weight, which must be uniform for all periods, and the values (QP) are totaled. The total value for the base period, which is 1930 in this case, is divided into the total values for all the periods and expressed as percentages.

SOCIO-ECONOMIC INDEXES

Frequently it is desired to measure variations in the level of living or other socio-economic factors of a number of areas or groups. Worksheet No. 105 illustrates a widely used method for making such indexes. It is desired to measure the socio-economic levels of the farmers of Rhode Island by counties. Five series of data are selected for the sample; (1) number of farm automobiles, (2) number of farm motorcycles, (3) number of tractors on farms, (4) number of farm homes lighted with electricity and (5) number of farm telephones. Other series no doubt should be included to make an adequate index, but these five farm conveniences will illustrate the principle just as well. All five of these items contribute to the productive power, pleasure, or satisfaction of farm families, and therefore, tend to measure their socio-economic status.

At once the student faces the difficulty of variation in the size of counties and of variation in the amount of farm lands per county. If the index is to measure accurately the level of living of farmers in counties of various areas and numbers of farms such data as these must be reduced to some satisfactory common denominator. A common denominator widely used in such cases is per-capita figures, or the population of the area. Since in this case we are dealing with farms, urban population must be excluded and only farm population used. Such a common base of comparison would be adequate in this case.

Another and perhaps a better measure would be the average number of each article or service per farm. These figures are

shown in the right-hand column under each county in Worksheet No. 105, and are obtained by dividing the original data for each series for each county by the number of farms in that county. These ratios are summed and their county average for the state used as a base.

$$2.510 + 2.841 + 3.100 + 2.711 + 2.901 = 14.063.$$

$14.063 \div 5 = 2.8126$ = county average for the state. This is used as the base.

Bristol County has the lowest level and Newport County the highest. Providence County is slightly below the state average and Kent and Washington Counties slightly above the average.

WORKSHEET NO. 105

COMPUTATION OF SOCIO-ECONOMIC INDEX FOR FARMERS IN RHODE ISLAND, 1940

Accommodations	Counties									
	Bristol		Kent		Newport		Providence		Washington	
	Total No.	No. per Farm	Total No.	No. per Farm	Total No.	No. per Farm	Total No.	No. per Farm	Total No.	No. per Farm
Automobile	96	.466	342	.753	343	727	942	.703	416	.766
Motorcycles	84	.408	233	513	252	534	600	.448	284	.523
Tractors	40	.194	123	.271	183	.388	338	.252	186	.343
Dwellings Electrically Lighted	182	.893	387	.852	415	.879	1,114	.832	413	.761
Telephones	113	.549	205	.452	270	.572	683	.476	276	.508
No. of Farms	206		454		472		1,339		543	
Totals		2 510		2.841		3.100		2 711		2.901
Index		89 3		101 0		110 2		96 4		103 1

Source: *U. S. Census of Agriculture*, Vol. 1, Part 1, Table X, 1940.

TESTS FOR INDEX NUMBERS

Three rather complicated tests have been devised to check the dependability of index numbers. They are (1) the time reversal test, (2) the factor reversal test, and (3) the circular test. There is some criticism of the adequacy and value of these tests. They are not necessary to the elementary student's understanding of index numbers and are therefore omitted from this text. The student who may wish to investigate them will find adequate treatments in the references at the end of this chapter.

SUMMARY

1. Index numbers are specialized types of averages.
2. The making of index numbers requires seven decisions as follows:
a. What to measure.
b. The periods or areas to be measured.
c. The series of data to be included in the index.
d. Selecting the base.
e. Determining the type of average to be used.
f. Selecting the weights for the several series.
g. Testing the reliability of the index.
3. Index numbers may be computed to measure:
a. Changes from one time period to another.
b. Changes from one area to another.
4. The base period or area should be an average or normal period or area. The mean of the periods covered or the mean of the areas covered may be used as the base.
5. In time series indexes it is preferable to have the base as near the present time as is possible to obtain a normal or average value.
6. Aggregative indexes are weighted by *quantities*.
7. Averages of price relatives indexes are weighted by *values*. Value is price time quantity.
8. The base of an index number may be changed or switched to another period or area by dividing the index through by the value of the particular period or area which one wishes to serve as the new base.
9. Index numbers are usually not based on general random sampling but on stratified purposive sampling.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapters XIX, XX. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapters XX, XXI. Prentice-Hall, Inc., New York, 1948.

Fisher, Irving, *The Making of Index Numbers*. Houghton Mifflin Company, Boston, Massachusetts, 1922.

King, W. I., *Index Numbers Elucidated*. Longmans, Green & Company, New York, 1930.

Mills, F. C., *Statistical Methods, Revised*, Chapters VI, IX. Henry Holt & Company, New York, 1938.

Persons, W. M., *The Construction of Index Numbers*. Houghton Mifflin Company, Boston, Massachusetts, 1928.

Riggleman, J. R., and Frisbee, I. N., *Business Statistics*, Chapter IX. McGraw-Hill Book Company, New York, 1951.

STUDY QUESTIONS

1. Distinguish between index numbers and relatives. What reasons are there for this distinction?
2. What problems must be decided before an index number can be made?
3. What are the chief problems in deciding what to measure? Explain in detail.
4. What are the difficulties in deciding on the time units for which measurements shall be made?
5. What rules should be followed in selecting the data for an index number? Explain fully.
6. What period or area should be chosen as the base period or base area? Explain fully.
7. What are the relative advantages of the various types of averages as applied to index numbers? Which would you advise using and why?
8. What is meant by weighting an index number? Can an index number be unweighted? Why?
9. For what areas or periods should weights be chosen? Why?
10. What kind of weights does an aggregative index require? Why?
11. What kind of weights does a weighted average of price relatives require? Why?
12. What kind of weights does a quantity or volume or production index require? Explain fully.
13. How may the dependability of an index number be tested? Explain in detail.
14. What is an aggregative index?
15. What is an average of relatives index?
16. How may the base period of a completed index number be shifted to another period or area?
17. Why is this shifting of index bases useful?
18. What is meant by the statement that the average person consumes more index numbers than any other kind of statistics?
19. What uses of indexes are made in marketing? Illustrate.
20. What indexes are used in personnel management?
21. How may indexes be used in budgeting?
22. Indicate the uses of index numbers in production.
23. Check 20 index numbers in the *Survey of Current Business*.
24. Check 20 index numbers in the *Federal Reserve Bulletin*.
25. Check the *Statistical Abstract of the United States* for index numbers.

EXERCISES [1]

ANNUAL AVERAGE PRICES OF BUILDING MATERIALS IN U.S. 1942–1950

Year	Southern Pine (per 1000 ft.)	White Lead (per 100 lbs.)	Steel (per ton) composit	Cement (per bbl.)	Brick (per 1000)
1942	28.90	$ 6.48	$53.00	1.67	$13.02
1943	34.42	6.50	53.00	1.67	13.43
1944	40.45	6.50	53.00	1.70	14.29
1945	41.63	6.50	54.60	1.77	15.89
1946	48.08	8.11	60.00	1.85	18.13
1947	69.97	14.67	68.40	2.06	19.87
1948	74.07	18.04	78.20	2.32	22.86
1949	81.14	23.29	84.92	2.41	23.95
1950	89.32	31.54	91.33	2.50	26.12

Quantities 1947

Weights	Southern Pine	5,000,000,000 ft.
	White Lead	360,000,000 lbs.
	Steel	88,000,000 tons
	Cement	190,000,000 bbls.
	Brick	10,000,000,000 bricks

Source: *Survey of Current Business*

From the above data compute the following index numbers:

1. Unweighted aggregative of actual prices.
2. Weighted aggregative of actual prices, weighted by *quantities*.
3. Unweighted arithmetic average of price relatives.
4. Weighted arithmetic average of price relatives, weighted by *values*.
5. Weighted geometric average of price relatives. Weight the logarithms of the relatives with the *value* weights.
6. After the index numbers are computed switch the bases of No. 2 and No. 4 to the last year used as a new base of 100.
7. Plot all five of the index numbers on millimeter graph paper with time on the X-axis.

[1] Any three to five years of the above data are sufficient for one laboratory exercise. The above data may thus be used for several semesters.

CHAPTER 19

PRACTICAL APPLICATION OF INDEX NUMBERS

Consumption of Statistics

Statistical Tables. The financial page of the average metropolitan newspaper each day contains from fifteen to thirty tables which analyze and display data on (1) New York Stock Exchange, (2) New York Curb Stocks, (3) Midwest Stocks, (4) New York Bond Prices, (5) Investment Trust, Banks, and Insurance, (6) National Livestock, (7) Chicago Livestock, (8) Kansas City Livestock, (9) Local Livestock, (10) New York Cotton Futures, (11) New Orleans Spot Cotton, (12) New York Cotton Seed Oil; (13) Chicago Grains, (14) Kansas City Grains, (15) Local Market Grains, (16) New York Poultry, (17) Arkansas Poultry, (18) New York Sugar, (19) New York Coffee, (20) Chicago Butter & Eggs, (21) New York Non-ferrous Metals, (22) Dow-Jones Averages, (23) New York Times Indexes, (24) Foreign Exchange Rates, and frequently from six to ten other tables. The consumption of statistics in the form of tables is approximately ten times as large as all other forms combined. This is the reason that this text spends so much space on the construction, meaning, and use of statistical tables.

The business section of most of the larger metropolitan papers contains from one to three small charts or graphs. The sports page of most metropolitan newspapers contains from ten to fifteen tables on current athletic and sports events, but carries a chart or graph very rarely.

Business magazines of the factual type run from ninety to ninety-five percent of their statistics in tables. "The Federal Reserve Bulletin" each month contains from 140 to 150 tables and from eight to ten charts and graphs. Sales, advertising, marketing, personnel, and management magazines carry a somewhat larger percentage of charts and graphs than the "Federal Reserve Bulletin" and the "Survey of Current Business"; but in them, also, tables predominate ten to one.

Many of the statistical tables in both the popular and the more factual business publications are composed *largely of index numbers* or other percentage figures. Frequently when the tables consist of raw data they are short numbers of from three to six digits and are read in 1000, in 1,000,000 or some other large unit which gives them the form and use of indexes of current business conditions. Hundreds, if not thousands, of such indexes, simple and composite, are employed by businessmen for each coefficient of correlation, determination, and regression computed. In sheer volume the most popular and widely consumed statistic, or computed number, in the business world, except totals and averages, is the index number. Some understanding of their computation, form, and use is a most essential part of a business education.

The world is literally full of index numbers. Many of them have proved their value through decades of use, and each year new ones appear. The largest single source of such measures is the Federal Government of the United States. The student at this point would do well to check briefly the sources of data and statistics supplied in Chapter 24 as to the vast extent of the government's contributions in the field of practical and applied statistics. Two of the largest, most convenient, and satisfactory of these sources are the "Federal Reserve Bulletin" published monthly by the Board of Governors of the Federal Reserve System at a cost of $3.00 a year or twenty cents a copy, and the "Survey of Current Business," published monthly by the United States Department of Commerce at a cost of $3.00 a year, or twenty-five cents a copy.

The most useful source on applied index numbers for the majority of firms is the "Survey of Current Business." It usually contains about twenty-five large pages of factual analysis of cur-

rent business problems. Volume 31, No. 1, January 1951, is typical with analyses of the following problems: (1) Consumer demand, (2) Finished Goods Output, (3) Controls, (4) Agricultural prices and income, (5) Parity revisions, (6) Business capital expenditures, (7) Investment programs, (8) Industrial trends, (9) National income and profits, (10) Gross national product in constant dollars, (11) War changes in production, (12) Current financial position of corporations, (13) Corporate holdings of liquid assets, (14) Bond yields vs. stock yields, (15) Private investments abroad, (16) New and revised statistical series, and many others.

The second part of the "Current Survey" consists of 40 pages of monthly business statistics composed of 282 separate index numbers and 2,287 statistical series of data which may be used as current indexes in the management of business whether the problems being considered fall in the fields of production, advertising, marketing, personnel, finance, budgeting, or general management.

The Survey of Current Business

1. General Business Indicators — Federal Reserve Index
    1. 72 specific indexes
    2. 158 additional series of data which serve as indexes
2. Business Population (firms)
    1. 35 separate series of data serving as indexes
3. Commodity prices
    1. Prices paid by farmers, 18 indexes
    2. Retail prices, 16 indexes
    3. Wholesale prices, 52 indexes
4. Construction and Real Estate
    1. Construction costs, 21 indexes
    2. Construction materials, 2 indexes
5. Domestic Trade — 57 Indexes
    1. Advertising, 31 separate series
    2. Postal services, 4 series
    3. Personal consumption, 19 separate series
    4 Retail trade, 92 separate series
    5. Wholesale trade, 6 series
6 Employment and Population
    1. Labor conditions, 90 series
    2. Wages, 144 separate series
7. Finance
    1. Banking, 58 separate series
    2. Consumer Credit, 25 series
    3. Federal Government Financing, 50 series
    4. Life Insurance, 42 series
    5. Monetary statistics, 22 series
    6. Profits and dividends, 13 series
    7. Securities issued, 51 series
    8. Commodity markets, 2 series
    9. Securities markets
        1. Bonds, 36 series
        2. Stock, 39 series
8. International Transactions of U.S.
    1. 12 indexes on foreign trade
    2. 126 separate series of data
9. Transportation and Communications
    1. Airlines, 6 series
    2. Express, 2 series
    3. Local freight, 3 series
    4. Class 1 Railroads, 31 series and 18 indexes
    5. Waterways, 5 series
    6 Travel, 10 series, 1 index
    7. Communications, 15 series
10. Chemicals and Allied Products
    1. Chemicals, 54 separate series
    2. Fertilizers, 14 series
    3. Naval stores, 6 series
    4. Miscellaneous, 4 series
    5. Fats, oils, by-products, 62 series
    6. Paints, 5 series
    7. Synthetic plastics and Resin products, 11 series
11. Electric Power and Gas
    1. Electric power, 19 series
    2. Gas, 18 series

12. Foodstuffs and Tobacco
    1. Alcoholic beverages, 23 series
    2. Dairy products, 28 series
    3. Fruits and vegetables, 9 series
    4. Grain and grain products, 61 series
    5. Livestock, 16 series
    6. Meats, 21 series
    7. Poultry and eggs, 8 series
    8. Miscellaneous foods, 28 series
    9. Tobacco, 18 series
13. Leather and leather products
    1. Hides and skins, 7 series
    2. Leather, 9 series
    3. Leather manufactures, 16 series
14. Lumber and Manufactures
    1. Lumber, all types, 11 series
    2. Soft woods, 29 series
    3. Hardwood floors, 10 series
15. Metals and Manufactures
    1. Iron and steel, 70 series
    2. Non-ferrous metals, 43 series
    3. Heating Appliances, 21 series
    4. Machinery, 10 series
    5. Electric Equipment, 14 series
16. Petroleum, Coal, and Products
    1. Coal, 30 separate series
    2. Coke, 9 separate series
    3. Petroleum and Products, 62 separate series
17. Pulp, Paper, and Printing, 76 series
18. Rubber and Products, 22 series
19. Stone, Clay, and Glass Products, 40 series
20. Textile products, 74 series
21. Transportation equipment
    1. Aircraft, 2 series
    2. Motor vehicles, 17 series
    3. Railway equipment, 25 series
    4. Industrial Electric Motors and Trucks, 3 series

The reason for developing the contributions of this particular source on current index numbers so fully is that it is the most complete and easiest available reliable source on business data in existence. There is scarcely a problem arising in current business management on which pertinent, reliable and decisive information is not found in the Survey of Current Business. Much of its material is republished in other Federal publications, and it carries material from other governmental sources such as the publications of the "Bureau of Labor Statistics," the "Federal Reserve Bulletin," and the "Departments of Agriculture" and the "Treasury." But in "The Survey" the information is consolidated in easily usable form. Many private statistical services use this source extensively.

Consolidated "Supplements" of the "Survey of Current Business" have been published in 1932, 1936, 1938, 1940, 1942, 1947, 1949 which bind in one volume the accumulated data of the preceding years making it more easily available for the computation of long run secular trends, indexes of seasonal variation, and detailed studies in cyclical fluctuations. Much of the data employed in the illustrations and problems of this text were taken from "The Survey of Current Business" and its "Supplements." In the solution of the many and varied problems of business management it is of superior value.

Most of its indexes are computed by the method of weighted

averages of price relatives. The bases most frequently used are the means of the period 1935–1939, 1910–1914, and 1926. Many of the weights are for 1939. As time passes, however, both the bases and the weights are revised and brought nearer to date. Fortunately, many of these indexes are designed so that they may be combined or compared with a maximum of ease and dependability. An abundance of footnotes on each page in the regular monthly issues and of longer explanations in the backs of the "Supplements" explain the nature of each index and its limitations and qualities. For the busy businessman who must have adequate answers quickly this source is highly recommended.

When economic problems require banking, monetary and financial indexes, and data, the "Federal Reserve Bulletin" is most valuable. It is also easily available and is dependable. Since Chapter 24 develops in fuller detail all the principal sources of statistical data supplied by the government, state as well as national, and of how to locate them, the remainder of this chapter will be devoted to some of the more widely used private services, and specialized governmental indexes.

THE INDEX OF WHOLESALE PRICES

Since 1890 The Bureau of Labor Statistics has supplied an index of wholesale prices. It has been expanded and modified frequently. The base has been shifted from 1913 to 1926, and a revised list of weights used based on the 1923–1925 period. At present it contains 887 series and is made up in "total" and "subdivisions." The principal subdivisions are (1) Farm products, (2) Foods, (3) Hides and leather products, (4) Textile products, (5) Fuel and lighting products, (6) Metals and metal products, (7) Building materials, (8) Chemicals and allied products, (9) House furnishing goods, (10) Raw materials, and (11) Manufactured products.

The original source of this index is the *Monthly Labor Review*, but it is also published in the *Federal Reserve Bulletin* and in "The Survey of Current Business." It is a widely used index and is employed to change current money values to constant dollar values, as is illustrated in Problem No. 1, page 518.

PROBLEM NO. 1

PROCESS OF DEFLATING DOLLAR VALUES BY DIVIDING THEM BY THE PRICE LEVEL INDEX. COMPUTATION OF REAL VALUE OF PERSONAL INCOME FROM 1939–1950 *

Year	Total Personal Income in Billions	Wholesale Price Index	Deflated Personal Income
1939	72.6	77.1	94.2
1940	78.3	78.6	99.6
1941	95.3	87.3	109.2
1942	122.7	98.8	124.2
1943	150.3	103.1	146.7
1944	165.9	104.0	159.5
1945	171.9	105.8	162.5
1946	177.7	121.1	146.7
1947	191.0	152.1	125.6
1948	209.5	165.1	126.9
1949	206.1	155.0	133.0
1950	223.2	161.5	138.2
If 1939 is taken as equal 100%, 1950 is only 146.7.			
1939	100.0		100.0
1950	307.4		146.7

* Source: *Federal Reserve Bulletin*, May, 1951, pp. 563–565.

This device is most useful in times of large inflation or deflation in reducing money sales to physical volume sales, and money wages to real wages and money profits to real profits. This computation is most essential in preparing physical budgets from dollar budgets.

INDEX OF COST OF LIVING INDEX

The Bureau of Labor Statistics has published a current cost of living index since 1913. It is based on food prices in 51 large cities, costs of other goods and services in 21 large cities and other data from 20 smaller cities. It contains 198 series, uses the 1935–1939 averages as a base and the following base period weights:

Foods	33.9
Clothing	10.5
Rent	18.1
Fuel, electricity and ice	6.4
Household furnishings	4.2
Miscellaneous	26.9
	100.0

The weights are changed slightly from time to time as the proportions and costs of family expenditures change.

FARM INDEXES

The Bureau of Agricultural Economics has supplied two farm indexes since 1910. They are (1) The prices farmers pay and (2) The prices farmers receive. They are used as the basis of government support to farmers' "parity" prices. The base period for the prices that farmers pay is the average of 1910–1914. The base for the prices that farmers receive is the average of the period, August, 1909 to July, 1914

PROBLEM NO. 2 *

Year	Prices Received	Prices Paid	Ratio of Recd. to Paid
1945	202	181	111.6
1946	233	203	114.8
1947	278	246	113.0
1948	287	264	108.7
1949	251	253	99.2
1950	256	261	98.1

* Source: *Oklahoma Current Farm Economics*, Vol. 24, No. 2, April, 1951.

FACTORY EMPLOYMENT AND PAY ROLLS INDEXES

Since 1919 the Bureau of Labor Statistics has supplied monthly indexes of Factory Employment and Pay Rolls which are published in the Federal Reserve Bulletin.

INDEXES OF STATISTICAL SERVICES

The following are some of the widely used indexes produced and sold to business firms by private statistical agencies.

1. Dun and Bradstreet's *Wholesale Food Index*, 1916 to date. It is based on a "market basket" of thirty-one foods in prices per pound and is expressed in dollars and cents instead of percentages.

2. The Guarantee Trust Company of New York compiles and publishes each month an index on business activity. It covers the period 1919 to date and is an excellent device for measuring general cyclical fluctuations.

3. Babson's Reports Inc. publishes a wide variety of statistical services including (1) Physical volume of business, (2) Bond prices, (3) Industrial stock prices, and (4) Industrial commodity prices.

4. Standard & Poor's Services and Publications cover the following areas:

(1) Corporation Records, six volumes
(2) Dividend Record covering 7,000 issues
(3) New Issues Service
(4) Listed Stock Reports, over 1200 individual reports on 1600 issues
(5) Bond Report, over 500 reports on 800 active issues
(6) Statistical Section, covers business, finance, stocks, bonds, commodity prices, employment, foreign trade, production by industries, inventories, and many other basic factors, all accumulated for several years
(7) Railroad Securities Service
(8) Bond Investment Service
(9) Register of Directors and Executives and many other services including hundreds of index numbers. Their Industry Surveys include 46 surveys of some 1000 companies.

5. *The Real Estate Analyst* published by Roy Wenzlick & Company of St. Louis carries scores of real estate indexes on all varieties of real estate activity and construction for national and regional areas.

6. *Moody's Investors' Service* publishes Moody's Industrials, Railroads, Public Utilities, Governments, and Banks-Insurance. Each of these consists of a semi-weekly publication which gives complete coverage of corporate factual data in the field named, including summaries of all financial reports, new issues, and the like. With each of these publications comes the well known *Moody's Manual*, covering financial background and descriptive material for past years. The company also publishes *Moody's Bond Record* (semi-monthly) and *Moody's Dividend Record* (semi-weekly).

The company also offers an extensive variety of investment advisory and supervisory facilities to both individuals and institutions. These include: Personal Management Service; Bank Supervisory Service; Pension Fund Advisory Service; and Investors' Advisory Service.

Moody's Stock Survey and *Moody's Bond Survey* are issued weekly for investors in these two fields. The organization also does tax valuation work, special industry surveys, and financial research projects of all types.

Other sources of information on index numbers and statistical services in this text are supplied in Chapters 14 and 24. The omission of many other well known and excellent sources of index numbers is due only to lack of space. These are given as representative of types of services and indexes rather than an exhaustive survey of the entire field.

Part Three
The Analysis of Small Samples

CHAPTER 20

THE ANALYSIS OF SMALL SAMPLES

The theory and methods presented in Part One, although they may be employed with relatively small samples, are designed primarily for large bodies of data and attain their highest degree of dependability in large samples. In samples of fewer than thirty items, the error is usually so large that it invalidates the dependability of the computations to a considerable degree. The reliability of samples tends to improve in proportion to the square root of their size. To obtain means, standard deviations, regression, and correlation coefficients that are quite dependable for purposes of forecasting, samples must be of sufficient size to permit only a small error.

In the field of the social sciences, economics, education, farm management, home economics, marketing, finance, population problems, sociology, government, taxation, history, psychology, labor problems, office practice, military strategy and operations, shipping, etc., etc., it is usually quite easy to obtain large samples. The various censuses, corporation and governmental reports, school reports, and the like, supply large volumes of data. In this field, information may be obtained by questionnaires in large quantity which may be quite dependable. In some cases large samples also may be obtained in the fields of the physical

and biological sciences. In the various treatments administered to fruit flies, mice, rats, rabbits, guinea pigs, horses, cattle, and even to human beings in many cases, very large quantities of data may be obtained and large sample methods used. Even in physics and chemistry, sometimes large volumes of data are available.

In many situations, however, where the data must be gathered by costly laboratory or field experiments that often require not only large expenditures of money but also long periods of preparation and development, it is impossible to obtain more than a few items of data in many weeks or even months and years. In all such cases it is quite necessary to have an adequate technique for the analysis of small samples. Even if it is possible to obtain large samples by the expenditure of considerable time and money, if smaller samples can be made to serve the purpose approximately as well, they are a very real economy and advantage. In fact, the techniques of analyzing small samples are so valuable and necessary that in recent years mathematical statisticians have given them much thought and carried them to a high degree of perfection. Such methods are so widely used today that it is advisable even in an elementary study to introduce the student to some of the more simple measures.

Making decisions on the basis of inadequate data is not a new experience in human affairs. Often a jury has to reach a verdict on the basis of limited circumstantial evidence. Every day of our lives we have to make important decisions in business, politics, social affairs, and war, with less than all the facts. The analysis of small samples, therefore, falls within the need of all of us.

WHAT IS A SMALL SAMPLE?

There is no clear demarcation line between large and small samples. The division is not a point or line, but rather a zone in which the one shades off gradually into the other. Some statisticians have arbitrarily chosen 30 as the division point, but 29 items are approximately as good as 31 and even 35 is not far removed from 30. The transition zone may well be considered as falling between 25 items and 50 items. One item added to 25 in-

creases the sample 4% but one item added to 50 increases the sample only 2%. Another reason for somewhat arbitrarily placing the upper edge of the small sample zone at 50 is that in small sample replications based on the Latin square the combinations 5×5, 6×6, and 7×7 are quite common. Larger Latin squares are used but are not as frequent.

PURPOSE OF SMALL SAMPLES

The main reasons for using small samples are *necessity* or *economy*. In the physical and biological sciences in which much of the data are obtained from elaborate time consuming laboratory or field experiments it is often impossible to amass a large volume of data. An entire summer or year may be required to obtain data on field crops, insects, pests, and blights. In the case of trees, several years may be required. For cattle, sheep, hogs, horses, and other animals several generations may be necessary for some problems. Certain controlled experiments in livestock feeding, in the use of new drugs on human beings, and in dietetics and metabolism yield only small quantities of data over a long period of time. If such problems had to be solved on the basis of large samples — hundreds or thousands of items — little progress could be made. In other cases fairly large samples ultimately could be assembled, but only at prohibitive expense. If the number of laboratory assistants, the amount of laboratory or field equipment, or raw materials can be reduced 50% to 90%, from twice to ten times as many studies can be made. In cases in which either necessity or excessive costs stand in the way of scientific advancement the small sample methods are of great value.

HOW SMALL SAMPLES ARE OBTAINED

In Chapter 13 four methods of sampling were explained. "Purposive" sampling and "Stratified Purposive" sampling were illustrated as applying to large bodies of data. In both of those cases certain *controls* were introduced to make a smaller sample more truly representative of its population. The superior con-

trolled, small "Gallup" and "Fortune" polls of only a few thousand items were compared with the much larger but inferior former "Literary Digest" polls of millions of items. In collecting small samples additional, much more exact and refined, controls are used which make it possible to measure certain characteristics of elusive populations with a large degree of accuracy from samples of only a few properly chosen items. In Chapter 23 the planning of a statistical study was emphasized. Although much careful forethought is essential to effective large sample studies, it is doubly necessary in small sample analysis. In many cases, small samples simply cannot be found or taken. They must be made.

Experimental Design

Satisfactory small sampling procedures must conform to two principles. First, *the particular and specific variation or variations which it is desired to measure must be separated from each other and from all residual variations.* Second, *an adequate basis for computing error must be established.* The most effective devices so far discovered for achieving these two ends are known as *experimental design.* The experiment which is to supply the desired data is planned and set up in a form which will segregate and reveal the particular type of variation which is to be measured. This method has been more fully developed in the fields of agriculture than elsewhere, although the principles have a wide application.

If it is desired to measure the difference in yield between two varieties of wheat, the experiment is set up to eliminate or to compute out all other causes of variations as far as possible. Wheat yields may vary because of differences in (1) soils, (2) fertilizers, (3) time of planting, (4) condition of seed bed, (5) moisture in soil, (6) amount of seed per square foot, (7) evenness of seed distribution, (8) quality of seed, and many other factors. To measure accurately real differences in yield due to *variety* all other factors must be held as near constant as possible. It is the same type of problem the physicist faces in the laboratory when he attempts to measure the relationship between the temperature and the pressure of a gas. The physicist must hold all other fac-

tors, including volume, equal. The statistician must design his experiment so as to eliminate all other variations or to segregate them and to compute them out of his results.

One side of the field may be poorer soil than the other, or drainage may be better, or exposure to wind or heat may be unequal, or many other similar factors may obscure the true results. A poor yielding variety on good soil might produce more than a good yielding variety on poor soil. The experiment must be designed to eliminate the effects of all such variations as between the two varieties of wheat.

Randomized Blocks or Rows

The field may be laid off in a number of blocks each of which contains the same number of rows as in Fig. 107. If, for instance,

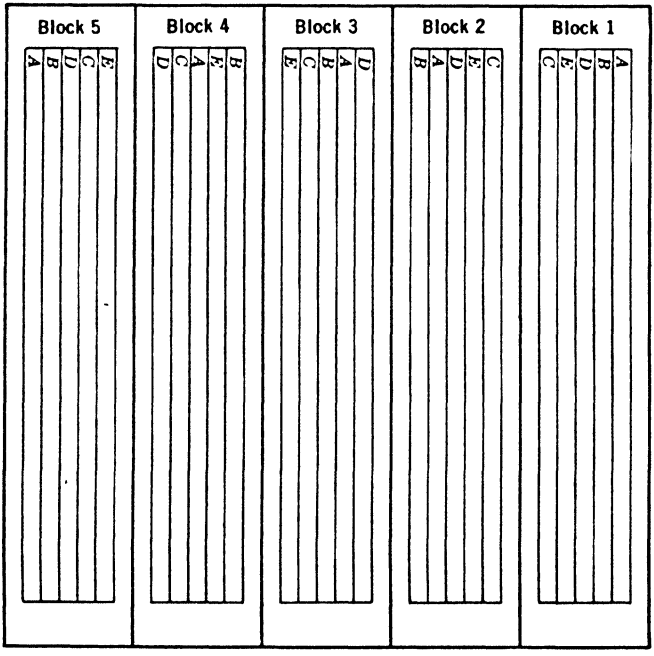


FIG. 107. Randomized blocks in field test. By this method the five varieties are distributed among the five blocks in such a random order that all varieties appear in all possible locations.

the experiment was designed to test differences in yield of five varieties of wheat, there would be five rows in each block, distributed in the several blocks in random order. If Block 1 was in the more favorable side of the field which continually grew poorer toward the opposite side including Block 5, variety *A* of wheat would be planted at random in all the various locations in the five blocks so that the advantages and disadvantages of location would be cancelled out; and variety *B* and all the others would have equal soil advantages with *A*. One limitation of the *row method* is that the field may also vary in soil or other respects diagonally so that there is an unequal variation within the rows from one end to the other. If the rows were very long, this might affect unequally the several varieties. This difficulty might be partially overcome by running the rows diagonally across the field or otherwise varying their contours, direction, or length.

The Latin Square

Under many conditions some form of the Latin square is a more effective control in experimental design. The Latin square is based on the relationship of n^2. If, for instance, four types of fertilizer were to be applied to one variety of cotton the square would be laid off 5×5 with as many rows as columns and with one block in each row for each kind of fertilizer and one check block with no fertilizer. If the four fertilizers were designated as *A*, *B*, *C*, and *D*, and the no fertilizer block as *E*, the Latin square could be set up as follows:

	Columns				
Rows	1	2	3	4	5
1	*A*	*C*	*B*	*D*	*E*
2	*D*	*E*	*C*	*A*	*B*
3	*B*	*A*	*D*	*E*	*C*
4	*E*	*B*	*A*	*C*	*D*
5	*C*	*D*	*E*	*B*	*A*

By this device the field or plot is divided up into as many rows as columns in order that each type of fertilizer may be applied to each section of the field so that whatever relationship or variation there might be between kind of fertilizer and soil condition would be equally distributed. All kinds of fertilizer would be applied on all types of soils.

This design could be applied to five types of fertilizer used with five varieties of cotton so that each variety of cotton would be planted once with each kind of fertilizer. If the five varieties of cotton were each indicated by an Arabic digit as 1, 2, 3, 4, and 5, the Latin square would be as follows:

	Columns				
Rows	1	2	3	4	5
1	A_1	B_2	C_3	D_4	E_5
2	B_4	C_5	D_1	E_2	A_3
3	C_2	D_3	E_4	A_5	B_1
4	D_5	E_1	A_2	B_3	C_4
5	E_3	A_4	B_5	C_1	D_2

This form of the Latin square is not desirable, as R. A. Fisher points out, because the letters fall in *too* uniform a pattern. The fact that all the *E*'s run in one diagonal line through the middle of the square violates the principle of *random* sampling which should be followed strictly. The *A*'s and the *D*'s are only a little less biased in location. All the letters and numbers should be scattered throughout the square in random order. This random design is necessary to establish a correct basis for random error. A strictly random order may be obtained for the several letters and numbers by shifting by chance the location of either the rows or columns or both. The Latin square on page 527 is made from the one above by shifting Row 5 in the square above to the position of Row 2, and shifting Row 2 in the square above to the position of Row 4, and shifting Row 4 to Row 5.

	Columns				
Rows	1	2	3	4	5
1	A_1	B_2	C_3	D_4	E_5
2	E_3	A_4	B_5	C_1	D_2
3	C_2	D_3	E_4	A_5	B_1
4	B_4	C_5	D_1	E_2	A_3
5	D_5	E_1	A_2	B_3	C_4

If Column 1 in this square changed position with Column 4, or Column 2 changed position with Column 5, a still better randomization would, perhaps, be obtained.

Latin squares may be set up for any value of n^2, but in actual practice the number of varieties and variables which it is convenient to include in one problem usually limits it to 4×4, 5×5, 6×6, 7×7, or at most 8×8.

Computation of Error

In Chapter 13 problems of sampling and error were considered. It was shown that standard errors are based on random sampling and pure-chance variation as expressed in the normal curve of error. It is, of course, likely that a small sample of from four to ten items, or even from 12 to 30 items will not describe a very smooth normal curve. The sample is *too* small. But the computation of a dependable measure of error still depends on the normal random distribution of the means in the form of the normal curve. The experimental design from which the small sample is obtained must be so managed that it is free from any bias or abnormality which would violate the assumptions underlying the normal distribution curve.

EFFECT OF SMALL SAMPLES ON STANDARD MEASURE

The fact that a sample is small and its error large need not invalidate it if we are able to measure that error with sufficient

accuracy and to make necessary adjustments for the error. An excellent example of this fact is the magnetic needle of the surveyor's or mariner's compass. The fact that the earth's magnetic poles do not coincide with the poles of the earth's axis prevents the magnetic needle from pointing directly north and south in most parts of the earth. But the fact that mariners and surveyors are able to compute accurately the amount of error enables them to steer their ships and run their lines as accurately as if there were no error in the compass. The same principle applies to all statistical computations. The increased error because of small samples need not invalidate our work if we can compute accurately the amount of adjustment necessary to compensate for it.

Our first problem is to determine how large the errors are for all the principal statistical measures, means, standard deviations, and coefficients of regression and correlation for various sizes of small samples. These relationships are illustrated and, perhaps, to some extent, demonstrated by the computations in the following table. Nine large samples of 105 items were taken from

TABLE 49

COMPARISON OF 9 LARGE AND 9 SMALL SAMPLES FROM STILLWATER, OKLAHOMA, GRADE SCHOOL CHILDREN

	9 Samples of 105				9 Samples of 26			
	Largest	—	Smallest	Range	Largest	—	Smallest	Range
$\overline{X}$	52.96	—	52.31	.65	53.38	—	51.81	1.54
$\sigma_{\bar{x}}$	.488	—	.439	.049	1.06	—	.83	.22
σ_x	5.00	—	4.48	.52	5.26	—	4.5	.81
σ_σ	.345	—	.315	.030	.737	—	.63	.107
$\overline{Y}$	66.00	—	64.29	1.71	70.23	—	62.54	7.69
$\sigma_{\bar{y}}$	1.99	—	1.57	.42	5.18	—	2.96	2.22
σ_y	20.28	—	15.12	5.16	21.25	—	12.71	8.54
σ_σ	1.403	—	1.042	.361	4.25	—	2.54	1.71
r	.907	—	.826	.081	.950	—	.749	.201
r^2	.823	—	.682	.141	.902	—	.562	.340
σ_r	.031	—	.017	.014	.088	—	.019	.069
b_{yx}	3.69	—	2.90	.79	4.51	—	2.59	1.92

RATIO OF RANGE OF 26 ITEM SAMPLES TO 105 ITEM SAMPLES

Measure	Range of Errors, Samples 105	Range of Errors, Samples 26	Ratio of Errors	Measure	Range of Errors, Samples 105	Range of Errors, Samples 26	Ratio of Errors
$\overline{X}$	.65	1.54	2.37	σ_y	5.16	8.54	1.66
$\sigma_{\overline{x}}$	.049	.22	4.49	σ_σ	.361	1.71	4.74
σ_x	.52	.81	1.56	r	.81	.201	2.46
σ_σ	.030	.107	3.57	r^2	.141	.340	2.34
$\overline{Y}$	1.71	7.69	4.50	σ_r	.014	.069	4.93
$\sigma_{\overline{y}}$	.42	2.22	5.30	b_{yx}	.79	1.92	2.43

the Stillwater, Oklahoma, grade school population, and nine smaller samples of 26 items each were also taken from the same parent population; and the arithmetic mean, standard error of the mean, standard deviation, standard error of the standard deviation, the coefficients of correlation and determination, the standard error of the coefficient of correlation and the regression coefficient were computed for both height and weight of children in both sets of samples.

It will be noted that in every case the range of error is larger for the 26 item samples than for the 105 item samples. The errors for the various measures of the smaller samples are from 1.56 to 4.93 times as large as the corresponding errors of the larger samples. This clearly demonstrates (1) that the larger the sample is the smaller is the error; (2) that larger samples are to be preferred to smaller ones if the labor and cost are not prohibitive; and (3) that if we can compute the error of a small sample with a high degree of accuracy, it may be used to great advantage in many cases.

The following generalizations may be made from the computations in the above tables:

1. The means of small samples tend to fall in about equal numbers on either side of the true mean, but the smaller the samples are, the more likely the means are to scatter over a wider range.

2. The standard deviations of small samples tend to be smaller than the true standard deviation of the population from which

they are drawn, and the smaller the samples, the smaller the standard deviation tends to be.

3. Since the standard deviations of small samples tend to be smaller than the parameter of the population, all errors based on these smaller standard deviations tend to be too small.

4. Unless corrections are made for the errors of small samples, they always tend to understate the actual error.

Fortunately there is available an accurate method of computing the increased error of small samples which is easy to apply. It may be stated briefly as follows:

Let M = the true mean of the population
$\bar{X}$ = the mean of a single sample
$\sigma_{\bar{x}}$ = the standard error of the mean

$$T = \frac{\bar{X} - M}{\sigma_{\bar{x}}},$$

or the difference between the sample mean and the population mean in terms of the standard error of the mean. When the standard error of the mean, $\sigma_{\bar{x}}$, is based on a large number of samples as is shown in Chapter 13, page 388, or is computed from the true standard deviation of the population, T is a correct and adequate measure of the deviation of $\bar{X}$ from M. But when $\sigma_{\bar{x}}$ is computed from one small sample, it is *not* a correct measure of this deviation. *It must be corrected for smallness of sample.* This correction is possible by the following method:

Let s = the standard deviation of the small sample
$s_{\bar{x}}$ = the standard error of the mean of the small sample
$x = \bar{X} - M$, the difference between the sample mean and the true population mean, and $t = \frac{x}{s_{\bar{x}}}$,

or the measure of the deviation between $\bar{X}$ and M in terms of the small standard error of the small sample mean. Then we may derive t, the measure for the small sample, from T, the measure for the large sample:

$$T = \frac{x}{\sigma_{\bar{x}}} \qquad t = \frac{x}{s_{\bar{x}}} \qquad T \div t = \frac{x}{\sigma_{\bar{x}}} \div \frac{x}{s_{\bar{x}}} = \frac{s_{\bar{x}}}{\sigma_{\bar{x}}}$$

The ratio of t to T is $\left(\frac{s_{\bar{x}}}{\sigma_{\bar{x}}}\right)$, but this ratio of the standard error of the small sample, $s_{\bar{x}}$, to the standard error of the true mean, $\sigma_{\bar{x}}$, *is known*, having been discovered and measured for all sizes of small samples by experiments and tests based on pure chance by the mathematical statisticians.

Therefore, $$t = \frac{x}{\sigma_{\bar{x}}} \div \frac{s_x}{\sigma_{\bar{x}}} = \frac{x}{s_{\bar{x}}}.$$

These corrections do not have to be computed by the student or statistician who uses them. For small samples of various sizes, they may be read from Table 50.

It is not possible to prove that there could never be an exception to a sample. The sample would have to be infinitely large to do this. It would have to include the entire population. Otherwise the chance would always remain that once in a thousand, or in ten thousand or in a million times there would be an exception. But it can be logically and safely inferred that if the differences are sufficiently few and far apart they may be accounted for by chance sampling and do not prove that the sample does not adequately represent the population. If the differences between two samples are sufficiently great, if they are so great that they cannot be accounted for by chance sampling within certain wide limits, it is safe to infer that the samples represent two different populations. From the statistical standpoint this method is made easier if an assumption of no difference between the samples is made and this assumption is proved false by showing that the difference *is so great* that it cannot logically be accounted for by chance sampling.

THE NULL HYPOTHESIS

The method of approaching this analysis is to set up a hypothesis of no difference, a *null hypothesis*. Suppose that an experiment has been set up to study the results of two methods of teaching reading in the second grade. A given number of pupils with the same intelligence quotients and cultural background, age, and other characteristics, would be chosen to participate in

TABLE 50[1]

TABLE OF *t*-VALUES FOR SMALL SAMPLES

n	$P = .9$	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.792	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

[1] Source: Table 36 is reprinted from Table 4 of R. A. Fisher, *Statistical Methods for Research Workers*, Sixth Edition, p. 166, Oliver and Boyd, Ltd., Edinburgh, 1936, by kind permission of the authors and publishers.

P = Probability

n = Number of degrees of freedom in computation of t.

All the values in the bottom line are *standard values* based on infinitely large samples.

the experiment. The pupils would be divided into two equal groups. For a given period, perhaps one term, each group would be taught to read by a method different from that used with the other group, and comparisons would be made between the two groups at the end of the term to determine which class had gained most in vocabulary and speed and comprehension in reading. The experimenter *would assume that there was no difference* in the results of the two methods, and then subject his assumption to proof. He would assume the null hypothesis and then proceed to disprove it, if he could. If the tests at the end of the term showed a large difference in student reading attainment, the *null hypothesis would have been discarded* and a superior method of teaching reading demonstrated. If the tests show only a small difference in student reading attainment, the *null hypothesis would have been sustained* and no superior method of teaching reading discovered. Any planned experiment would be set up on the same principle. If we wished to prove that keeping bright electric lights on hens all night would result in less hen time spent on the roost and more hen time spent in exercise on the floor eating and in egg production, we would begin by assuming the null hypothesis, or that such a procedure would make no difference in egg production. We would then select two equal groups of hens identical in all particulars, permitting one group to be in the dark at night while the other group was kept in the light. The comparative number of eggs produced by the end of a fixed period would determine whether our hypothesis of no difference was sustained or discarded.

DEGREES OF FREEDOM

Earlier in this chapter the fact that the standard deviation of small samples tends to be *too* small was presented. A device which will aid in correcting this bias in error is the practice of using $(N - 1)$ or $(N - m)$ for N in dividing the squared deviations from the mean. The explanation of this practice rests upon what the mathematicians call "degrees of freedom." This idea is based on the fact *that the sum of the deviations from a mean must always total zero* (0). This point may be demonstrated as follows:

$X - \bar{X} = x$
$9 - 10 = -1$
$12 - 10 = +2$
$8 - 10 = -2$
$7 - 10 = -3$
$13 - 10 = +3$
$11 - 10 = +1$
$60 \qquad 0$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{60}{6} = 10$$

$\Sigma x = 0$ (always)

Since the deviations from the arithmetic mean *must* equal zero, it follows that the freedom of variation of such deviations is limited. Since their sum must equal zero, one of the deviations is not free to vary, but must compensate algebraically for the value of the others. If in a distribution of six items, five deviations from the mean were $-7, +2, +1, +8, -9$, the other item MUST be a $+5$ to make the sum of deviations zero. It is not a question of which deviation is not free to vary. It is merely a fact that if all the others are free to vary, this particular one is *not* free. The degrees of freedom, as they are called, are one less than the number of items. For the sum of the squared variations of a single variable, the statement is $(N - 1)$. For cases in which there is included in a grand total the sum of the squared deviations of two or more variables, the formula requires $(N - M)$ in which M = number of variables. The principle of "degrees of freedom" in the sum of deviations from a mean is always correct theoretically, but it is of little practical significance in treating large samples. For instance, $(500 - 1)$ or 499, is little different from 500, but $(5 - 1)$, or 4, is twenty percent less than five. In small samples, therefore, it is very necessary to restrict our computations by "degrees of freedom," signified by (df), in order not to overstate our error.

TEST OF SIGNIFICANCE OF DIFFERENCES BETWEEN MEANS

The use of the table of t-values in the analysis of small samples, may be illustrated by taking samples of 6 items each from the first, second, and third grades of the heights of Stillwater, Oklahoma, school children. Up to this point it has been assumed that

the heights of the children of the first six grades composed a single population suitable for statistical analysis. In fact, up to this point, it has been subjected to a great deal of analysis, from means to regression and correlation coefficients. But now we wish to question the assumption that it is a single population and subject this assumption to proof. We wish to show that it is, in fact, six separate populations, as far as height is concerned — a separate population for each grade. We shall attempt this test by means of small samples of 6 items each from the first three grades. The same type of tests could be applied to all six grades.

WORKSHEET NO. 106

	Grade 1 Heights	*Grade 2 Heights*	*Grade 3 Heights*
	48	49	50
	43	48	52
	47	50	52
	44	48	54
	46	47	48
	43	46	51
Sums	271	288	307
$\bar{X}$	45.167	48.0	51.167
Σx^2	22.83	10.0	20.833
s^2	4 633	2.0	4.167
s	2.152	1.414	2.04
$s_{\bar{x}}$	.9624	.6324	.9123

In which:

Σx^2 = sum of squared deviations from means
s = standard deviations of small samples
s^2 = squared standard deviations
$s_{\bar{x}}$ = standard error of mean of small samples
s_{1-2} = standard deviation of differences of two samples

Then, by FORMULA NO. 117

$$s_{1-2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}} = \sqrt{\frac{22.83 + 10}{6 + 6 - 2}}$$

$$= \sqrt{\frac{32.83}{10}} = \sqrt{3.283} = 1.8$$

$$s_{1-2} = \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2[(N_1 - 1) + (N_2 - 1)]}}$$

This alternate Formula No. 118 may be used.

and FORMULA No. 119

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{1-2}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \frac{48.0 - 45.16}{1.8} \sqrt{\frac{36}{12}}$$

$$= \frac{2.84}{1.8} \sqrt{3} = 1.58 \times 1.732 = 2.737$$

If Formula No. 118 is used, also use Formula No. 120

$$t = \frac{X_1 - X_2}{s_{1-2}}$$

Turning to Table 50 where $n = 5$ and $P = .05$, we find $t = 2.571$. This figure, 2.571, means that for a sample of 6 ($N - 1$ is 5), the probability is 5%, or in only one case out of 20 that this difference could have occurred from chance. In other words, a $t = 2.571$ in a sample of 6 items means that the chances are 20 to 1 that the two small samples were drawn from separate parent populations. This means that the heights of first-grade children and second-grade children are two separate universes. If the t-value had been $P = .01$ for 5 degrees of freedom, or 4.032, it would have indicated that there was only 1 chance in 100 that this large difference between the sample means could have occurred from chance sampling. Statisticians designate 1 chance out of 20, or 95% probability, as *significant*. A 1 out of 100 chance or a 99% probability they call *highly significant*. In the problem above, $t = 2.737$, which is more than the *significant* limit of 2.571, but is less than the *highly significant* value of 4.032. Our t of 2.737 based on samples of 6 items is perhaps not conclusive proof that the two grades are separate populations, but it is a strong indication that this is the case. A sample of 10 items would no doubt yield conclusive proof. Here is a case in which two tiny samples of 6 items each have revealed significant information about the heights of school children.

If we wish to make the comparison between grades two and three, the computations are as follows:

$$s_{1-2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}} = \sqrt{\frac{20.833 + 10}{6 + 6 - 2}}$$

$$= \sqrt{\frac{30.833}{10}} = \sqrt{3.8033} = 1.76$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{1-2}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \frac{51.167 + 48}{1.76} \sqrt{\frac{36}{12}}$$

$$= \frac{3.167}{1.76} \sqrt{3} = 1.8 \times 1.732 = 3.118$$

Looking again at Table 50 where $n = 5$ and $P = .05$, we find $t = 2.571$ which is considerably smaller than the $t = 3.118$ in our problem. We conclude, therefore, that the difference in the heights of second-grade and third-grade school children is *significant*, but is not *highly significant*.

If, however, we wish to make a comparison between the first and third grades, we find that the difference is highly significant.

$$s_{1-2} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}} = \sqrt{\frac{22.567 + 20.833}{6 + 6 - 2}}$$

$$= \sqrt{\frac{43.400}{10}} = \sqrt{4.34} = 2.083$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{1-2}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \frac{51.167 - 45.167}{2.083} \sqrt{\frac{36}{12}}$$

$$= \frac{6}{2.083} \sqrt{3} = 2.89 \times 1.732 = 5.005$$

Checking again with Table 50, we find for $n = 5$ and $P = .05$ that $t = 2.571$. The t-value of 5.005 in the problem above is nearly twice this amount and is, therefore, *significant*. In fact, it is larger than the *highly significant* value of $t = 4.032$. We conclude, therefore, that the two small samples for the first grade and the third grade could *not* have been taken from the same parent population. We have learned from our three small samples that the grade school population is *not* a single homogeneous population but is a conglomerate of several populations. Samples from the heights of the pupils in the other grades would, no doubt, have revealed the same facts. The point in this analysis is not that this information could not have been obtained from large samples from each grade. Of course, it could have been learned from large samples and with greater certainty than from small

ones. But the point is that it was obtained from very small samples of 6 items each with the aid of Table 50. This illustrates the real economy of time and money that may be made from the careful analysis of small samples. It opens to statistical analysis the whole field of planned experimental research, which must often depend entirely on small samples. Although these small samples were taken at random they were not produced by experimental design.

TEST OF REGRESSION COEFFICIENTS

Small samples may be used not only to test the similarity and reliability of means to determine whether the samples were taken from the same parent population, but also to test the degree of similarity between regression coefficients. In connection with the small samples of 6 items used above to test the similarity of sample means, we may raise the question as to whether the relationship between height and weight is the same for these three grades. Does the weight of second-grade children increase in proportion to the increase of their height at approximately the same ratio that holds for first-grade children? Do they tend to grow faster in height than they do in weight? To what extent do the regression coefficients between height and weight for these three grades coincide? This test also can be made from small samples.

WORKSHEET NO. 107

	1st Grade		2d Grade		3rd Grade	
	X Height	Y Weight	X Height	Y Weight	X Height	Y Weight
	48	56	49	55	50	49
	43	48	48	50	52	62
	47	58	50	52	52	57
	44	46	48	49	54	59
	46	50	47	47	48	50
	43	46	46	49	51	58
Sums	271	304	288	302	307	335

X^2	Y^2	X^2	Y^2	X^2	Y^2
2,304	3,136	2,401	3,025	2,500	2,401
1,849	2,304	2,304	2,500	2,704	3,844
2,209	3,364	2,500	2,704	2,704	3,249
1,936	2,116	2,304	2,401	2,916	3,481
2,116	2,500	2,209	2,209	2,304	2,500
1,849	2,116	2,116	2,401	2,601	3,364
12,263.00	15,536.00	13,834	15,240.00	15,729.00	18,839.00
12,240.17	15,402.33	13,824	15,200.67	15,708.17	18,704.17
$\Sigma x_1^2$22.83	$\Sigma y_1^2$133.67	$\Sigma x_2^2$10	$\Sigma y_2^2$39.33	$\Sigma x_3^2$20.83	$\Sigma y_2^2$134.83

XY	XY	XY
2,688	2,695	2,450
2,064	2,400	3,224
2,726	2,600	2,964
2,024	2,352	3,186
2,300	2,209	2,400
1,978	2,254	2,958
14,780	14,510	17,182
13,730.67	14,496	17,140.83
$\Sigma x_1 y_1$49.33	$\Sigma x_2 y_2$14.0	$\Sigma x_3 y_3$41.16

FORMULA No. 120 (See Chapter 9)

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$$

FORMULA No. 121

$$t = \frac{b_1 - b_2}{\sigma_{b_1 - b_2}}$$

FORMULA No. 122

$$\sigma_{b_1 - b_2} = \sqrt{\frac{S_{y_1}^2}{\Sigma x_1^2} + \Sigma x_2^2}$$

FORMULA No. 123

$$S_{y_1-y_2}^2 = \frac{\Sigma(Y_1 - Y'_1)^2 + \Sigma(Y_2 - Y'_2)^2}{N_1 + N_2 - 4} = \frac{\Sigma(z_1)^2 + \Sigma(z_2)^2}{N_1 + N_2 - 4}$$

Computation of t for Regression Coefficients

Squared deviations and product deviations required for computations, taken from Worksheet No. 107.

Σx_1^2	22.83	Σy_1^2	133.67	$\Sigma x_1 y_1$	49.33
Σx_2^2	10.00	Σy_2^2	39.33	$\Sigma x_2 y_2$	14.00
Σx_3^2	20.83	Σy_3^2	134.83	$\Sigma x_3 y_3$	41.16

$$b_1 = \frac{\Sigma x_1 y_1}{\Sigma x_1^2} = \frac{49.33}{22.83} = 2.161 \qquad b_2 = \frac{\Sigma x_2 y_2}{\Sigma x_2^2} = \frac{14.00}{10.00} = 1.400$$

$$b_3 = \frac{\Sigma x_3 y_3}{x_3^2} = \frac{41.16}{20.83} = 1.976$$

Computation of Σz_1^2

FORMULA No. 124

$$\Sigma(z^2) = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$$

$$\Sigma(z_1^2) = 133.67 - \frac{(49.33)^2}{22.83} = 133.67 - \frac{2{,}433.4489}{22.83}$$
$$= 133.67 - 106.58 = 27.09$$

$$\Sigma(z_2^2) = 39.33 - \frac{(14.00)^2}{10.00} = 39.33 - \frac{196.00}{10.00}$$
$$= 39.33 - 19.60 = 19.73$$

$$\Sigma(z_3^2) = 134.83 - \frac{(41.16)^2}{20.83} = 134.83 - \frac{1{,}694.1456}{20.83}$$
$$= 134.83 - 81.33 = 53.50$$

Computations for S_y^2

$$S_{y_1-y_2}^2 = \frac{\Sigma(z_1^2) + \Sigma(z_2^2)}{N_1 + N_2 - 4}$$

$$S_{y_1-y_2}^2 = \frac{27.09 - 19.73}{12 - 4} = \frac{46.82}{8} = 5.85$$

$$S_{y_1-y_3}^2 = \frac{27.09 - 53.50}{12 - 4} = \frac{80.59}{8} = 10.07$$

$$S_{y_2-y_3}^2 = \frac{19.73 - 53.50}{12 - 4} = \frac{73.23}{8} = 9.15$$

$$\sigma_{b_1-b_2} = \sqrt{\frac{S_y^2}{\Sigma x_1^2} + \frac{S_y^2}{\Sigma x_2^2}} = \sqrt{\frac{5.85}{22.83} + \frac{5.85}{10.00}} = \sqrt{\frac{11.70}{32.83}}$$

$$= \sqrt{.3564} = .597$$

$$\sigma_{b_1+b_3} = \sqrt{\frac{S_y^2}{\Sigma x_1^2} + \frac{S_y^2}{\Sigma x_3^2}} = \sqrt{\frac{10.07}{22.83} + \frac{10.07}{20.83}} = \sqrt{\frac{20.14}{43.66}}$$

$$= \sqrt{.4613} = .680$$

$$\sigma_{b_2+b_3} = \sqrt{\frac{S_y^2}{\Sigma x_2^2} + \frac{S_y^2}{\Sigma x_3^2}} = \sqrt{\frac{9.15}{10.00} + \frac{9.15}{20.83}} = \sqrt{\frac{18.30}{30.83}}$$

$$= \sqrt{.5936} = .770$$

Completion of Computation of t

$$t = \frac{b_1 - b_2}{\sigma_{b_1-b_2}} = \frac{2.161 - 1.400}{.597} = \frac{.761}{.597} = 1.275$$

$$t = \frac{b_1 - b_3}{\sigma_{b_1-b_3}} = \frac{2.161 - 1.976}{.680} = \frac{.185}{.680} = .272$$

$$t = \frac{b_1 - b_2}{\sigma_{b_1-b_3}} = \frac{1.976 - 1.400}{.770} = \frac{.576}{.770} = .748$$

Checking the probability column, .05, in Table 50 for sample $N - 2$, or 4, we find $P = 2.774$. Our t-value for differences between b_1 and b_2 is only 1.275 which is less than one-half the requirement for a significant difference in regression lines. We conclude, therefore, that as far as this sample indicates there is no significant difference in the increase of the weights of children, as related to their heights between the first grade and the second grade. The t-values for the differences between the second and the third grades and between the first and the third grades are still smaller and are, therefore, not significant. This result indicates that the rate of increase in weight per inch of height is about the same regardless

of the age among these children. t shows only non-significant values for these regression lines.

Computation of the significance of difference between two regression lines for hyperbolic curves from Chapter 10, Worksheet No. 45.

$\Sigma(z_1^2) = .063887$ for the 1932–1937 problem
$\Sigma(z_2^2) = .182370$ for the 1941–1946 problem

$$S_{y_1-y_2}{}^2 = \frac{\Sigma z_1{}^2 + \Sigma z_2{}^2}{N + N - 4} = \frac{.063887 + .182370}{15 + 15 - 4} = \frac{.246257}{26.} = .0094714$$

$$\sigma_{b_1-b_2} = \sqrt{\frac{S_y{}^2}{\Sigma(x_1{}^2)} + \frac{S_y{}^2}{\Sigma(x_2{}^2)}} = \sqrt{\frac{.0094714}{2.562501} + \frac{.0094714}{1.490799}}$$

$$= \sqrt{\frac{.0189428}{4.053300}} = \sqrt{.004673425} = .068356$$

$$t = \frac{b_1 - b_2}{\sigma_{b_1-b_2}} = \frac{.45674 - .32264}{.068365} = \frac{.13410}{.068365} = 1.9615$$

In checking in Table 50 for the $P = .05$ level of significance for a small sample of 26 degrees of freedom, we find in the line in which $n = 25$, the t-value of 2.060. Since the ratio of our problem is only 1.9615, we conclude that it is not significant. In other words, from these small samples we conclude that their two regression lines are not sufficiently different to be confident that they were drawn from separate populations. They probably came from the same population.

TEST OF CORRELATION COEFFICIENTS

Computing r for each of our small samples from the deviations in Worksheet No. 107 by Formula No. 54 in Chapter 9, we have:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{49.33}{\sqrt{22.83 \times 133.67}} = \frac{49.33}{\sqrt{3{,}051.6861}} = \frac{49.33}{55.24} = .893$$

$$r = \frac{14.0}{\sqrt{10.0 \times 39.33}} = \frac{14.0}{\sqrt{393.3}} = \frac{14.0}{19.84} = .706$$

$$r = \frac{41.16}{\sqrt{20.83 \times 134.83}} = \frac{41.16}{\sqrt{2{,}808.5089}} = \frac{41.16}{53.00} = .777$$

FORMULA NO. 125

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

$$t = \frac{.893\sqrt{6-2}}{\sqrt{1-.797}} = \frac{.893 \times 2}{\sqrt{.203}} = \frac{1.786}{.450} = 3.969$$

$$t = \frac{.706\sqrt{6-2}}{\sqrt{1-.498}} = \frac{.706 \times 2}{\sqrt{.502}} = \frac{1.412}{.708} = 1.994$$

$$t = \frac{.777\sqrt{6-2}}{\sqrt{1-.604}} = \frac{.777 \times 2}{\sqrt{.396}} = \frac{1.554}{.630} = 2.467$$

Table 50 gives a t-value of 2.776 for a sample of $(N - 2)$, or 4, and a P-value of .05. In our small sample for the first grade, t equals 3.969, which is a little larger than the significant value for the table. But it is not sufficiently large to be highly significant. The values for the two other samples are not significantly different from zero (0). This value is, however, considerably smaller than the $t = 4.604$ for $P = .01$. We may conclude, therefore, that these samples are not sufficient to establish the fact of significant correlation between height and weight in the first three grades. There is, of course, the possiblity that larger samples might do so.

t-TEST OF PERCENTAGE OCCURRENCES

FORMULA NO. 126

$$\sigma_{p_1-p_2} = \sqrt{\frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2}}$$

in which p = favorable occurrences and q = unfavorable occurrences. (See Chapter 13 on Probabilities.)

In the distribution of class grades for large groups, 8% of the grades are normally A's. In one freshman class, only 4% of the grades were A's. The question arises as to whether this small percentage is a significant departure from the standard. The solution is as follows on page 546.

$$\sigma_{p_1-p_2} = \sqrt{\frac{.08 \times .92}{100} + \frac{.04 \times .96}{25}} = \sqrt{\frac{.0736}{100} + \frac{.0384}{25}}$$

$$= \sqrt{.000736 - .001536} = \sqrt{.008} = .0283$$

$$t = \frac{p_1 - p_2}{\sigma_{p_1-p_2}} = \frac{.08 - .04}{.0283} = \frac{.04}{.0283} = 1.413$$

In Table 50 the value of $t = 2.060$ for a sample of 25; for $P = .05$. Even the t-value for a very large sample for $P = .05$ is 1.95996. We conclude, therefore, that the difference is not significant. This percentage test may be applied to large samples as well as small ones.

USE OF SIGNIFICANT DIFFERENCES

Careful scientific thinking is based on accurate measurements and comparisons. The t-test, or critical ratio as it is sometimes called, is quite useful in making accurate comparisons. In measuring the relative effectiveness of two poisons in killing boll weevil, of two methods of teaching reading, of two ways of training apprentice machinists or brick layers, of two approaches in salesmanship or in any other area in which samples expressed in numerical quantities can be obtained, this device is useful. The t-test measures well whether sample differences are so small that they may be disregarded or so large that they are highly important to clear thinking and practical management.

The t-test may also be used to determine how much variation in a product is permissible. The output of automatic and semi-automatic machines may vary considerably. The number of imperfect units such as five per hundred, two per hundred or one per hundred or less may be determined by the use of significant differences.

SUMMARY OF TESTS BETWEEN MEASURES OF SMALL SAMPLES

1. The amount of corrected error for small samples of various sizes up to 25 items is given in Table 50, known as *Table of t-Values*, or simply the *t-table*.

2. By the methods shown above the difference between only *two* samples at one time may be measured for only one statistic at a time.

3. The degree to which the difference between two sample means is significant may be measured.

4. The degree to which the difference between the *b*-values of the regression lines of two samples is significant may be measured.

5. The degree to which the *r*-values of the coefficient of correlation of two samples is significant may be measured.

6. The degree to which the difference between two percentages is significant may be measured.

This method is the most simple and limited in range of all the methods of testing the significance of the statistics of a small sample. Only two samples can be used at once. Separate tests are necessary for each statistic for each pair of samples. The method, however, may be easily and quickly applied to a few samples, but if there are a large number of pairs of samples the task becomes long and tedious.

CHI-SQUARE MEASURE OF VARIATION

The degree to which a given sample may vary from an accepted or theoretical standard is measured by Chi-Square. It is a useful device for checking hypotheses against observed frequencies. It may be used with large samples as well as small ones, but it can be applied to quite small series.

FORMULA NO. 127

$$\chi^2 = \frac{(X_1 - m)^2}{m} + \frac{(X_2 - m)^2}{m}$$

in which

m = expected occurrences
X_1 = 1st set of occurrences
X_2 = 2d set of occurrences

One might raise the question as to the sex distribution of grade-school children. In a normal population, the sex distribution is a

1 : 1 or 50–50 relationship. If we take a random sample of the sex of the children in 10 rooms by selecting all the boys and girls in each room whose names begin with the letter A, we have the following distribution:

WORKSHEET NO. 108

School Rooms

Sex	1	2	3	4	5	6	7	8	9	10	Totals
Boys	1	2	0	1	0	3	1	5	2	3	18
Girls	4	0	1	2	0	2	1	2	1	1	14
Totals	5	2	1	3	0	5	2	7	3	4	32

Various types of analysis may be applied to these small samples. A percentage comparison may be made.

	Number	*Percentage*
Boys	18	56.25
Girls	14	43.75
Total	32	100.00

If the same percentage analysis should be applied to the separate rooms, the fluctuations would be wide and irregular. The samples are too small. If, however, there were some definite and permanent relationship between names and sex, and the sex names were evenly distributed so that in each room there were two pupils whose names began with A and one was a boy and the other was a girl, so that in this respect all rooms were uniform, the analysis of one room would be sufficient to determine the sex ratio in all rooms. Such uniformity in sex-name relations does not exist, and only an average of several rooms would give a dependable result.

One might consider the average number of boys and girls per room, whose names began with the letter A.

$\frac{18}{10}$ boys = 1.8 boys per room

$\frac{14}{10}$ girls = 1.4 girls per room

Still another way of analyzing the data is to compare the results of these samples with an assumed or accepted theory of distribution. Since the sex of children seems to be determined by chance, and the chances are approximately equal that any child may be male or female, one may set up the proposition that the sex of a group of school children on the average should be 50% males and 50% females. Taking this as the parameter of the population, one may compare the sampled rooms with this normal distribution to determine whether these school rooms depart from the normal to a significant degree.

Solution of Chi-Square

$$\chi^2 = \frac{(X_1 - m)^2}{m} + \frac{(X_2 - m)^2}{m} = \frac{(18 - 16)^2}{16} + \frac{(14 - 16)^2}{16}$$

$$= \frac{2^2}{16} + \frac{2^2}{16} = \frac{4 + 4}{16} = \frac{8}{16} = .5$$

Since Chi-Square is the measure between a sample and an accepted theory or standard, the question arises as to how large χ^2 must be before it indicates a significant variation. Again, we are indebted to R. A. Fisher for Table 51, the χ^2 table, the limits of significant values for samples of various sizes.

It is evident from Table 51 that when $\chi^2 = 0$, $P = 1$, and when $\chi^2 = \infty$, $P = 0.00$. For all intermediate values of χ^2, P varies with n. In our sample of two groups of grade-school children, χ^2 is equal to .5. For it to be significant for this size of sample, it would have to be at least 3.841, the value indicated in Table 51. Since our χ^2-value is .5, less than one-seventh of the value required to indicate a significant difference, we conclude that the sex variation in our sample is caused only by chance sampling.

By means of Chi-Square it is possible to measure the degree to which a given frequency distribution varies from the normal expectation. If we take the frequency distribution of the heights of 106 grade-school children as given in Worksheet No. 109, and compare this rather uneven distribution with the normal curve of error by means of computing χ^2, we can determine whether it departs from normal to a significant degree.

TABLE 51 *

CHI-SQUARE VALUES

n	P = .99	.95	.50	.10	.05	.02	.01
1	.000157	.00393	.455	2.706	3.841	5.412	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	1.145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.342	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	27.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	5.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	33.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.652	41.566	44.314
26	12.198	15.379	25.336	35.563	38.885	42.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

* This table is reproduced through the courtesy of R. A. Fisher and his publishers, Oliver and Boyd, Edinburgh. The entries are taken from Table III, *Statistical Methods for Research Workers.*

This computation is shown in Worksheet No. 109. The student will notice that the first three classes of the original frequency distribution are combined in Worksheet No. 109 as one class, "Under 46.5"; and that the last four classes of the original frequency distribution are combined as one class, "Over 58.4." This consolidation of the two tail groups is made to prevent an over-magnification of the effects of a small absolute deviation, which may easily occur in small frequencies. For instance, when the last four classes of the original frequency are considered separately their χ^2 value is 3.978 but when combined the value is only .700. In the first case the over magnification of the small absolute deviations + 4.0, − 1.3, and + 0.1 resulted in an expansion of χ^2 by (3.978 − .700) or 3.278 which is more than one-fourth of the total of 12.251 by the combined method. Used separately, $\chi^2 = 17.778$. With the small end classes combined it is reduced to 12.251 which more nearly represents the actual deviation between the observed and theoretical frequencies.

WORKSHEET NO. 109

COMPUTATION OF χ^2 FOR TESTING GOODNESS OF FIT OF THE NORMAL CURVE OF ERROR TO THE DISTRIBUTION OF 106 STILLWATER, OKLAHOMA, GRADE-SCHOOL CHILDREN

1 Class Intervals	2 Observed Frequencies f_1	3 Theoretical Frequencies f_2	4 $(f_1 - f_2)$	5 $(f_1 - f_2)^2$	6 $\frac{(f_1 - f_2)^2}{f_2}$
Under 46.5	7	10.5	− 3.5	12.25	1.167
46.5–48.4	18	10.4	+ 7.6	57.76	5.554
48.5–50.4	13	14.4	− 1.4	1.96	.136
50.5–52.4	14	17.7	− 3.7	13.69	.773
52.5–54.4	23	17.5	+ 5.5	30.25	1.728
54.5–56.4	10	14.2	− 4.2	17.64	1.242
56.5–58.4	7	10.1	− 3.1	9.61	.951
Over 58.4	14	11.2	+ 2.8	7.84	.700
	106	106.0			12.251

FORMULA NO. 128

$$\chi^2 = \Sigma\left[\frac{(f_1 - f_2)^2}{f_2}\right] = 12.251$$

If we check in Table 37 for $(n - 3)$, 5, and $P = .01$, we find $\chi^2 = 15.086$, which is considerably larger than the figure 12.251 for our distribution. We conclude, therefore, that our frequency distribution does not differ significantly from a normal distribution. Such deviation from the normal curve, as it manifests, may be accounted for by chance sampling.

This device for checking the conformity of samples with accepted or standard hypotheses is of great value in agricultural, biological, chemical, and other laboratory sciences in which it is difficult or impossible to obtain large samples. χ^2 does not reveal a great deal of information about the data, but it does indicate whether the deviations from the hypothesis are sufficient to be significant for further study. In developing a new strain of plants or a new breed of animals, one must always determine whether the peculiar characteristics of the new order are indicative of a separate strain or simply caused by chance sampling. χ^2 may also be used in checking the conformity of economic, sociological, psychological, and other social science data.

THE CHI-SQUARE TEST OF HOMOGENEITY

In many fields of statistical study it is desirable to test the similarity, likeness, or homogeneity of a total group of data. This may be done by sub-dividing the total group into sub-classes according to some standard classification, as equality of purchasing power, class grades, courses failed, incomes, length of service, rank, or other such accepted classifications, or by arbitrarily dividing the total group into sub-classes to test their homogeneity. The failures of the four standard college classes in a university are so tested in Worksheet No. 110.

The question arises as to whether this student body is an homogeneous group as far as failures are concerned. To test the assumption that it is an homogeneous group, theoretical or ex-

pected failures and non-failures for each class are computed by applying the percentage of failures and non-failures for the total group to each class. The expected or theoretical frequencies for students not failing are computed by taking the percentage of

WORKSHEET NO. 110

COMPARISON OF OBSERVED AND THEORETICAL FREQUENCIES OF STUDENTS ENROLLED AND FAILURES AND NON-FAILURES BY FOUR COLLEGE CLASSES

1	2 Seniors	3 Juniors	4 Sopho-mores	5 Freshmen	6 Totals	7 Percent
1 Total Number of Students	634	708	1,030	1,543	3,915	100.0
2 Actual Number of Students without Failures (f_1)	487	286	262	246	1,281	32.7203
3. Theoretical Number of Students without Failures (f_2)	207	232	337	505	1,281	
4 $(f_1 - f_2)$	+ 280	+ 54	− 75	− 259	0	
5 $(f_1 - f_2)^2$	78,400	2,916	5,625	67,081		
6. $\frac{(f_1 - f_2)^2}{f_2}$	378 7	12 6	16.7	132 8		
7 Actual Number of Students with Failures (f_1)	147	422	768	1,297	2,634	
8. Theoretical Number of Students with Failures (f_2)	427	476	693	1,038	2,634	67 2797
9. $(f_1 - f_2)$	− 280	− 54	+ 75	+ 259	0	
10 $(f_1 - f_2)^2$	78,400	2,916	5 625	67 081		
11. $\frac{(f_1 - f_2)^2}{f_2}$	183.6	6.1	8.1	64.6	803 2	

those students, 1,281, as against 3,915, the total number of students. This percentage is 32.7203. By multiplying the total number in each class by this percentage, the theoretical frequencies of 207 seniors, 232 juniors, 337 sophomores, and 505 freshmen — who, if the group is homogeneous, are expected not

to fail — will be obtained. These figures are shown in line 3. In like manner, if the total number of students in each class is multiplied by the average percentage of failures, 67.2797, the expected or theoretical number of failures for each class is obtained. These figures appear in line 8.

The deviations between the actual, or f_1, and the theoretical, or f_2 -values are shown in lines 4 and 9. These deviations are squared and divided by the theoretical frequency, (f_2), of their class. The final results appear in lines 6 and 11 and together equal 803.2, the total χ^2 value.

Since $$\chi^2 = \frac{f_1 - f_2^2}{f_2},$$

$$\chi^2 = \frac{(280)^2}{207} + \frac{(54)^2}{232} + \frac{(75)^2}{337} + \frac{(259)^2}{505} + \frac{(280)^2}{427} + \frac{(54)^2}{476} + \frac{(75)^2}{693} + \frac{(259)^2}{1038} = 803.2$$

The value of χ^2 computed for these data is 803.2. According to *Table 51, Chi-Square Values,* the value of χ^2 for 3 degrees of freedom and $p = .01$ is 11.341. Since the value of χ^2 we obtained for college classes is so much larger, one must conclude that these four college classes do not form a homogeneous group. Their difference is *too* great to be accounted for by chance sampling. It is a *real* difference. Seniors and juniors are a more highly selected group than freshmen, and therefore are better adjusted to college work.

SUMMARY

1. A small sample may be arbitrarily designated as a sample containing less than 30 items although the demarcation between small and large samples is a zone rather than a point.

2. Small sample analyses are made necessary by the application of statistics to experimental laboratory sciences from which often only small samples can be obtained.

3. The mean of a small sample tends to be more erratic or less dependable than the mean of a larger sample.

4. The standard deviation of a small sample tends to be too small, and, therefore, to give a standard error of estimate that is too small.

5. The statistics computed from small samples become dependable when adequate adjustments are made for their tendency to be too small. The so called *t-table* gives the ratios of dependability of small samples of various sizes.

6. A statistical result is said to be *significant* when this result could have occurred only one time in twenty as the result of chance sampling.

7. A statistical result is said to be *highly significant* when the result could have occurred only one time in one hundred as the result of chance sampling.

8. The adjustments provided in the *t*-table of ratios may be applied to the means, standard deviations, regression coefficients, coefficients of correlation, percentages, and the difference between means of small samples. But *t*-table ratios can be applied to differences between only two statistics at one time.

9. Chi-square is a statistical device which enables one to measure the significance of the difference among more than two statistics at one time. It is more extensive and flexible than the methods resting on the *t*-table ratios.

10. Chi-square is useful in comparing data with an established or acceptable standard or theory, or comparing a given frequency distribution with the normal curve.

11. Chi-square is useful in measuring the homogeneity of a sample such as a marketing area, a population group, or other social, political, or economic body.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter XXIX. Irwin, Inc., Chicago, 1941.

Fisher, R. A., *Statistical Methods for Research Workers*, Chapters IV, V. Oliver and Boyd, Ltd., Edinburgh, 1925.

Mills, F. C., *Statistical Methods, Revised*, Chapter XVIII. Henry Holt & Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapter XXI. John Wiley & Sons, New York, 1942.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, Chapter VI. McGraw-Hill Book Company, New York, 1940.

Richardson, C. H., *An Introduction to Statistical Analysis*, Chapters X, XI. Harcourt, Brace & Company, New York, 1943.

Snedecor, G. W., *Statistical Methods*, Chapters III, IV. Collegiate Press, Inc., Ames, Iowa, 1938.

REVIEW QUESTIONS

1. What is a small sample? How many items must a sample contain before it ceases to be small?
2. What are the characteristics of the means of small samples? Why?
3. What are the characteristics of the standard deviations of small samples? Why?
4. What are the characteristics of the standard errors of small samples? Why?
5. What is the meaning and use of t?
6. How is t derived?
7. What statistical measures and relations may be tested by t?
8. What is Chi-square?
9. What are the uses of χ^2 and what are its limitations?
10. Why are small samples advantageous? Are they better than large samples?
11. In what fields are small samples most necessary?

EXERCISES

1. Population samples from six small frontier mining camps

	Camps					
	1	2	3	4	5	6
Males	21	12	32	17	46	29
Females	6	7	12	5	27	22

Compute χ^2.

2. Data on second generation of mixed white and red peas on eight vines

	1	2	3	4	5	6	7	8
Colored	29	32	22	41	19	31	37	40
White	7	7	6	11	5	8	9	10

Compute χ^2.

3.

Class Intervals	f
4– 9.9	2
10–15.9	5
16–21.9	4
22–27.9	18
28–33.9	7
34–39.9	12
40–45.9	1

Compute test of abnormality from normal curve.

CHAPTER 21

THE ANALYSIS OF VARIANCE

In Chapter 7, the theory and methods of computing the standard deviation were developed and explained. In later chapters we have seen that statisticians make wide use of the standard deviation in measures of correlation, error, variation, and regression. We shall now modify this measure for other uses. The standard deviation is the square root of the sum of the squared deviations between the mean and the individual items divided by the number of items. As a measure of variation and deviation, better results for many purposes may be obtained by using the square of the standard deviation.

$$\sigma x = \sqrt{\frac{\Sigma x^2}{N}} = \text{Standard deviation}$$

FORMULA NO. 129

$$\sigma x^2 = \frac{\Sigma x^2}{N} = V = \text{Variance}$$

Variance is the name given the square of the standard deviation by R. A. Fisher, who developed its theory and uses. Since the squared deviations from the mean are always positive in sign and may be easily manipulated mathematically and have come into general use as a measure of deviation and variation, it seems logical to retain them in their squared form throughout the comparisons and computations. V, or σ^2, is as easy to manipulate as σ and, as we shall see, has certain advantages of simplicity. In Chapters 13 and 20 we learned methods of determining whether two samples had been taken from the same parent popu-

lation, or were from separate populations. One of these measures was the test for significance in *t*-tables. Another method was that of Chi-Square. With these methods in mind, the student is prepared to consider a more adequate and complex method which will measure the significance of the difference between several means at once. This is the analysis of variance.

DETAILED EXPLANATION OF VARIANCE

The purpose of the following illustration is to hold the analysis of variance, as it were, under a microscope, so that the student may see clearly every detail.

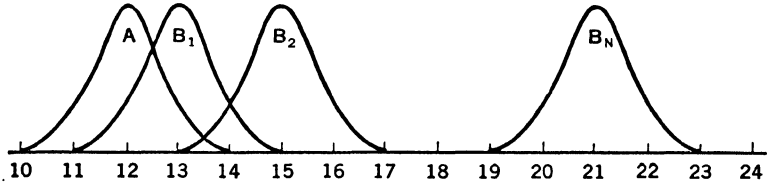


FIG. 108. Illustrating the divergence of samples in the analysis of variance

Assuming that we wish to determine whether two samples of corn are taken from two separate populations or are in reality only two samples from the same parent population, we must measure the divergence of their means in relation to their total deviation from their common mean. The distribution A extends from 10 to 14, inclusive, and the distribution of B_1 from 11 to 15, inclusive, on the X-scale. If B_1 also extended from 10 to 14, inclusive, its mean and range would be identical with the mean and range of A. Under this condition it would be evident that the two samples were identical and, therefore, could have been taken from the same parent population. The null hypothesis would be sustained.

Since B_1 is not identical with A, but is one unit to the right, it *might* be that this increase in size of B_1 is sufficient to indicate strongly that it came from a separate population. However, if the difference between A and B_1 is too small to be significant, it is certain that if B_1 were moved farther to the right, to the position of B_2, it would be *so* different from A that the difference would be

significant. If we continued to move B_1 farther still to the right, it would finally reach the position of B_n where the difference between A and B_n would clearly be significant. The purpose of the analysis of variance is not to determine the *exact point* on X where the difference between A and B is significant, but to determine *whether* the difference *is* significant at the point where B is located.

WORKSHEET NO. 111

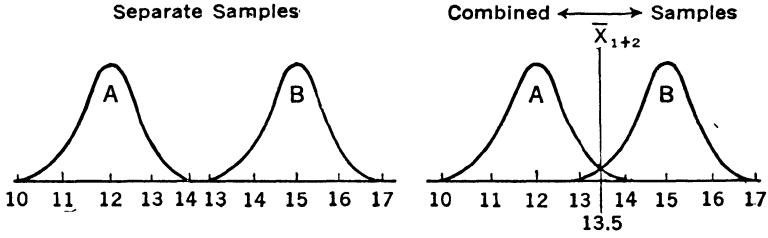


FIG. 109. Combination of two samples in analysis of variance

$X - \overline{X} = x$	x^2	$X - \overline{X} = x$	x^2	$X - \overline{X} = x$	x^2
10 − 12 = 2	4	13 − 15 = 2	4	10 − 13.5 = − 3.5	12.25
				11 − 13.5 = − 2.5	6.25
11 − 12 = 1	1	14 − 15 = 1	1	11 − 13.5 = − 2.5	6.25
				12 − 13.5 = − 1.5	2.25
				12 − 13.5 = − 1.5	2.25
11 − 12 = 1	1	14 − 15 = 1	1	12 − 13.5 = − 1.5	2.25
12 − 12 =		15 − 15 =		12 − 13.5 = − 1.5	2.25
12 − 12 =		15 − 15 =		13 − 13.5 = − 0.5	.25
				13 − 13.5 = − 0.5	.25
12 − 12 =		15 − 15 =		13 − 13.5 = − 0.5	.25
				14 − 13.5 = + 0.5	.25
12 − 12 =		15 − 15 =		14 − 13.5 = + 0.5	.25
				14 − 13.5 = + 0.5	.25
				15 − 13.5 = + 1.5	2.25
13 − 12 = 1	1	16 − 15 = 1	1	15 − 13.5 = + 1.5	2.25
				15 − 13.5 = + 1.5	2.25
13 − 12 = 1	1	16 − 15 = 1	1	15 − 13.5 = + 1.5	2.25
				16 − 13.5 = + 2.5	6.25
14 − 12 = 2	4	17 − 15 = 2	4	16 − 13.5 = + 2.5	6.25
				17 − 13.5 = + 3.5	12.25
10|120	12	10|150	12	20|270	69.00
12		15		13.5	

$\overline{X}_1 = 12$ $\overline{X}_2 = 15$ $\overline{X}_{1+2} = 13.5$

(1) The total sum of squares for the combined samples is 69.0

(2) $(\bar{X}_{1+2} - \bar{X}_1)^2 = (13.5 - 12.0)^2 = (-1.5)^2 = 2.25$
$(\bar{X}_{1+2} - \bar{X}_2)^2 = (13.5 - 15.0) = (1.5)^2 = 2.25$
2.25×10 (the number of items in 1st sample)
$= 22.5$ = sum of squares *between* 1st sample mean and total group mean.
2.25×10 (the number of items in 2nd sample)
$= 22.5$ = sum of squares *between* 2nd sample mean and total group mean.
$22.5 + 22.5 = 45.0$ = total of squared difference between the two sample means and the total group mean.
Sum of squares between samples = 45.0

(3) Total X^2 (squared deviations *within* 1st sample) = 12
Total X^2 (squared deviations *within* 2nd sample) = 12
Total sum of squares of deviations *within* the two samples = 24.0

Total Sum of Squares of combined samples = 69
Sum of Squares *between* samples = 45
Sum of Squares *within* samples = 24

	df	Sum of Squares	Mean Squares
Total	19	69	
Between	1	45	45.0
Within	18	24	1.333

$$F = \frac{\text{Variance between group averages}}{\text{Variance within groups}} = \frac{45.}{1.333} = 33.8$$

The total *df* is 19 because the combined samples contained 20 items (20 − 1) = 19. The degree of freedom between samples is 1 because there are 2 samples and (2 − 1) = 1.

The remaining 18 degrees, the difference between 19 and 1, or 18, are allotted to the individual items within the two samples. The proof that 18 is the correct figure is that each sample has 10 items, and

Samples	Total	*df*
A	10	$(N - 1) = 9$
B	10	$(N - 1) = 9$
		18

We, perhaps, now have sufficient insight into the methods of the analysis of variance to undertake the solution of a practical problem. It will be recalled that in Chapter 20 we questioned for the first time our original assumption that the heights of children in the first six grades were a single homogeneous population. Perhaps we are dealing with six populations instead of one. Let us now test this assumption more completely. We shall set up first the null hypothesis that there are no significant differences among the heights of children in the first six school grades. Second, from our school data we shall draw a small random sample of 8 items for each of the six grades and proceed to test the hypothesis of no significant differences in height.

WORKSHEET NO. 112

ANALYSIS OF VARIANCE OF HEIGHTS OF 48 CHILDREN DISTRIBUTED EQUALLY AMONG THE FIRST SIX GRADES OF STILLWATER SCHOOLS

Child	Grades						Totals
	1	2	3	4	5	6	
1	46	47	51	51	56	60	311
2	48	50	54	55	57	59	323
3	46	47	51	52	59	60	315
4	43	48	48	57	56	60	312
5	44	50	49	54	54	61	312
6	50	49	55	51	59	59	323
7	45	48	52	53	56	59	313
8	41	49	53	54	54	56	307
Totals	363	388	413	427	451	474	2,516

1st Step. Computation of Total Variance for 6 Grades

FORMULA NO. 51 $\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$ See Chapter 9.

$\Sigma Y^2 = (46)^2 + (48)^2 + (46)^2 + (43)^2 + \cdots + (56)^2 = 133{,}188$

$\Sigma Y = 2{,}516$

$\frac{(\Sigma Y)^2}{N} = \frac{(2{,}516)^2}{48} = \frac{6{,}330{,}256}{48} = 131{,}880.33 = \text{correction}$

$\Sigma y^2 = 133{,}188 - 131{,}880.33 = 1{,}307.67$

2d Step. Computation of Variance between the means of Grades

The sum of (Grade Totals)^2 = $(363)^2 + (388)^2 + \cdots + (474)^2$
= 1,063,288.

This total is divided by (f), 8, the number of items in each grade sample, in order to give the squares of the class means, each weighted by the number of observations upon which it is based.

FORMULA NO. 130

$$\frac{\Sigma[(\Sigma Y_1)^2 + (\Sigma Y_2)^2 + \cdots + (\Sigma Y_n)^2]}{f} - \frac{(\Sigma Y)^2}{N} = \Sigma y_c^2$$

In which

(ΣY_1), (ΣY_2), (ΣY_n) = Class totals
(ΣY) = Grand total of all classes
Σy_c^2 = Sum of squared deviations of class means from common mean
f = Number of items in a class

$$\frac{1{,}063{,}288}{8} = 132{,}911$$

132,911 − 131,880.33 = 1,030.67 = Total squared deviations between classes.

COMPUTATION OF F, TEST OF SIGNIFICANCE OF DIFFERENCES AMONG GRADE MEANS

	Degrees of Freedom	Sum of Squares	Mean Squares
Total	47	1,307.67	
Between Grades	5	1,030.67	206.16
Within Grades	42	277.00	6.6

$$F = \frac{206.16}{6.6} = 31.24$$

This worksheet shows the analysis of variance in its simplest form which is very easy for even the beginning student to compute. In Table 52 we may locate the F-value for 42 and 5 degrees

of freedom and compare it with our F of 31.24, which proves to be highly significant.

The method of locating the F-value desired in Table 52 is to look across the top of the table along n, till one arrives at the column of figures directly under the number of degrees of freedom in the *between classes* variable $(N - 1)$. Then follow down that column of F-values until one is opposite the number in the left-hand column of n^2, which is the number of degrees of freedom in the *within classes* variables. For our problem of classes of school children, $n_1 = 5$; that is, the fifth column on the top of Table 38, $n_2 = 42$, which is located on the left-hand side of the table. The top figure in light type, 2.44, is the 5% level of significance. The lower bold-face type figure of 3.49 is the 1% level of highly significant value. Our F-value of 31.24 is more than nine times this highly significant value. This proves that these 6 samples are taken from 6 separate populations. We must conclude, therefore, that as far as height is concerned, each grade of the 6 grades of the common schools is a separate universe. The amount of height a child gains in a year definitely separates him in height from the child a year younger or a year older. From the standpoint of a completely accurate analysis of our sample, this discovery throws our data into an entirely new light.

SUMMARY OF BASIC RELATIONS OF ANALYSIS OF VARIANCE

A glance at the data in Worksheet No. 112, showing the heights of children for the six primary grades reveals that there is considerable variation in height within each class. The range of each of the six grades is

Grade	Range
1	41–50
2	47–50
3	48–55
4	51–57
5	54–59
6	56–61

TABLE 52

VALUES OF F[1]

5% OR 95% IN LIGHT-FACE TYPE,
1% OR 99% IN **Bold-Face Type**

n_1 degrees of freedom
for **greater** variance

n_2	1	2	3	4	5	6	7	8	9	10	11	12
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**
2	18.51 **98.49**	19.00 **99.01**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**
3	10.13 **34.12**	9.55 **30.81**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.70 **4.14**	2.64 **4.00**	2.59 **3.89**	2.55 **3.80**	2.51 **3.73**	2.48 **3.67**
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.66 **4.03**	2.59 **3.89**	2.54 **3.78**	2.49 **3.69**	2.45 **3.61**	2.42 **3.55**
17	4.45 **8.40**	3.59 **6.11**	3.20 **5.18**	2.96 **4.67**	2.81 **4.34**	2.70 **4.10**	2.62 **3.93**	2.55 **3.79**	2.50 **3.68**	2.45 **3.59**	2.41 **3.52**	2.38 **3.45**
18	4.41 **8.28**	3.55 **6.01**	3.16 **5.09**	2.93 **4.58**	2.77 **4.25**	2.66 **4.01**	2.58 **3.85**	2.51 **3.71**	2.46 **3.60**	2.41 **3.51**	2.37 **3.44**	2.34 **3.37**
19	4.38 **8.18**	3.52 **5.93**	3.13 **5.01**	2.90 **4.50**	2.74 **4.17**	2.63 **3.94**	2.55 **3.77**	2.48 **3.63**	2.43 **3.52**	2.38 **3.43**	2.34 **3.36**	2.31 **3.30**
20	4.35 **8.10**	3.49 **5.85**	3.10 **4.94**	2.87 **4.43**	2.71 **4.10**	2.60 **3.87**	2.52 **3.71**	2.45 **3.56**	2.40 **3.45**	2.35 **3.37**	2.31 **3.30**	2.28 **3.23**
21	4.32 **8.02**	3.47 **5.78**	3.07 **4.87**	2.84 **4.37**	2.68 **4.04**	2.57 **3.81**	2.49 **3.65**	2.42 **3.51**	2.37 **3.40**	2.32 **3.31**	2.28 **3.24**	2.25 **3.17**
22	4.30 **7.94**	3.44 **5.72**	3.05 **4.82**	2.82 **4.31**	2.66 **3.99**	2.55 **3.76**	2.47 **3.59**	2.40 **3.45**	2.35 **3.35**	2.30 **3.26**	2.26 **3.18**	2.23 **3.12**
23	4.28 **7.88**	3.42 **5.66**	3.03 **4.76**	2.80 **4.26**	2.64 **3.94**	2.53 **3.71**	2.45 **3.54**	2.38 **3.41**	2.32 **3.30**	2.28 **3.21**	2.24 **3.14**	2.20 **3.07**
24	4.26 **7.82**	3.40 **5.61**	3.01 **4.72**	2.78 **4.22**	2.62 **3.90**	2.51 **3.67**	2.43 **3.50**	2.36 **3.36**	2.30 **3.25**	2.26 **3.17**	2.22 **3.09**	2.18 **3.03**
25	4.24 **7.77**	3.38 **5.57**	2.99 **4.68**	2.76 **4.18**	2.60 **3.86**	2.49 **3.63**	2.41 **3.46**	2.34 **3.32**	2.28 **3.21**	2.24 **3.13**	2.20 **3.05**	2.16 **2.99**
26	4.22 **7.72**	3.37 **5.53**	2.98 **4.64**	2.74 **4.14**	2.59 **3.82**	2.47 **3.59**	2.39 **3.42**	2.32 **3.29**	2.27 **3.17**	2.22 **3.09**	2.18 **3.02**	2.15 **2.96**

Degrees freedom lesser variance

14	16	20	24	30	40	50	75	100	200	500	∞
245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,258**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**
19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**
8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**
5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**
4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**
3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**
3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**
3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**
3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**
2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**
2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**
2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**
2.55 **3.85**	2.51 **3.78**	2.46 **3.67**	2.42 **3.59**	2.38 **3.51**	2.34 **3.42**	2.32 **3.37**	2.28 **3.30**	2.26 **3.27**	2.24 **3.21**	2.22 **3.18**	2.21 **3.16**
2.48 **3.70**	2.44 **3.62**	2.39 **3.51**	2.35 **3.43**	2.31 **3.34**	2.27 **3.26**	2.24 **3.21**	2.21 **3.14**	2.19 **3.11**	2.16 **3.06**	2.14 **3.02**	2.13 **3.00**
2.43 **3.56**	2.39 **3.48**	2.33 **3.36**	2.29 **3.29**	2.25 **3.20**	2.21 **3.12**	2.18 **3.07**	2.15 **3.00**	2.12 **2.97**	2.10 **2.92**	2.08 **2.89**	2.07 **2.87**
2.37 **3.45**	2.33 **3.37**	2.28 **3.25**	2.24 **3.18**	2.20 **3.10**	2.16 **3.01**	2.13 **2.96**	2.09 **2.89**	2.07 **2.86**	2.04 **2.80**	2.02 **2.77**	2.01 **2.75**
2.33 **3.35**	2.29 **3.27**	2.23 **3.16**	2.19 **3.08**	2.15 **3.00**	2.11 **2.92**	2.08 **2.86**	2.04 **2.79**	2.02 **2.76**	1.99 **2.70**	1.97 **2.67**	1.96 **2.65**
2.29 **3.27**	2.25 **3.19**	2.19 **3.07**	2.15 **3.00**	2.11 **2.91**	2.07 **2.83**	2.04 **2.78**	2.00 **2.71**	1.98 **2.68**	1.95 **2.62**	1.93 **2.59**	1.92 **2.57**
2.26 **3.19**	2.21 **3.12**	2.15 **3.00**	2.11 **2.92**	2.07 **2.84**	2.02 **2.76**	2.00 **2.70**	1.96 **2.63**	1.94 **2.60**	1.91 **2.54**	1.90 **2.51**	1.88 **2.49**
2.23 **3.13**	2.18 **3.05**	2.12 **2.94**	2.08 **2.86**	2.04 **2.77**	1.99 **2.69**	1.96 **2.63**	1.92 **2.56**	1.90 **2.53**	1.87 **2.47**	1.85 **2.44**	1.84 **2.42**
2.20 **3.07**	2.15 **2.99**	2.09 **2.88**	2.05 **2.80**	2.00 **2.72**	1.96 **2.63**	1.93 **2.58**	1.89 **2.51**	1.87 **2.47**	1.84 **2.42**	1.82 **2.38**	1.81 **2.36**
2.18 **3.02**	2.13 **2.94**	2.07 **2.83**	2.03 **2.75**	1.98 **2.67**	1.93 **2.58**	1.91 **2.53**	1.87 **2.46**	1.84 **2.42**	1.81 **2.37**	1.80 **2.33**	1.78 **2.31**
2.14 **2.97**	2.10 **2.89**	2.04 **2.78**	2.00 **2.70**	1.96 **2.62**	1.91 **2.53**	1.88 **2.48**	1.84 **2.41**	1.82 **2.37**	1.79 **2.32**	1.77 **2.28**	1.76 **2.26**
2.13 **2.93**	2.09 **2.85**	2.02 **2.74**	1.98 **2.66**	1.94 **2.58**	1.89 **2.49**	1.86 **2.44**	1.82 **2.36**	1.80 **2.33**	1.76 **2.27**	1.74 **2.23**	1.73 **2.21**
2.11 **2.89**	2.06 **2.81**	2.00 **2.70**	1.96 **2.62**	1.92 **2.54**	1.87 **2.45**	1.84 **2.40**	1.80 **2.32**	1.77 **2.29**	1.74 **2.23**	1.72 **2.19**	1.71 **2.17**
2.10 **2.86**	2.05 **2.77**	1.99 **2.66**	1.95 **2.58**	1.90 **2.50**	1.85 **2.41**	1.82 **2.36**	1.78 **2.28**	1.76 **2.25**	1.72 **2.19**	1.70 **2.15**	1.69 **2.13**

Degrees freedom lesser variance

[1] Snedecor, G. W., *Statistical Methods*, pp. 184–187, 1940. Reprinted by permission.

TABLE 52

VALUES OF F[1] (Continued)

5% OR 95% IN LIGHT-FACE TYPE,
1% OR 99% IN **Bold-Face Type**
n_1 degrees of freedom
for **greater** variance

n_2	1	2	3	4	5	6	7	8	9	10	11	12
27	4.21 **7.68**	3.35 **5.49**	2.96 **4.60**	2.73 **4.11**	2.57 **3.79**	2.46 **3.56**	2.37 **3.39**	2.30 **3.26**	2.25 **3.14**	2.20 **3.06**	2.16 **2.98**	2.13 **2.93**
28	4.20 **7.64**	3.34 **5.45**	2.95 **4.57**	2.71 **4.07**	2.56 **3.76**	2.44 **3.53**	2.36 **3.36**	2.29 **3.23**	2.24 **3.11**	2.19 **3.03**	2.15 **2.95**	2.12 **2.90**
29	4.18 **7.60**	3.33 **5.42**	2.93 **4.54**	2.70 **4.04**	2.54 **3.73**	2.43 **3.50**	2.35 **3.33**	2.28 **3.20**	2.22 **3.08**	2.18 **3.00**	2.14 **2.92**	2.10 **2.87**
30	4.17 **7.56**	3.32 **5.39**	2.92 **4.51**	2.69 **4.02**	2.53 **3.70**	2.42 **3.47**	2.34 **3.30**	2.27 **3.17**	2.21 **3.06**	2.16 **2.98**	2.12 **2.90**	2.09 **2.84**
32	4.15 **7.50**	3.30 **5.34**	2.90 **4.46**	2.67 **3.97**	2.51 **3.66**	2.40 **3.42**	2.32 **3.25**	2.25 **3.12**	2.19 **3.01**	2.14 **2.94**	2.10 **2.86**	2.07 **2.80**
34	4.13 **7.44**	3.28 **5.29**	2.88 **4.42**	2.65 **3.93**	2.49 **3.61**	2.38 **3.38**	2.30 **3.21**	2.23 **3.08**	2.17 **2.97**	2.12 **2.89**	2.08 **2.82**	2.05 **2.76**
36	4.11 **7.39**	3.26 **5.25**	2.86 **4.38**	2.63 **3.89**	2.48 **3.58**	2.36 **3.35**	2.28 **3.18**	2.21 **3.04**	2.15 **2.94**	2.10 **2.86**	2.06 **2.78**	2.03 **2.72**
38	4.10 **7.35**	3.25 **5.21**	2.85 **4.34**	2.62 **3.86**	2.46 **3.54**	2.35 **3.32**	2.26 **3.15**	2.19 **3.02**	2.14 **2.91**	2.09 **2.82**	2.05 **2.75**	2.02 **2.69**
40	4.08 **7.31**	3.23 **5.18**	2.84 **4.31**	2.61 **3.83**	2.45 **3.51**	2.34 **3.29**	2.25 **3.12**	2.18 **2.99**	2.12 **2.88**	2.07 **2.80**	2.04 **2.73**	2.00 **2.66**
42	4.07 **7.27**	3.22 **5.15**	2.83 **4.29**	2.59 **3.80**	2.44 **3.49**	2.32 **3.26**	2.24 **3.10**	2.17 **2.96**	2.11 **2.86**	2.06 **2.77**	2.02 **2.70**	1.99 **2.64**
44	4.06 **7.24**	3.21 **5.12**	2.82 **4.26**	2.58 **3.78**	2.43 **3.46**	2.31 **3.24**	2.23 **3.07**	2.16 **2.94**	2.10 **2.84**	2.05 **2.75**	2.01 **2.68**	1.98 **2.62**
46	4.05 **7.21**	3.20 **5.10**	2.81 **4.24**	2.57 **3.76**	2.42 **3.44**	2.30 **3.22**	2.22 **3.05**	2.14 **2.92**	2.09 **2.82**	2.04 **2.73**	2.00 **2.66**	1.97 **2.60**
48	4.04 **7.19**	3.19 **5.08**	2.80 **4.22**	2.56 **3.74**	2.41 **3.42**	2.30 **3.20**	2.21 **3.04**	2.14 **2.90**	2.08 **2.80**	2.03 **2.71**	1.99 **2.64**	1.96 **2.58**
50	4.03 **7.17**	3.18 **5.06**	2.79 **4.20**	2.56 **3.72**	2.40 **3.41**	2.29 **3.18**	2.20 **3.02**	2.13 **2.88**	2.07 **2.78**	2.02 **2.70**	1.98 **2.62**	1.95 **2.56**
55	4.02 **7.12**	3.17 **5.01**	2.78 **4.16**	2.54 **3.68**	2.38 **3.37**	2.27 **3.15**	2.18 **2.98**	2.11 **2.85**	2.05 **2.75**	2.00 **2.66**	1.97 **2.59**	1.93 **2.53**
60	4.00 **7.08**	3.15 **4.98**	2.76 **4.13**	2.52 **3.65**	2.37 **3.34**	2.25 **3.12**	2.17 **2.95**	2.10 **2.82**	2.04 **2.72**	1.99 **2.63**	1.95 **2.56**	1.92 **2.50**
65	3.99 **7.04**	3.14 **4.95**	2.75 **4.10**	2.51 **3.62**	2.36 **3.31**	2.24 **3.09**	2.15 **2.93**	2.08 **2.79**	2.02 **2.70**	1.98 **2.61**	1.94 **2.54**	1.90 **2.47**
70	3.98 **7.01**	3.13 **4.92**	2.74 **4.08**	2.50 **3.60**	2.35 **3.29**	2.23 **3.07**	2.14 **2.91**	2.07 **2.77**	2.01 **2.67**	1.97 **2.59**	1.93 **2.51**	1.89 **2.45**
80	3.96 **6.96**	3.11 **4.88**	2.72 **4.04**	2.48 **3.56**	2.33 **3.25**	2.21 **3.04**	2.12 **2.87**	2.05 **2.74**	1.99 **2.64**	1.95 **2.55**	1.91 **2.48**	1.88 **2.41**
100	3.94 **6.90**	3.09 **4.82**	2.70 **3.98**	2.46 **3.51**	2.30 **3.20**	2.19 **2.99**	2.10 **2.82**	2.03 **2.69**	1.97 **2.59**	1.92 **2.51**	1.88 **2.43**	1.85 **2.36**
125	3.92 **6.84**	3.07 **4.78**	2.68 **3.94**	2.44 **3.47**	2.29 **3.17**	2.17 **2.95**	2.08 **2.79**	2.01 **2.65**	1.95 **2.56**	1.90 **2.47**	1.86 **2.40**	1.83 **2.33**
150	3.91 **6.81**	3.06 **4.75**	2.67 **3.91**	2.43 **3.44**	2.27 **3.14**	2.16 **2.92**	2.07 **2.76**	2.00 **2.62**	1.94 **2.53**	1.89 **2.44**	1.85 **2.37**	1.82 **2.30**
200	3.89 **6.76**	3.04 **4.71**	2.65 **3.88**	2.41 **3.41**	2.26 **3.11**	2.14 **2.90**	2.05 **2.73**	1.98 **2.60**	1.92 **2.50**	1.87 **2.41**	1.83 **2.34**	1.80 **2.28**
400	3.86 **6.70**	3.02 **4.66**	2.62 **3.83**	2.39 **3.36**	2.23 **3.06**	2.12 **2.85**	2.03 **2.69**	1.96 **2.55**	1.90 **2.46**	1.85 **2.37**	1.81 **2.29**	1.78 **2.23**
1,000	3.85 **6.66**	3.00 **4.62**	2.61 **3.80**	2.38 **3.34**	2.22 **3.04**	2.10 **2.82**	2.02 **2.66**	1.95 **2.53**	1.89 **2.43**	1.84 **2.34**	1.80 **2.26**	1.76 **2.20**
∞	3.84 **6.64**	2.99 **4.60**	2.60 **3.78**	2.37 **3.32**	2.21 **3.02**	2.09 **2.80**	2.01 **2.64**	1.94 **2.51**	1.88 **2.41**	1.83 **2.32**	1.79 **2.24**	1.75 **2.18**

Degrees freedom lesser variance

14	16	20	24	30	40	50	75	100	200	500	∞	n_2
2.08 **2.83**	2.03 **2.74**	1.97 **2.63**	1.93 **2.55**	1.88 **2.47**	1.84 **2.38**	1.80 **2.33**	1.76 **2.25**	1.74 **2.21**	1.71 **2.16**	1.68 **2.12**	1.67 **2.10**	27
2.06 **2.80**	2.02 **2.71**	1.96 **2.60**	1.91 **2.52**	1.87 **2.44**	1.81 **2.35**	1.78 **2.30**	1.75 **2.22**	1.72 **2.18**	1.69 **2.13**	1.67 **2.09**	1.65 **2.06**	28
2.05 **2.77**	2.00 **2.68**	1.94 **2.57**	1.90 **2.49**	1.85 **2.41**	1.80 **2.32**	1.77 **2.27**	1.73 **2.19**	1.71 **2.15**	1.68 **2.10**	1.65 **2.06**	1.64 **2.03**	29
2.04 **2.74**	1.99 **2.66**	1.93 **2.55**	1.89 **2.47**	1.84 **2.38**	1.79 **2.29**	1.76 **2.24**	1.72 **2.16**	1.69 **2.13**	1.66 **2.07**	1.64 **2.03**	1.62 **2.01**	30
2.02 **2.70**	1.97 **2.62**	1.91 **2.51**	1.86 **2.42**	1.82 **2.34**	1.76 **2.25**	1.74 **2.20**	1.69 **2.12**	1.67 **2.08**	1.64 **2.02**	1.61 **1.98**	1.59 **1.96**	32
2.00 **2.66**	1.95 **2.58**	1.89 **2.47**	1.84 **2.38**	1.80 **2.30**	1.74 **2.21**	1.71 **2.15**	1.67 **2.08**	1.64 **2.04**	1.61 **1.98**	1.59 **1.94**	1.57 **1.91**	34
1.98 **2.62**	1.93 **2.54**	1.87 **2.43**	1.82 **2.35**	1.78 **2.26**	1.72 **2.17**	1.69 **2.12**	1.65 **2.04**	1.62 **2.00**	1.59 **1.94**	1.56 **1.90**	1.55 **1.87**	36
1.96 **2.59**	1.92 **2.51**	1.85 **2.40**	1.80 **2.32**	1.76 **2.22**	1.71 **2.14**	1.67 **2.08**	1.63 **2.00**	1.60 **1.97**	1.57 **1.90**	1.54 **1.86**	1.53 **1.84**	38
1.95 **2.56**	1.90 **2.49**	1.84 **2.37**	1.79 **2.29**	1.74 **2.20**	1.69 **2.11**	1.66 **2.05**	1.61 **1.97**	1.59 **1.94**	1.55 **1.88**	1.53 **1.84**	1.51 **1.81**	40
1.94 **2.54**	1.89 **2.46**	1.82 **2.35**	1.78 **2.26**	1.73 **2.17**	1.68 **2.08**	1.64 **2.02**	1.60 **1.94**	1.57 **1.91**	1.54 **1.85**	1.51 **1.80**	1.49 **1.78**	42
1.92 **2.52**	1.88 **2.44**	1.81 **2.32**	1.76 **2.24**	1.72 **2.15**	1.66 **2.06**	1.63 **2.00**	1.58 **1.92**	1.56 **1.88**	1.52 **1.82**	1.50 **1.78**	1.48 **1.75**	44
1.91 **2.50**	1.87 **2.42**	1.80 **2.30**	1.75 **2.22**	1.71 **2.13**	1.65 **2.04**	1.62 **1.98**	1.57 **1.90**	1.54 **1.86**	1.51 **1.80**	1.48 **1.76**	1.46 **1.72**	46
1.90 **2.48**	1.86 **2.40**	1.79 **2.28**	1.74 **2.20**	1.70 **2.11**	1.64 **2.02**	1.61 **1.96**	1.56 **1.88**	1.53 **1.84**	1.50 **1.78**	1.47 **1.73**	1.45 **1.70**	48
1.90 **2.46**	1.85 **2.39**	1.78 **2.26**	1.74 **2.18**	1.69 **2.10**	1.63 **2.00**	1.60 **1.94**	1.55 **1.86**	1.52 **1.82**	1.48 **1.76**	1.46 **1.71**	1.44 **1.68**	50
1.88 **2.43**	1.83 **2.35**	1.76 **2.23**	1.72 **2.15**	1.67 **2.06**	1.61 **1.96**	1.58 **1.90**	1.52 **1.82**	1.50 **1.78**	1.46 **1.71**	1.43 **1.66**	1.41 **1.64**	55
1.86 **2.40**	1.81 **2.32**	1.75 **2.20**	1.70 **2.12**	1.65 **2.03**	1.59 **1.93**	1.56 **1.87**	1.50 **1.79**	1.48 **1.74**	1.44 **1.68**	1.41 **1.63**	1.39 **1.60**	60
1.85 **2.37**	1.80 **2.30**	1.73 **2.18**	1.68 **2.09**	1.63 **2.00**	1.57 **1.90**	1.54 **1.84**	1.49 **1.76**	1.46 **1.71**	1.42 **1.64**	1.39 **1.60**	1.37 **1.56**	65
1.84 **2.35**	1.79 **2.28**	1.72 **2.15**	1.67 **2.07**	1.62 **1.98**	1.56 **1.88**	1.53 **1.82**	1.47 **1.74**	1.45 **1.69**	1.40 **1.62**	1.37 **1.56**	1.35 **1.53**	70
1.82 **2.32**	1.77 **2.24**	1.70 **2.11**	1.65 **2.03**	1.60 **1.94**	1.54 **1.84**	1.51 **1.78**	1.45 **1.70**	1.42 **1.65**	1.38 **1.57**	1.35 **1.52**	1.32 **1.49**	80
1.79 **2.26**	1.75 **2.19**	1.68 **2.06**	1.63 **1.98**	1.57 **1.89**	1.51 **1.79**	1.48 **1.73**	1.42 **1.64**	1.39 **1.59**	1.34 **1.51**	1.30 **1.46**	1.28 **1.43**	100
1.77 **2.23**	1.72 **2.15**	1.65 **2.03**	1.60 **1.94**	1.55 **1.85**	1.49 **1.75**	1.45 **1.68**	1.39 **1.59**	1.36 **1.54**	1.31 **1.46**	1.27 **1.40**	1.25 **1.37**	125
1.76 **2.20**	1.71 **2.12**	1.64 **2.00**	1.59 **1.91**	1.54 **1.83**	1.47 **1.72**	1.44 **1.66**	1.37 **1.56**	1.34 **1.51**	1.29 **1.43**	1.25 **1.37**	1.22 **1.33**	150
1.74 **2.17**	1.69 **2.09**	1.62 **1.97**	1.57 **1.88**	1.52 **1.79**	1.45 **1.69**	1.42 **1.62**	1.35 **1.53**	1.32 **1.48**	1.26 **1.39**	1.22 **1.33**	1.19 **1.28**	200
1.72 **2.12**	1.67 **2.04**	1.60 **1.92**	1.54 **1.84**	1.49 **1.74**	1.42 **1.64**	1.38 **1.57**	1.32 **1.47**	1.28 **1.42**	1.22 **1.32**	1.16 **1.24**	1.13 **1.19**	400
1.70 **2.09**	1.65 **2.01**	1.58 **1.89**	1.53 **1.81**	1.47 **1.71**	1.41 **1.61**	1.36 **1.54**	1.30 **1.44**	1.26 **1.38**	1.19 **1.28**	1.13 **1.19**	1.08 **1.11**	1,000
1.69 **2.07**	1.64 **1.99**	1.57 **1.87**	1.52 **1.79**	1.46 **1.69**	1.40 **1.59**	1.35 **1.52**	1.28 **1.41**	1.24 **1.36**	1.17 **1.25**	1.11 **1.15**	1.00 **1.00**	∞

Degrees freedom lesser variance

[1] Snedecor, G. W., *Statistical Methods*, pp. 184–187, 1940. Reprinted by permission.

This variation inside of each grade sample means that if all six grade samples were thrown into one large sample of 48 items, a large part of the total variation in that total sample would be due to the *intraclass variations.* It is also evident, however, that the average height of each succeeding class is much greater than that of the previous class. The total and average differences between classes are:

Grade	Total	Mean
1	363	45.4
2	388	48.5
3	413	51.6
4	427	53.4
5	451	56.4
6	474	59.2

From these two analyses it is evident that part of the total variation is due to variation *within* the classes, and part of it is the result of variation *between* the classes. If there were no variation *between* the classes, all the variation would be *within* the several grades. If, on the other hand, there were no variation within any class, all the variation would be between the grades. If, for instance, all first-grade children were 45 inches tall, all second grade pupils were 48 inches tall, all third-grade pupils were 52 inches tall, and all other grades followed the same pattern, all the variation would be *between* grades. But since both kinds of variation occur at the same time, the problem of variance is to determine (1) how much of the total is due to *within* class variation, and (2) how much of the total is due to *between* class variation. This computation is made by (1) assuming the null hypothesis of no difference among classes and then computing the total variance of all the data of all classes from their common mean, and then (2) computing the amount of variation *between* the classes, and (3) subtracting the sum of *between* variation from the *total* variation to obtain the *within* variation. The last step is to divide the mean *between* variation by the mean *within* variation to obtain the ratio between the two portions of the total variance.

$$F = \frac{\text{Variance between group averages}}{\text{Variance within group}}$$

If this ratio, F, is larger than could logically be accounted for on the basis of chance sampling, it should be concluded that the separate class samples are from different populations. If, on the other hand, this ratio, F, is so small that the differences among the classes can be logically accounted for on the bases of chance sampling, it should be concluded that the samples are not significantly different, but might have all been taken from the same population.

The estimation of the significance of the difference between only two means at one time can be made by the methods developed in Chapters 13 and 20, and generally known as the Critical Ratio, but the analysis of variance enables one to make comparisons among a larger number of means in one computation.

As will be indicated in later examples, more than one type of *between* measurement may be made for the data at the same time. These refinements greatly increase the accuracy of the measurements of the several *between* types of variation.

WORKSHEET NO. 113

SUMMARY OF CABBAGE SEED TREATMENT TESTS IN EXPERIMENTS IN SIX STATES

States	Number of Seeds Emerging			Total
	Treatments			
	Zno	Sem.	None	
Delaware	57	56	46	
Florida	73	74	52	
Louisiana	62	66	40	
Maryland	70	68	55	
New York	79	78	72	
Virginia	60	67	53	
Totals	401	409	318	1,128
Means	66.83	68.17	53.00	62.67

First Step. Computation of total sum of squares for six states

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$X^2 = (57)^2 + (73)^2 + (62)^2 + \cdots + (53)^2 = 72{,}766$$

$$\frac{(\Sigma X)^2}{N} = \frac{(1{,}128)^2}{18} = \frac{1{,}272{,}384}{18} = 70{,}688$$

$$\Sigma x^2 = 72{,}766 - 70{,}688 = 2{,}078$$

Second Step. Computation of sum of squares between means.

$$\frac{\Sigma[(\Sigma X_1)^2 + (\Sigma X_2)^2 + (\Sigma X_3)^2]}{f} = \frac{160{,}801 + 167{,}281 + 101{,}124}{6}$$

$$= \frac{429{,}206}{6} = 71{,}534.3 - 70{,}688 = 846.33$$

In the first step all of the data are thrown together in one sample on the theory of the null hypothesis that there is no significant difference in the results whether the cabbage seeds are treated or not. It is assumed that just as many of the seeds will grow when they are not treated as when they are. On this assumption, the data are considered *one* sample and the squared deviations of all the items from their *common* mean are computed. Their sum is 2,078. This is the total sum of squares in the data. This total is composed of two parts. One part of this total is the result of variation *within* each treatment. For instance, in the treatment Zno, the data vary all the way from 57 to 79. In the treatment Sem., the variations are in the range of 56 to 78, and in the treatment None, the variation spreads from 46 to 72. It is evident that part of the total variance of 2,078 is due to this variation *within* the several treatments. But it is also evident that there is a considerable variation *between* the means of the three treatments which are 66.83 for Zno, 68.17 for Sem., 53.00 for None. The problem of the analysis of variance is to compute the mean amount of variance which is the result of *within variation* as compared with *between variation.* The ratio between the *within mean squares* and the *between mean squares* is the measure of significance.

Method of computing the *Critical Ratio,* or the difference between two means, was developed in Chapters 13 and 20 by means of the formula:

$$t = \frac{\bar{X}_1 - X_2}{\sigma_D}$$

This formula limits one to the comparison of only two means at a time. The analysis of variance provides a method of comparing the variation among a group of several means at one time, or by one computation. It also provides the means of splitting the total variance into two parts, (1) that variation which is the result of scatter of items *within* classes, and (2) that variation which is due to divergence *between the means* of the several classes. The *t*-ratio, or the Critical Ratio, should be used when one wishes to make comparisons between only *two* classes at a time. The analysis of variance must be used to make an over-all comparison among a group of means, or comparisons among several means at once. The analysis of variance may also be used to compare only two means, but the Critical Ratio is limited to two at one time.

The sum of the squared deviations within the three separate treatments computed from the treatment means may be verified as follows:

Treatment	*Squared Deviations Within Treatments*
Zno	362.83
Sem.	284.84
None	584.00
Total	1,231.67

Computed by the method of variance, the total within squared deviations are also 1,231.67 as shown below.

TABLE 53

COMPUTATION OF *F*, TEST OF SIGNIFICANCE OF DIFFERENCE AMONG TREATMENT MEANS

	Degrees of Freedom	Sum of Squares	Mean Squares
Total	17	2,078	
Between Treatments	2	846.33	423.16
Within Treatments	15	1,231.67	82.11

$$F = \frac{423.16}{82.11} = 5.153$$

Whether the within-treatments variation is computed for each class (treatment) mean separately and then summed, or is computed by subtracting the between-variation from the total variation, the results are identical. In this case, both are 1,231.67. This test is a check on the accuracy of the computations.

In this case, the results are significant. In checking on Table 52 — Values of F for 2 degrees of freedom (df) across the top of the table, and 15 degrees of freedom (df) down the left-hand side of the table, one finds for the 5% level of significance, 3.68 and for the 1% level of significance, 6.36. Since our F-value is 5.15, we conclude that our results are significant, but not quite large enough to be highly significant. A larger sample would, perhaps, give a result which would be highly significant. But from this small sample it may be concluded that the cabbage seeds which were treated were significantly better in germination than those which were not treated.

It is possible to use the analysis of variance on a wide variety of sociological, economic, and business problems, such as population groups, advertising, current ratios, sales expenses, and class differences in education. In most cases the analysis of variance would not be sufficiently complete treatment for data in the social sciences, but it might be used as a preliminary method to discover whether the data would justify a more complete analysis.

WORKSHEET NO. 114

ANALYSIS OF VARIANCE OF SIZE OF FAMILIES IN LARGE CITIES, SMALL CITIES, AND RURAL COUNTIES IN THE CORN BELT

Rows	Large Cities	Small Cities	Rural Counties	Total
1	3.84	3.60	3.92	11.36
2	3.68	3.48	4.15	11.31
3	3.71	3.51	4.23	11.45
4	3.74	3.18	4.52	11.44
5	4.00	3.35	3.97	11.32
6	3.96	3.46	3.87	11.29
7	3.69	3.25	4.36	11.30
8	3.82	3.60	3.88	11.30
Totals	30.44	27.43	32.90	90.77

1st Step. Computation of total sum of squares, 3 groups

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$\Sigma X^2 = (3.84)^2 + (3.68)^2 + \cdots + (3.88)^2 = 345.8633$$

$$\Sigma X = 90.77$$

$$\frac{(\Sigma X)^2}{N} = \frac{(90.77)^2}{24} = \frac{8{,}239.1929}{24} = 343.2997$$

$$\Sigma x^2 = 345.8633 - 343.2997 = 2.5636, \text{ total variance}$$

2d Step. $\dfrac{\Sigma(\text{Class Sums})^2}{f} - \text{Correction} =$

$$\frac{(30.44)^2 + (27.43)^2 + (32.90)^2}{8} = \frac{2761.4085}{8} = 345.1761$$

$345.1761 - 343.2997 = 1.8764$, between-groups variance

3d Step. Sum of squares between-rows

$$\frac{\Sigma(\text{Row Totals})^2}{f} - \text{Correction} =$$

$$(11.36)^2 + (11.31)^2 + (11.45)^2 + \cdots + (11.30)^2 = 1{,}029.9283$$

$$\frac{1{,}029.9283}{3} = 343.30943$$

$343.30943 - 343.2997 = .0097$, between-rows variance

COMPUTATION OF F, TEST OF SIGNIFICANCE OF DIFFERENCE BETWEEN GROUPS, BETWEEN ROWS, AND EXPERIMENTAL ERROR

	Degrees of Freedom	*Sum of Squares*	*Mean Squares*
Total	23	2.5636	
Between Groups	2	1.8764	.9382
Between Rows	7	.0097	.0014
Experimental Error	14	.6775	.0484

$$\text{Groups } F = \frac{.9382}{.0484} = 19.4 \quad \text{Highly significant}$$

$$\text{Rows } F = \frac{.0014}{.0484} = 0.03 \quad \text{not significant}$$

EXPERIMENTAL ERROR

An additional step is introduced in this problem in order to check the uniformity of the rows of data. If there were any real relationship between the items of data in a row, if the tie between the large city, small city, and rural county family size in each row were logically associated, it is possible for a large portion of the total variance to be associated with row size instead of group size. To measure how much of the total is associated with row relations, the third step is introduced into the problem. In this particular problem row relations have no logical and no actual significance as proved by the test. In many problems, however, in which the student will work, this cross-relation will be very important and should be measured and subtracted from the within variance which will hereafter be called Experimental Error. It is the total error caused by chance sampling and other errors all lumped together as a total unexplained residual. It is always the *denominator* of the F-ratio against which all computed variances are measured.

In setting up a problem for solution by the analysis of variance, the experimental design should eliminate as far as possible all extraneous factors. If one wished to compute the effects of various amounts of artificial light in hen houses on egg production in the winter season, hens should be chosen for the experiment which are as far as possible equal in age, weight, kind of feed, and previous treatment and condition. They should be of the same breed and otherwise identical. If one group is to be subjected to only eight hours of light while the other group is to be given sixteen or twenty hours of light per day, in order to test the effect of light on egg production, all other factors should be held constant. If all the other factors are not held constant, it may be that part of the difference in egg production is due to some of these other variables and is not due to difference in the number of light hours per day. If, for instance, one group of hens were one year old and the other group were three years old, the difference in egg production might be due primarily to age and not to light. If the feed of one group were superior to that

of the other group, the difference in egg production might be due to feed and not to light. It is evident that if the effect of longer days on egg production due to increased artificial light is to be measured accurately, age, feed, breed, and all other factors must be held constant.

In spite of all effort to hold all extraneous factors constant, it is often impossible to do so. Fortunately, a statistical device is available whereby such uncontrolled differences may be computed out and the results held constant in spite of them. In Worksheet No. 112 on the heights of school children, no such correction or control was used. In Worksheet No. 114 on size of families, the variation between rows caused by any factors was computed and subtracted out of the total variance *before* the ratio between size of family variance and *experimental* error was computed. In this case, the variance among the rows was quite small, .0014, but in many problems it may be quite large. In any case, in which there is a logical basis for assuming a correlation between the several columns it is more accurate procedure to compute it and remove it from the within variance before F is computed for the columns. Worksheet No. 114 is, therefore, a better method than Worksheet No. 112.

WIDER POSSIBILITY OF STUDY

The method of computing out the influence of the extraneous and unequal factors enables one to use more data and to widen his research field and to work on problems which would otherwise be excluded. For instance, if it were impossible to obtain a sufficient number of hens of the same age or breed, one could use hens of various ages and breeds by *computing out the average effect of the unequal* factors, and thus arrive at a satisfactory result. Often this is the only method on which one can proceed.

After all these unequal and extraneous elements have been computed out of the variance that one can get at, all the rest of the variance is lumped together and called *Experimental Error.* Theoretically, experimental error should include only that error which is due to chance sampling, but actually in most problems

it also includes other elements of variance which the statistician is not able to segregate and remove. Experimental error is in any case the base or denominator of the F-ratio of significance.

The principles involved in the egg production problem would, of course, apply equally well in all problems in the physical, biological, or social sciences. The principles are universal.

PRACTICAL RESEARCH PROBLEM SIGNIFICANCE OF TREATMENTS OF COTTON SEED

In Worksheet No. 115 a much more complex and practical problem is introduced in which five separate elements of variance are analyzed; namely, series, varieties, chemicals, varieties and chemicals, and experimental error. This is an actual problem worked out by Dr. K. Starr Chester and Dr. W. W. Ray of the Botany Department of Oklahoma A. and M. College, to determine the significance of treating cotton seed for disease resistance. Delinted-graded cotton seed and ordinary cotton seed with the lint left on it by the gin were both subjected to identical chemical treatments to determine the effect on plants emerging and surviving. The chemical treatments proved to be highly significant in preventing disease, especially on the delinted cotton.

The computations for Worksheet No. 115 are given in detail below as a guide to the student in organizing such a complex problem. There are five steps in the solution.

1st Step.

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = (81)^2 + (90)^2 + (30)^2 + \cdots (17)^2 = 127{,}703$$

$$\frac{(\Sigma X)^2}{N} = \frac{(2005)^2}{48} = \frac{4{,}020{,}025}{48} = 83{,}751$$

$\Sigma x^2 = 127{,}703 - 83{,}751 = 43{,}952$, Total sum of square

2d Step.

$$\frac{(482)^2 + (466)^2 + (496)^2 + (561)^2}{12} = 84{,}185, \text{ Series sum of squares}$$

$$84{,}185 - 83{,}751 = 434$$

WORKSHEET NO. 115

THE PERCENTAGE OF GERMINATION OF FUZZY AND DELINTED-GRADED COTTON SEED TREATED WITH VARIOUS CHEMICALS AND PLANTED IN SOIL INFESTED WITH RHIZOCTONIA

Variety and Chemical	1	2	3	4	Total
Del-graded					
Ceresan	81	85	83	94	343
Spergon	90	80	91	99	360
Spergonex	30	18	16	29	93
Cyanamid	15	21	11	23	70
Sanoseed	26	25	29	28	108
Check	15	16	24	31	86
Variety Total					1,060
Fuzzy					
Ceresan	79	70	84	79	312
Spergon	81	70	83	67	301
Spergonex	15	31	8	13	67
Cyanamid	25	32	30	63	150
Sanoseed	12	10	14	18	54
Check	13	8	23	17	61
Series Total	482	466	496	561	
Variety Total					945
Grand Total					2,005

Totals

Ceresan	655**
Spergon	661**
Spergonex	160
Cyanamid	220
Sanoseed	162
Check	147

(Series) $\frac{2}{12}$ = 84,185

(Treatments) $\frac{2}{24}$ = 84,026

(Chem.) $\frac{2}{8}$ = 123,475

(V & C) $\frac{2}{4}$ = 125,357

SX = 2,005

SX^2 = 127,703

CF = 83,751

Variate	*df*	SD^2	*Mean Square*	*F-Value*		
				Calc.	*5%*	*1%*
Total	47	43,952				
Series	3	434	138	2.50	2.89	4.44
Varieties	1	275	275	4.74	4.14	7.47
Chemicals	5	39,724	7,945	137.00**	2.50	3.64
V & C Joint	5	1,607	321	5.54**	2.50	3.64
Error	33	1,912	58.0			

** These are the highly significant or important factors.

Worksheet No. 115 illustrates how a complex combination of data can be explained by the analysis of variance.

3d Step.

$$\frac{(1060)^2 + (945)^2}{24} = \frac{2{,}016{,}625}{24} = 84{,}026$$

$$84{,}026 - 43{,}751 = 275, \text{ Varieties sum of squares}$$

4th Step.

$$(655)^2 + (661)^2 + (160)^2 + (220)^2 + (162)^2 + (147)^2 = \frac{987{,}799}{8} = 123{,}475$$

$$123{,}475 - 83{,}751 = 39{,}724, \text{ Chemicals sum of squares}$$

5th Step.

$$(343)^2 + (360)^2 + (93)^2 + \cdots + (61)^2 = \frac{501{,}429}{4} = 125{,}357$$

$$125{,}357 - 83{,}751 = 41{,}606, \text{ Total variance of 12 sub-classes}$$

$$41{,}606 - (39{,}724 + .275) = 1{,}607$$

$$\left\{\begin{array}{l}\text{Total variance of 12 sub-}\\ \text{classes of 2 varieties and}\\ \text{6 chemicals}\end{array}\right\} - \left\{\begin{array}{l}\text{Chemicals plus}\\ \text{varieties sum of}\\ \text{squares}\end{array}\right\} = \left\{\begin{array}{l}\text{Joint varieties}\\ \text{and chemicals}\\ \text{sum of squares}\end{array}\right\}$$

TABLE 54

COMPUTATION OF F

Variate	*df*	Squared Deviations	Mean Squares	*F*-Values
Total	47	43,952		
Series	3	434	145	2.50
Varieties	1	275	275	4.74
Chemicals	5	39,724	7,945	137.00
V & C Joint	5	1,607	321	5.54
Error	33	1,912	58.0	

The five problems explained in this chapter offer the student a sufficiently complete understanding of the analysis of variance to enable him to use this method effectively. There are, of course, still more complex variations of analysis possible, but the student who expects to specialize in research will take at least a second course in statistics.

DETAILED ANALYSIS OF WORKSHEET NO. 115

Data

The data in this problem are the percentages of germination of (1) fuzzy and (2) delinted cotton seed, separate plots of which are treated with five chemicals as follows: (1) Ceresan, (2) Spergon, (3) Spergonex, (4) Cyanamid, (5) Sanoseed, and (6) one Check plot not treated. Separate plots of the fuzzy and of the delinted each receive the same chemical treatment.

Objectives of Analysis

The several purposes in mind in making this analysis are to determine (1) whether the several chemical treatments produce significantly different results on the (*a*) fuzzy and on the (*b*) delinted cotton; (2) whether there are any significant differences among the results produced by the five different chemicals; (3) whether there is any significant difference between the Check, or non-chemical, plot in each case and the plots on which chemically treated seeds were planted; (4) whether there is any combined or joint association of varieties and chemicals that is significant. On this point it may be said that there is a possibility that the chemicals might have a more direct and powerful effect on the delinted seed; (5) four separate plots were used, as is indicated by the four columns of data, to reduce the error of chance sampling by providing a larger number of cases. It may be said that a still larger number of items or plots, as could have been provided in a Latin Square 6×6, for each variety (fuzzy and delinted) would have been still better; (6) to segregate and measure Experimental Error as the basis of the F-ratios of significance.

Computations for Analysis

For these six objectives, six steps of analysis are required as follows:

First Step: Computation of Total Variance for all the data based on the null-hypothesis. Total sum of squares = 43,952.

Second Step: Computation of variance between the four columns of data, or four series. Series sum of squares = 434. This computation removes from the Experimental Error any element of variance due to variations in soil, rain, methods of cultivation, etc., among the plots or series (columns of data in this case). Theoretically, if there were no variations in the percentage of germinations due to plot, soil, rain, or other extraneous variations, the germination in the parallel plots should be the same. Since they are not identical, it must be assumed that they contain the effects of some extraneous factors which must be removed for an accurate measurement of the other desired variances.

Third Step: Computation of variance between varieties (fuzzy and delinted in this case). Variety sum of squares = 275. This result means that the chemicals were more effective on one of the varieties than on the other.

Fourth Step: Computation of variation among the effects of the chemicals. Chemicals sum of squares = 39,724, which is the larger portion of the total variance. It is quite evident that (1) the results on the chemically treated plots were quite superior to those of the Check plot on which no chemicals were used, and (2) that some of the chemicals were much more effective than others. The divergence among the effects of the chemicals is responsible for most of the total variance. This is the decisive factor in the problem.

Fifth Step: Computation of joint effects of varieties and chemicals. Joint V and C sum of squares = 1,607. This step in the analysis is an important one. Sometimes the relationship between two, or more, factors is multiplicative instead of additive. In such a case when two are combined the result may be greater than their sum. In this case the fuzzy cotton might not yield as readily to the benefits of the chemicals as does the delinted cotton. The extent to which this relationship holds in this problem will be revealed by a computation of the variance of all twelve chemical treatments under the two varieties. In this

case, the result is not significant, but in some types of problems it will be, and the student should know how to compute the value.

Sixth Step: By means of the five computations above the total variance and the separate portions of the total variance which are associated with or are the results of the several factors included in the problem are segregated and measured. The exact portion of the total variance related to each factor is now known. To complete the analysis it is necessary to reduce these separate portions of the total variance to ratios with a common base. Since the number of items or more exactly the degrees of freedom associated with the different factors in the problem are not equal, it is first necessary to reduce each *sub-total* of variance to *mean* variance so that all the separate portions of the variance will be on exactly the same basis and in identical units. The final step is to reduce all these separate *mean variances* to comparable ratios by dividing each one by the *mean* Experimental Error. The Experimental Error, as was stated above, is that portion of the total variance which is due to *chance sampling* and to all other unidentified variation.

TABLE 55

SIGNIFICANCE OF *F*-RATIOS
In Fuzzy and Delinted Cotton Seed Problem
F-Values

	Computed	Required for	
		5%	1%
Series	2.50	2.89	4.44
Varieties	4.74	4.14	7.47
Chemicals	137.00	2.50	3.64
V. and C. Joint	5.54	2.50	3.64

From the above table it is evident (1) that the variance of varieties is significant but *not* highly significant, (2) that chemicals variance is most highly significant, (3) that V and C Joint variance is highly significant, and (4) that Series or plot variance is *not* significant.

CONCLUSION

If the mean variance for a factor is greater than that which could be reasonably expected to occur from chance sampling, the difference is considered significant. "Significant" in its use in the statistical sense means a difference *so* large that it occurs by chance only one time out of 20, or only in 5% of the cases. "Highly significant" is a statistical term used to indicate a difference *so* great that it could occur by chance only one time in 100, or only in 1% of the cases. Experimental Error is that variation which is the result of chance sampling. The ratio, therefore, between mean Experimental Error and any other given mean variance indicates the degree to which the variance of this given factor could have occurred by chance. This is the *F*-ratio.

R. A. Fisher and George W. Snedecor have worked out the mathematical limits of chance sampling occurrences for each combination of degrees of freedom. These ratios for both the 5% level of chance and the 1% level of chance are given in Table 52 — Values of *F*. By comparing the computed *F*-ratio for any problem with the values in Table 52 for the same degrees of freedom, one can determine whether his computed *F*-ratio is sufficiently "highly significant." The higher the computed ratio is above the required ratio, the more certain one is of the dependability of his results.

SUMMARY

1. Variance is the square of the standard deviation, or the mean of the summed squared deviations from the mean.

2. The null-hypothesis is the assumption that there is no significant difference between two or more series of data.

3. The analysis of variance is a method of reducing to a ratio (1) the portion of total variance due to variation between the means of the group of data and (2) the portion of the total variance which falls within the group of data, with this second portion as the base of the ratio. The variance is said to be significant when the obtained ratio is so large that it could be accounted for by chance sampling only one time in twenty.

4. The method of analysis of variance makes it possible to make comparisons among any number of groups of data at once.

5. The analysis of variance is especially useful in computing the degree of significance among the results obtained from planned experiments, especially in agriculture, based on the Latin Square or other types of replications.

6. The analysis of variance does not give an exact measure of the percentage of relationship between two variables as provided by the coefficient of determination, but only a ratio which by means of the F-table may be listed as (1) not significant, (2) significant, (3) highly significant, or (4) clearly larger than a ratio of high significance.

7. The analysis of variance may be used in the fields of economics, sociology, education, psychology, political science or other fields of social science. It is a tool, however, which should be applied to carefully planned research projects providing exact data.

SELECTED REFERENCES

Fisher, R. A., *Statistical Methods for Research Workers*, Chapters VII, VIII. Oliver and Boyd, Ltd., Edinburgh, 1925.

Mills, F. C., *Statistical Methods, Revised*, Chapter XV. Henry Holt & Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapter XIX. John Wiley & Sons, Inc., New York, 1942.

Snedecor, G. W., *Statistical Methods*, Chapters X, XI. Collegiate Press, Inc., Ames, Iowa, 1946.

REVIEW QUESTIONS

1. Define variance. How is it related to σ? What does it measure?
2. Explain the principle underlying the null-hypothesis. Illustrate.
3. What is meant by "degrees of freedom"? In connection with what kinds of computations is it used? Why?
4. What is "within-group" variance?
5. What is "between-groups" variance?
6. What is the relationship of "between-groups" variance to "within-group" variance?
7. What is the meaning of the F-coefficient? How is it computed?
8. What is "Experimental Error"? How is it computed?
9. For what reason is the variance computed for the "rows," or crosswise as well as for the classes or groups?
10. What is the advantage of computing several "cross" or "inter" variance relations?

EXERCISES

1. Analysis of Variance.

WEIGHT GAINS OF 30 SWINE FED SIX TYPES OF RATIONS

1	2	3	4	5	6
35	40	39	27	25	22
19	29	27	7	31	18
41	46	20	13	46	36
15	41	29	33	49	39
30	34	45	35	24	20

2. Analysis of Variance.

AVERAGE DAILY GAINS OF 24 SWINE CLASSIFIED ACCORDING TO RATION AND WEIGHT AND BEGINNING OF PERIOD

	Rations			
Weight	1	2	3	4
1	1.40	1.31	1.40	1.96
2	1.79	1.30	1.47	1.77
3	1.72	1.21	1.37	1.62
4	1.47	1.08	1.15	1.76
5	1.26	1.35	1.22	1.88
6	1.28	0.95	1.48	1.50

3. Turnovers of inventory in Grocery Stores per year for 1, rural; 2 city, individual owner; 3, chain; 4, super market.

TYPE OF STORES

1	2	3	4
7	12	23	31
5	14	20	33
6	11	19	30
7	14	22	30
8	13	20	29

4. The germination of two varieties of fuzzy cotton seed treated with various chemicals.[1]

Variety and Treatment	Replications 1	2	3	4
Delinted–11*a*				
Check	703	724	785	695
Ceresan	647	761	736	840
152–6–B	747	746	780	728
Spergon	637	767	808	591
Spergonex	794	781	839	756
Du Bay 1228R	819	839	760	760
Rogers Acala				
Check	719	728	600	667
Ceresan	714	754	646	691
154–6–B	749	809	736	745
Spergon	699	668	754	650
Spergonex	450	347	230	365
Du Bay 1228R	784	741	756	680

5. The effect of depth of planting with Ceresan-treated fuzzy and graded acid-delinted cotton seed.[1]

Variety and Treatment	Depth	Replications			
Acala 8	Inches	1	2	3	4
Fuzzy	1.00	687	802	814	840
Graded-del.	1.00	808	827	830	921
Fuzzy	1.75	692	799	711	737
Graded-del.	1.75	852	858	872	900
Fuzzy	2.50	270	435	480	295
Graded-del.	2.50	492	626	578	596
Fuzzy	3.25	230	209	113	138
Graded-del.	3.25	292	279	291	316

[1] Dr. K. Starr Chester and Dr. W. W. Ray, Oklahoma A. and M. College, Stillwater, Oklahoma.

CHAPTER 22

THE ANALYSIS OF COVARIANCE

In Chapters 6 and 7, we dealt with only one variable at a time. We measured its mean, standard deviation, and variation, and that was all. In Chapter 8 regression was introduced as a method of measuring the relation between two or more variables at once. The analysis of variance is similar to the methods presented in Chapters 6 and 7. It measures one variable at a time. *Covariance is a combination of variance and regression* whereby the variance of a variable may be measured after its relationship with a second variable has been removed by regression analysis. Covariance is, therefore, a much more complete and powerful method of statistical analysis than mere variance. It enables the statistician to bring to bear on small samples the combined analysis of regression, correlation, standard error of estimate, and variance. Variance analysis is based on the squared standard deviation. *Covariance is based on the squared standard error of estimate.* Variance is a one-dimensional measure. Covariance is a two or more dimensional analysis.

In the previous chapters the analysis of variance indicated clearly that the height of children in each of the six grades was a separate population. This is a somewhat surprising and disconcerting discovery after we have computed so many measures on the basis of height for all six grades as one population. In seeking the reason for this variation in height between grades of school children, we discover that the common changing variable is age. The children in each grade are about a year older than those in the next lower grade. Perhaps increasing age with accompanying growth is the disturbing factor. The analysis of

WORKSHEET NO. 116

ANALYSIS OF COVARIANCE OF HEIGHTS OF 48 STILLWATER GRADE SCHOOL CHILDREN AND THEIR ASSOCIATED AGES

Child	GRADES											
	1st		2d		3d		4th		5th		6th	
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
1	79	46	86	47	106	51	108	51	128	56	161	60
2	83	48	92	50	101	54	116	55	131	57	144	59
3	80	46	90	47	97	51	110	52	131	59	143	60
4	72	43	85	48	96	48	118	57	121	56	151	60
5	75	44	92	50	104	49	113	54	120	54	154	61
6	82	50	91	49	107	55	108	51	141	59	140	59
7	75	45	89	48	102	52	110	53	127	56	141	59
8	72	41	87	49	104	53	111	54	130	54	134	56
Sums	618	363	712	388	817	413	894	427	1,029	451	1,168	474
ΣX^2	47,872		63,420		83,547		100,018		132,657		171,060	
ΣY^2		16,627		18,828		21,361		22,821		25,451		28,100
ΣXY	28,121		34,546		42,212		47,768		58,071		69,273	

variance will not adequately solve this problem, but the analysis of covariance will. Covariance introduces the trend relationship with age which again unifies our sample of school children.

In Worksheet No. 116 our solution is set up in the form of six separate but related regression problems. In Fig. 110, the data for each one of these six problems are plotted and the regression line for each grade as finally computed in Worksheet No. 117 is drawn in. Our ultimate comparison of these six grades is based on the size of the squared standard error of estimate for each separate grade. Grades 2, 4, and 6 have very small standard errors of estimate. Grades 1, 3, and 5 have larger deviations. On Fig. 111, the Total Regression Line and the six separate Grade Regression lines are shown. The sum of the six separate standard errors of estimate from the six separate regression lines is less than the total standard of estimate from the Total Regression Line. The reason for this discrepancy which is 200.336 − 170.704, or 29.632, is that the sum of the standard errors from the six separate regression lines is less than the standard error of all the items from the total regression line. The items of each class are closer to their own class line than to the total line.[1]

The X-values in Worksheet No. 116 are the ages of the children in months. The Y-values are the associated heights of the same 48 children. Since in the analysis of covariance it is essential to compute the regression and standard error of estimate for the paired values of each class, the (1) sums (2) ΣX^2, sums of the X's squared (3) ΣY^2, sums of the Y's squared, and (4) ΣXY, sums of products of the paired X- and Y-values are accumulated at the bottom of their respective class columns.

The First Step is the accumulation of all the required totals from the class column sums at the bottom of Worksheet No. 116. The sum of the X's (ΣX) for each class column is squared, $(\Sigma X)^2$. The total is squared, as $(618)^2 = 381{,}924$, $(712)^2 = 508{,}944$, and so on for all ΣX and all ΣY. All of these totals are required in order to obtain the deviations from means.

[1] Adapted from *Statistical Methods*, Chapter 12, by Professor George W. Snedecor of Iowa State College, and are used with the kind permission of Professor Snedecor and the Collegiate Press, Inc., Ames, Iowa.

1st Step. EXPERIMENT TOTALS

$(\Sigma X)^2$	$(\Sigma Y)^2$	$(\Sigma X)(\Sigma Y)$
381,924	131,769	224,334
506,944	150,544	276,256
667,489	170,569	337,421
799,236	182,329	381,738
1,058,841	203,401	464,079
1,364,224	224,676	553,632
4,778,658	1,063,288	2,237,460

$\Sigma X = 5{,}238$
$\Sigma Y = 2{,}516$
$\Sigma X^2 = 598{,}574$
$\Sigma Y^2 = 133{,}188$
$\Sigma XY = 279{,}991$

1		2		3	
X^2	Y^2	X^2	Y^2	X^2	Y^2
6,241	2,116	7,396	2,209	11,236	2,601
6,889	2,304	8,464	2,500	10,201	2,916
6,400	2,116	8,100	2,209	9,409	2,601
5,184	1,849	7,225	2,304	9,216	2,304
5,625	1,936	8,464	2,500	10,816	2,401
6,724	2,500	8,281	2,401	11,449	3,025
5,625	2,025	7,921	2,304	10,404	2,704
5,184	1,681	7,569	2,401	10,816	2,809
47,872	16,627	63,420	18,828	83,547	21,361

4		5		6	
X^2	Y^2	X^2	Y^2	X^2	Y^2
11,664	2,601	16,384	3,136	25,921	3,600
13,456	3,025	17,161	3,249	20,736	3,481
12,100	2,704	17,161	3,481	20,449	3,600
13,924	3,249	14,641	3,136	22,801	3,600
12,789	2,916	14,400	2,916	23,716	3,721
11,664	2,601	19,881	3,481	19,600	3,481
12,100	2,809	16,129	3,136	19,881	3,481
12,321	2,916	16,900	2,916	17,956	3,136
100,018	22,821	132,657	25,451	171,060	28,100

XY	XY	XY	XY	XY	XY
3,634	4,042	5,406	5,508	7,168	9,660
3,984	4,600	5,454	6,380	7,467	8,496
3,680	4,230	4,947	5,720	7,729	8,580
3,096	4,080	4,608	6,726	6,776	9,060
3,300	4,600	5,096	6,102	6,480	9,394
4,100	4,459	5,885	5,508	8,319	8,260
3,375	4,272	5,304	5,830	7,112	8,319
2,952	4,263	5,512	5,994	7,020	7,504
28,121	34,546	42,212	47,768	58,071	69,273

This page of squares and products of X and Y includes the detailed computations required for Worksheet No. 116. They are

$$(79)^2 = 6{,}241,\ (83)^2 = 6{,}889,$$

and so on through all the X- and Y-values to the last Y-value, $(56)^2 = 3{,}136$. The first product of X and Y is $79 \times 46 = 3{,}634$ for the first X- and Y-values in the first grade to the last X- and Y-value of $134 \times 56 = 7{,}504$ in the last pair of items in the sixth grade. These sums of squares and products are all required for the computation of the regression lines and standard errors of estimate for the six separate classes, as developed in the second section of the analysis on page 597.

2d Step. Correction Terms

$$\frac{(\Sigma X)^2}{N} = \frac{(5{,}238)^2}{48} = \frac{27{,}436{,}644}{48} = 571{,}596.75$$

$$\frac{(\Sigma Y)^2}{N} = \frac{(2{,}516)^2}{48} = \frac{6{,}330{,}256}{48} = 131{,}880.333$$

$$\frac{\Sigma X \cdot \Sigma Y}{N} = \frac{(5{,}238) \cdot (2{,}516)}{48} = \frac{13{,}178{,}808}{48} = 274{,}558.5$$

3d Step. Sums of Squares

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 598{,}574 - 571{,}596.75 = 26{,}977.25$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 133{,}188 - 131{,}880.333 = 1{,}307.667$$

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X) \cdot (\Sigma Y)}{N} = 279{,}991 - 274{,}558.5 = 5{,}432.5$$

PLOTTED DATA AND REGRESSION LINE FOR EACH OF THE SIX GRADES ANALYZED IN COVARIANCE IN WORKSHEET NO. 117

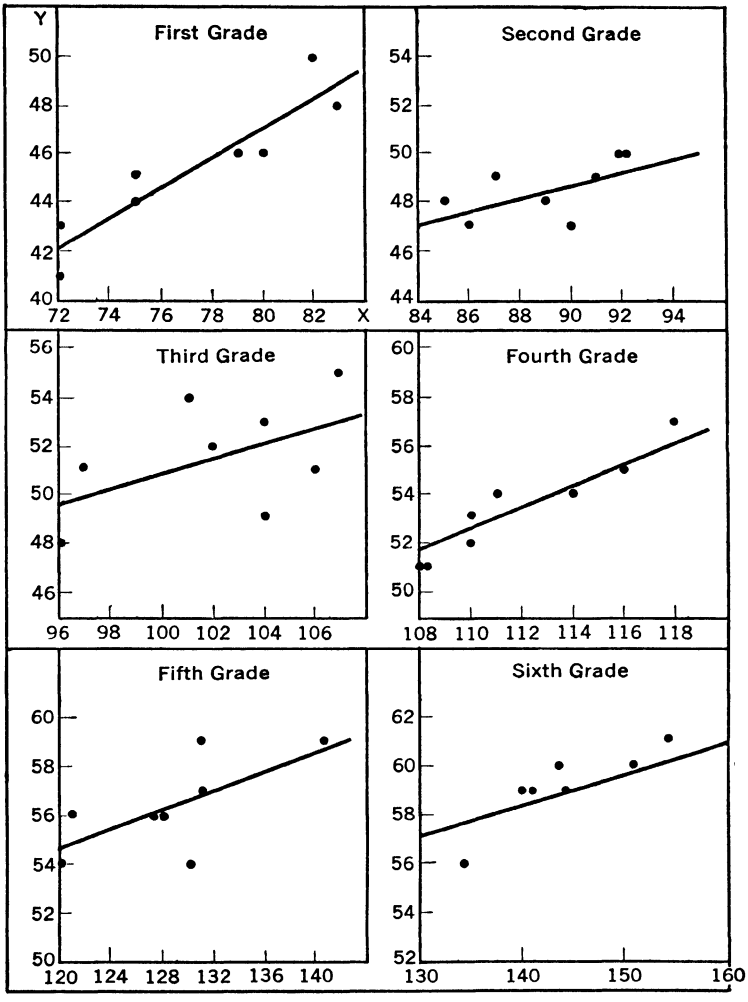


FIG. 110. Class regression lines in the analysis of covariance. (From Worksheet No. 117)

Since covariance is based on a comparison of the squared standard error of estimates of the samples, the student may get a clearer idea of these relationships by means of Fig. 110 and Fig. 111, which show the scatter of the items of each sample around its regression line. Fig. 111 shows the separate and the total regression lines.

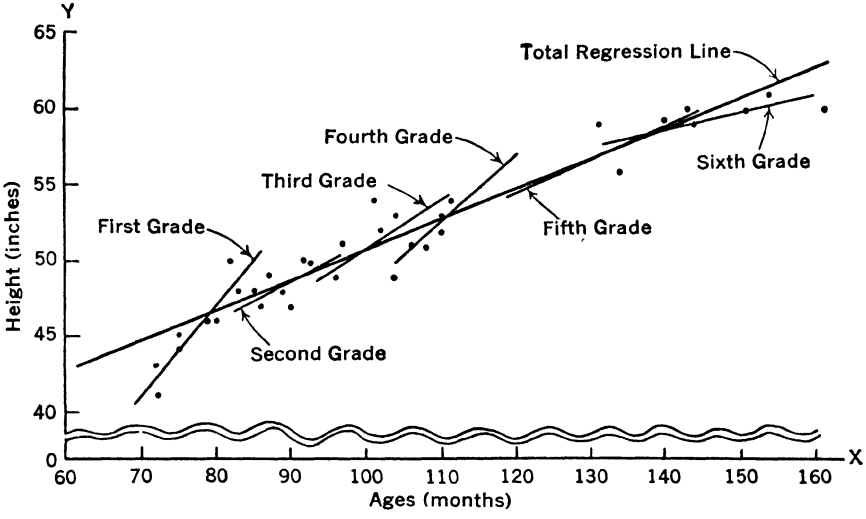


FIG. 111. Total regression line and separate class regression lines from the analysis of covariance of the ages (X), and heights (Y) of grade school children. (See Worksheets Nos. 116 and 117)

Regression Sums and Products

Age

$$\frac{4{,}778{,}658}{8} = 597{,}332.25 \qquad 597{,}332.25 - 571{,}596.75 = 25{,}735.5$$

Height

$$\frac{1{,}063{,}288}{8} = 132{,}911 \qquad 132{,}911 - 131{,}880.333 = 1{,}030.667$$

Products

$$\frac{2{,}237{,}460}{8} = 279{,}682.5 \qquad 279{,}682.5 - 274{,}558.5 = 5{,}124$$

4th Step. Squares of Errors of Estimate

$$\text{Total} \quad \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 1{,}307.667 - \frac{(5{,}432.5)^2}{26{,}977.25}$$

$$= 1{,}307{,}667 - \frac{29{,}512{,}056.25}{26{,}977.25} = 1{,}307.667 - 1{,}093.961$$

$$= 213.706$$

Within classes

$$\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 277.0 - \frac{(308.5)^2}{1{,}241.75}$$

$$= 277.0 - \frac{95{,}172.25}{1{,}241.75} = 277.0 - 76.643 = 200.336$$

The Second Step is the computation of the three *correction* values required to compute the deviations from the means for (1) ΣX^2, which are Σx^2; (2) for ΣY^2, which are Σy^2, and, for ΣXY, which are Σxy. This computation is the same as that used in computing the deviations from the mean in Chapters 8, 9, and 20.

The Third Step is the actual computation of the deviations from the means for Σx^2, Σy^2, and Σxy. In this case these deviations are $\Sigma x^2 = 26{,}977.25$, $\Sigma y^2 = 1{,}307{,}667$, and $\Sigma xy = 5{,}432.5$. These sums are the total variances for the entire group based on the assumption of the null-hypothesis of no differences between the classes.

REGRESSION SUMS AND PRODUCTS BETWEEN CLASSES

The second part of the Third Step is the computation of the deviations from the means for the *between* regression and standard error of estimate. These values are, for X (Age) = 25,735.5, for Y (Height) = 1,030.667, and for XY (Products) = 5,124; and are the portion of the total variances which are associated with differences between the classes.

The Fourth Step is the computation of the Squares of Errors of Estimate (1) for the total regression line and (2) for the within classes regression lines. The first figure is 213.706, which represents the sum of squares of the standard error of estimate for the

total regression line of all six classes thrown together as one sample. The second figure is 200.336 and represents the sum of all the squares of all six separate standard errors of estimate from the six separate class regression lines.

The purpose of this part of the analysis of covariance is the same as that of its counterpart in the analysis of variance which is to separate the total variance into two portions one of which is associated with variation *within* classes and the other of which is associated with variation or differences *between* the classes. In the case of the analysis of covariance the sum of squares of the Standard Error of Estimate is divided into two portions, that which is due to variation *within* classes and that which is due to variation *between* classes.

5th Step.

COVARIANCE AND TEST OF SIGNIFICANCE OF ADJUSTED CLASS MEANS

Source of Variation	Degrees of Freedom	Sums of Squares and Products			Errors of Estimate		
		Σx^2	Σxy	Σy^2	Sums of Squares	Degrees of Freedom	Mean Squares
Total	47	26,977.25	5,432.5	1,307.667	213.706	46	
Between Classes	5	25,735.5	5,124.0	1,030.667			
Within Classes	42	1,241.75	308 5	277.0	200 336	41	4.886
					13.470	5	2.694

$$F = \frac{2.694}{4.886} = .5514$$

6th Step. Correlation

Total $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{5{,}432.5}{\sqrt{26{,}977.25 \times 1{,}307.667}} = .9146$$

Between $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{5{,}124.0}{\sqrt{25{,}735.5 \times 1{,}030.667}} = .9945$$

Within $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{308.5}{\sqrt{1{,}241.75 \times 277.0}} = .5865$$

7th Step. Regression

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{5{,}432.5}{26{,}977.25} = .2014 \qquad \begin{matrix} \overline{X} = 109.125 \\ \overline{Y} = 52.417 \end{matrix}$$

$$a = \overline{Y} - b\overline{X}$$
$$= 52.417 - (.2014 \times 109.125)$$
$$= 52.417 - 21.777$$
$$= 30.64$$
$$Y = a + bX$$
$$Y = 30.64 + .2014X \quad \text{Total Regression Line}$$

The Fifth Step completes the computation of the mean squares of the standard error of estimate for both the *between* portion and the *within* portion of the total covariance and also the F-ratio between them. This F-value, in contrast with the F-value of variance, shows the significance of the variation between the Y-samples *after the regression with X in each class has been removed.* In this particular case F shows the significance of the variation in heights between classes *after the effect of age on heights has been computed out or removed.*

The Sixth Step is the computation of total correlation and the total regression line.

The conclusion of the analysis is that the highly significant differences among the six grades which were revealed by the analysis of variance are due almost entirely to differences in age. When the regression between age and height is computed and removed from the total variation the remaining variance between grades is quite insignificant.

At this point (end of Worksheet No. 116) the covariance analysis might be considered complete. F, r, b_{yx}, and S_y have all been computed. As far as the summary, or total, relations are concerned, it is complete. Additional information and confirmation concerning the correctness of our analysis and additional insight into the inter-class relations may be obtained from the same data by the methods illustrated in Worksheet No. 117. The covariance, regression, correlation, and standard error of each class may be obtained. The fact that the sum of the Σx, Σxy, and Σy^2 accumulated from the individual classes equals the Within Classes totals of Worksheet No. 117 proves the theoretical correctness of Worksheet No. 116 as well as the accuracy of the computations.

WORKSHEET NO. 117

1st Step. Separate Grade Regression Lines — Σx^2

No. 1

$$\Sigma x^2 = 47{,}872 - \frac{(618)^2}{8} = 47{,}872 - \frac{381{,}924}{8} = 47{,}872 - 47{,}740.5 = 131.5$$

No. 2

$$\Sigma x^2 = 63{,}420 - \frac{(712)^2}{8} = 63{,}420 - \frac{506{,}944}{8} = 63{,}420 - 63{,}368 = 52$$

No. 3

$$\Sigma x^2 = 83{,}547 - \frac{(817)^2}{8} = 83{,}547 - \frac{667{,}488}{8} = 83{,}547 - 83{,}436 = 111$$

No. 4

$$\Sigma x^2 = 100{,}018 - \frac{(894)^2}{8} = 100{,}018 - \frac{799{,}236}{8} = 100{,}018 - 99{,}904.5 = 113.5$$

No. 5

$$\Sigma x^2 = 132{,}657 = \frac{(1{,}029)^2}{8} = 132{,}657 - \frac{1{,}058{,}841}{8} = 132{,}657 - 132{,}355.125 = 301.875$$

No. 6

$$\Sigma x^2 = 171{,}060 - \frac{(1{,}168)^2}{8} = 171{,}060 - \frac{1{,}364{,}224}{8} = 171{,}060 - 170{,}528 = 532$$

2d Step. Sums of Grade y^2

No. 1

$$\Sigma y^2 = 16{,}627 - \frac{(363)^2}{8} = 16{,}627 - \frac{131{,}769}{8} = 16{,}627 - 16{,}471.125 = 155.875$$

No. 2

$$\Sigma y^2 = 18{,}828 - \frac{(388)^2}{8} = 18{,}828 - \frac{150{,}544}{8} = 18{,}828 - 18{,}818 = 10$$

No. 3

$$\Sigma y^2 = 21{,}361 - \frac{(413)^2}{8} = 21{,}361 - \frac{170{,}569}{8} = 21{,}361 - 21{,}321.125 = 39\ 875$$

No. 4

$$\Sigma y^2 = 22{,}821 - \frac{(427)^2}{8} = 22{,}821 - \frac{182{,}329}{8} = 22{,}821 - 22{,}791.125 = 29.875$$

No. 5

$$\Sigma y^2 = 25{,}451 - \frac{(451)^2}{8} = 25{,}451 - \frac{203{,}401}{8} = 25{,}451 - 25{,}425.125 = 25.875$$

No. 6

$$\Sigma y^2 = 28{,}100 - \frac{(474)^2}{8} = 28{,}100 - \frac{224{,}676}{8} = 28{,}100 - 28{,}084.5 = 15.5$$

3d Step. Sums of Grades xy

No. 1

$$\Sigma xy = 28{,}121 - \frac{618 \cdot 363}{8} = 28{,}121 - \frac{224{,}334}{8} = 28{,}121 - 28{,}041.75 = 79.25$$

No. 2

$$\Sigma xy = 34{,}546 - \frac{(712 \cdot 383)}{8} = 34{,}546 - \frac{276{,}256}{8} = 34{,}546 - 34{,}532 = 14$$

No. 3

$$\Sigma xy = 42{,}212 - \frac{(817 \cdot 413)}{8} = 42{,}212 - \frac{337{,}421}{8} = 42{,}212 - 42{,}177.625 = 34.375$$

No. 4

$$\Sigma xy = 47{,}768 - \frac{(894 \cdot 427)}{8} = 47{,}768 - \frac{381{,}738}{8} = 47{,}768 - 47{,}717.25 = 50.75$$

No. 5

$$\Sigma xy = 58{,}071 - \frac{(1029 \cdot 451)}{8} = 58{,}071 - \frac{464{,}079}{8} = 58{,}071 - 58{,}009.875 = 61.125$$

No. 6

$$\Sigma xy = 69{,}273 - \frac{(1168 \cdot 474)}{8} = 69{,}273 - \frac{553{,}632}{8} = 69{,}273 - 69{,}204 = 69$$

The four steps in Worksheet No. 117 are required to compute the regression coefficients, the coefficients of correlation and the standard errors of estimate for each of the six classes separately. The student will notice that the total values of Σx^2, Σxy, and Σy^2 at the bottom of the Fourth Step in Worksheet No. 117 are identical with those of the Within Classes at the bottom of the Fifth Step in Worksheet No. 116. These values are:

$$\Sigma x^2 = 1{,}241.875,\ \Sigma xy = 306.5, \text{ and } \Sigma y^2 = 277.0.$$

The sum of the parts equals the total. This result is to be expected and proves that the theory and methods of Worksheet No. 116 are correct.

The grade regression equations are required to plot the separate class regression line in Figs. 110 and 111.

4th Step.

REGRESSION AND CORRELATION DATA FOR EACH OF SIX GRADES

Grade	Degrees of Freedom	Sums of Squares of Products			Correlation Coefficient	Regression Coefficient	Errors of Estimate	
		Σx^2	Σxy	Σy^2			Sum of Squares	*df*
1	7	131.5	79.25	155.875	.5535	.60266	108.114	6
2	7	52.0	14.0	10.0	.6140	.26923	6.230	6
3	7	111.0	34.375	39.875	.5167	.30968	29.139	6
4	7	113.5	50.75	29.875	.8715	.44713	7.181	6
5	7	301.875	61.125	25.875	.6916	.20248	13.499	6
6	7	532.0	69.0	15.5	.7599	.12969	6.551	6
Sums	42	1,241,875	308.5	277.0	.5260	.24841	170.714	36

Grade Regression Equations

1. $Y = -1.18 + .60266X$
2. $Y = 24.54 + .26923X$
3. $Y = 20.00 + .30968X$
4. $Y = 3.41 + .44713X$
5. $Y = 30.33 + .20248X$
6. $Y = 40.32 + .12969X$

The net result of the work in Worksheets No. 116 and No. 117 is that we have identified and measured the factor responsible for the great variation in school children's heights and have shown that when this factor of age-growth is eliminated there is no other significant difference in the variation of those heights. The analysis of variance gave a highly significant F of 31.24. In co-

variance analysis, F is reduced to the non-significant value of .55. Covariance is necessary in the solution of those problems in which the variable we desire to measure is influenced by a secondary variable of wide fluctuations. A common example of this condition is the measurement of the gains of animals which are not of identical initial weight or age. A pig weighing 20 pounds will not ordinarily gain as much in a given time, week or month, as one weighing 40 or 60 pounds. To make accurate measurements of the significance of weight gains, the influence of the regression factor must be removed. Another example is the significance of gains in sales of various sizes. A firm with a capital of $10,000 could not expect to show the same gain in sales achieved by a corporation of $100,000 or $1,000,000 capital. The regression between sales and capital must be removed before the true significance of sales increase can be measured. The same method must be used in comparing any of the many factors dependent on size of population between cities, counties, or states.

ANALYSIS OF ERRORS OF ESTIMATE FROM AVERAGE REGRESSION WITHIN GROUPS (CLASSES)

Sources of Variation	Degrees of Freedom	Errors of Estimate	
		Sum of Squares	Mean Squares
1. Deviations from average (error) regression within lots, Fifth Step, Worksheet No. 116	41	200.336	4.886
2. Deviations from individual lots (grade) regressions, Fourth Step Worksheet No. 117	36	170.704	4.742
3. Differences among lot (grade) regressions	5	29.632	5.926

This difference between the Error of Estimate of the original within variation of 200.336 with 41 *df* and the totaled deviations from the six separate classes of 170.704 with 36 *df* is due to variance between the six regression lines and is due to Experimental Error. In a planned experiment based on a Latin Square some, perhaps most of it, could be removed.

COVARIANCE IN SOCIAL SCIENCES

The analysis of covariance is used primarily in biology, animal husbandry, agronomy, entomology, and related fields, but it may be used in economics, psychology, sociology, and education.

To answer the question, "Are there any significant differences in the number of eating and drinking places in (1) small, (2) medium, (3) large cities in Pennsylvania if population size is removed by regression computations? The answer appears in Worksheet No. 118.

WORKSHEET NO. 118

ANALYSIS OF COVARIANCE OF NUMBER OF EATING AND DRINKING PLACES (Y) AND POPULATION (X) OF SMALL, MEDIUM, AND LARGE CITIES IN PENNSYLVANIA, 1930

City	Small 2,500–5,000		Medium 15,000–25,000		Large 80,000 and up	
	X	Y	X	Y	X	Y
1	2.8	2	21.4	39	116.0	220
2	3.6	7	23.6	47	143.4	435
3	4.5	12	15.6	38	92.6	207
4	3.5	4	16.5	42	86.6	305
5	4.9	9	18.2	55	1,951.0	3,923
6	2.9	4	16.0	23	80.3	211
7	3.8	7	19.5	46	82.1	127
8	4.8	13	21.4	62	669.8	1,261
Sums	30.8	58	152.2	352	3,221.8	6,689
ΣX^2	123.20		2,955.98		4,318,315.41	
ΣY^2		528		16,452		17,414,199
ΣXY	243.5		6,822.4		8,659,241.0	

1st Step. Totals

$(\Sigma X)^2$	$(\Sigma Y)^2$	$(\Sigma X)(\Sigma Y)$
948.64	3,364	1,786.4
23,164.84	123,904	53,574.4
10,379,995.24	44,742,721	21,550,620.2
10,404,108.72	44,869,989	21,605,981.0

$\Sigma X = 3{,}404.8$
$\Sigma Y = 7{,}099.0$
$\Sigma X^2 = 4{,}321{,}394.59$
$\Sigma Y^2 = 17{,}431{,}179$
$\Sigma XY = 8{,}666{,}306.9$

2d Step. Correction Terms

$$\frac{(\Sigma X)^2}{N} = \frac{(3{,}404.8)^2}{24} = \frac{11{,}592{,}663.04}{24} = 483{,}027.62$$

$$\frac{(\Sigma Y)^2}{N} = \frac{(7{,}099)^2}{24} = \frac{50{,}395{,}801}{24} = 2.099.825.04$$

$$\frac{\Sigma X \cdot \Sigma Y}{N} = \frac{(3{,}404.8)(7{,}099)}{24} = \frac{24{,}170{,}675.2}{24} = 1.007{,}111.46$$

3d Step. Sums of Squares

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 4{,}321{,}394.59 - 483{,}027.62 = 3{,}838{,}366.97$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 17{,}431{,}179 - 2{,}099{,}825.04 = 15{,}341{,}353.96$$

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N} = 8{,}666{,}306.9 - 1{,}007{,}111.46$$

$$= 7{,}659{,}195.44$$

Regression Sums and Products

Population $\quad \frac{10{,}404{,}108.72}{8} = 1{,}300{,}513.59$

$$1{,}300{,}513.59 - 483{,}027.62 = 817{,}485.97$$

Eating and Drinking Places $\quad \frac{44{,}869{,}989}{8} = 5{,}608{,}748.62$

$$5{,}608{,}748.62 - 2{,}099{,}825.04 = 3{,}508{,}923.58$$

Products $\quad \frac{21{,}605{,}981.0}{8} = 2{,}700{,}747.62$

$$2{,}700{,}747.62 - 1{,}007{,}111.46 = 1{,}693{,}636.16$$

4th Step. Square of Errors of Estimate

Total $\quad \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 15{,}341{,}353.96 - \frac{(7{,}659{,}195.44)^2}{3{,}838{,}366.97}$

$$= 15{,}341{,}353.96 - 15{,}283{,}396 = 57{,}957.96$$

Within Classes

$$\Sigma y^2 = \frac{(\Sigma xy)^2}{\Sigma x^2} = 11{,}832{,}429 - \frac{(5{,}965{,}559)^2}{3{,}020{,}881} = 51{,}795.2$$

5th Step.

COVARIANCE AND TEST OF SIGNIFICANCE OF ADJUSTED CLASS MEANS

Source of Variance	Degrees of Freedom	Sums of Squares and Products			Errors of Estimate		
		Σx^2	Σxy	Σy^2	Sums of Squares	Degrees of Freedom	Mean Squares
Total	23	3,838,366.	7,659,195.	15,341,353.	57,957 96	22	
Between Cities	2	817,485.	1,693,636.	3,508,924.			
Within Cities	21	3,020,881	5,965,559	11,832,429	51,795.20	20	2,589.8
					6,162.76	2	3,081.4

$$F = \frac{3{,}081.4}{2{,}589.8} = 1.19 \text{ — Not significant}$$

VARIANCE ANALYSIS OF NUMBER OF EATING AND DRINKING PLACES IN PENNSYLVANIA CITIES, SMALL, MEDIUM, AND LARGE, 1930

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares
Total	23	15,341,353	
Between Cities	2	3,508,924	1,754,462
Within Cities	21	11,832,429	563,425

$$F = \frac{1{,}754{,}462}{563{,}425} = 3.117 \text{ — Almost significant}$$

This F-value of 3.117 for variance is within a few points of a significant difference between these cities, but the Covariance F-value of 1.19 is non-significant. This result shows that when we eliminate the difference in the number of eating and drinking places that is caused by population differences, Pennsylvania cities show no significant variation. In proportion to size, the small cities have about as many eating and drinking places as the larger cities.

LATIN SQUARE

As has already been indicated, the use of variance and covariance analysis is adapted primarily to planned experiments, especially in agriculture, although these methods are not at all

limited to that field. Whatever the field may be in which they are used, the best results are obtained from planned experiments. It is not likely that data picked up at random or which were assembled for other purposes will give dependable results or contribute to scientific discovery just because they are run through the variance or covariance mill. More and more researchers are coming to understand that to obtain dependable results of scientific value the data must be obtained from carefully prepared experiments set up in conformity with the most scientific methods available. Scientific knowledge doesn't "jes' grow"; it is wrought out by great labor.

	Columns							
Rows	1	2	3	4	5	6	7	8
1	C	B	A	D	E	H	F	G
2	D	A	F	C	B	G	E	H
3	B	C	D	A	G	E	H	F
4	E	H	G	F	C	B	A	D
5	A	G	B	H	F	C	D	E
6	H	E	C	G	D	F	B	A
7	F	D	E	B	H	A	G	C
8	G	F	H	E	A	D	C	B

FIG. 112. Idealized Latin Square

As was indicated in Worksheet No. 114 the data values in the rows as well as in the columns should be as free from outside and biased variation as possible. It is possible to eliminate all or most of this bias by computing the variance associated with the rows as was done in this case.

If, however, such bias can be prevented from entering the data at all, it is so much the better. Sources of such bias in field experiments in agriculture are unequal fertility of soil among the various plots, unequal drainage, more severe wind pressure on one side of the plants or plots, and other such inherent physical characteristics of the field, many of which cannot be wholly eliminated.

The Latin Square is a research device perfected by Fisher for counteracting or canceling out these diversities and variations. It is designed to have as many rows as columns, so that each treatment appears in various portions of the plot. If there are n rows, there are n columns. The number of separate treatments is, therefore, n^2, arranged so that each treatment appears once but only once in each column and in each row.

In this square there are eight treatments (fertilizers or time of planting or depth of planting or kind of seed or condition of seed or any other of eight different conditions) each of which appears once and once only in each row and in each column. The eight are so evenly distributed over the field that any difference in soil, drainage, wind, or other physical condition inherent in the field would affect all treatments in the same way. This arrangement makes the experimental base uniform. If one side of the plot is poorly drained, this bad condition is equally distributed among all varieties because some of all varieties are in the poorly drained land. If one side of the field is especially fertile, this beneficial condition is evenly distributed among all varieties, because some of all varieties are planted on the fertile side of the field. The good and the bad conditions of the physical environment are evenly distributed among all varieties because all varieties are scattered all over the plot or field. In both statistical theory and method this is an ideal procedure. All planned experiments as far as possible should follow this principle.

If it is impossible or inconvenient to set up a true Latin Square, similar results may be obtained from a Randomized Block System shown in Fig. 107.

Of course, a Latin Square may be of any size, but those from 3^2 to 6^2 are most frequently used.

MEANING AND USE OF COVARIANCE

1. Covariance analysis combines the methods of the analysis of variance with correlation and regression analysis. Variance analysis measures the degree of significance of the variation among two or more classes as to a *single variable* such as (1) gain in weight, or (2) growth height, (3) or incomes, (4) or sales, (5) or tensile strength, or any other variable which may change from class to class. In each, comparison is made between the several classes or groups for *only one variable at a time.*

Covariance analysis measures the degree of significance of the variation among two or more classes as to a single variable *after its relationship with a second variable has been computed by re-*

gression methods and removed from the deviations or variation of the first variable. In testing the effects which a given ration produces in the gain in weight of pigs, it is desirable to have all the pigs in the several pens of the same age, weight, breed, etc. Often this is impossible because of the lack of a sufficient number of identical pigs. This lack of identical pigs would be a serious handicap in the use of the analysis of variance but it need not at all hinder one in reaching dependable results with the analysis of covariance. All that is required in the covariance analysis is to compute the regression between original weight, (X_2), and gains, (X_1) or (Y), and remove this amount of variation from the total variation for each pen or group. With the effect of this outside variable removed a dependable comparison can be made between the remaining sums of variation among the several groups of pigs. The analysis of covariance enables one to use with dependable results much data which otherwise must be discarded, or if used, employed with a large degree of error.

2. Covariance analysis as here developed assumes straight line regression, which will not apply accurately in many cases. It is, however, possible to use curvilinear regression in the analysis of covariance just as it is in correlation analysis. This work is somewhat beyond the level of an elementary course but is available in more advanced texts.[1]

3. Covariance analysis would be useful in measuring (1) the degree of significance of variation in profits among grocery stores of varying sizes in different cities, (2) the incomes of machine shops of varying sizes in different communities, (3) the production of eggs by hens of varying ages for different breeds, (4) the resistances of materials of varying thicknesses in different uses, (5) the yields of wheat or other crop with varying degrees of rainfall on different soils.

These examples will indicate a few of the many fields in which this complex but powerful research device may be used profitably. It is so important in the analysis of small samples obtained from

[1] Fisher, R. A., *Statistical Methods for Research Workers*, Oliver & Boyd, Ltd., Edinburgh, 1936. Snedecor, George W., *Statistical Methods*, Chapter 13, Collegiate Press, Ames, Iowa, 1946.

planned experiments in agriculture and in the biological sciences that a brief presentation of its theory and methods is required in an elementary text.

SUMMARY

1. Covariance is a method of analysis which combines the methods of the analysis of variance with the methods of regression analysis and correlation analysis. It is one of the most powerful and exact of all the methods of statistical analysis and gives dependable results only when applied to data obtained from carefully planned experiments.

2. The analysis of variance is based on the squared standard deviation, but the analysis of covariance is based on the squared standard errors of estimate.

3. The analysis of covariance removes from the total variation of the factors to be compared that portion of the variation of each factor which is due to the influence or association of a second or outside factor.

4. The relationship between the two variables may be either straight line or curvilinear. The examples presented in this text are only straight line and are, therefore, not applicable to clearly curvilinear relationships.

5. Regression coefficients and coefficients of correlation may be computed for each class or section of paired variables as well as for the total group of data.

6. The Latin Square is a device used in planned experiments which permits each variable to be subjected to all positions and influences of random variations in order to nullify or eliminate such extraneous factors from the results. It is especially useful in agricultural experiments.

7. The analysis of covariance provides four results as follows:

a. The coefficients of correlation between the two variables in each block and the total of all the blocks.

b. The regression lines for each block and for the total data of all blocks combined.

c. The standard error of estimate for each block and for the total data of all blocks, and

d. The significance of the residual variations between all the blocks of data after the influence of the independent variable has been removed.

SELECTED REFERENCES

Brumbaugh, M. A., and Kellogg, L. S., *Business Statistics*, Chapter XXIX. Irwin, Inc., Chicago, 1941.

Croxton, F. E., and Cowden, D. J., *Applied General Statistics*, Chapter XIII. Prentice-Hall, Inc., New York, 1948.

Fisher, R. A., *Statistical Methods for Research Workers*, Chapter VIII. Oliver and Boyd, Ltd., Edinburgh, 1936.

Mills, F. C., *Statistical Methods, Revised*, Chapter XV. Henry Holt & Company, New York, 1938.

Pearson, F. A., and Bennett, K. R., *Statistical Methods Applied to Agricultural Economics*, Chapter XIX. John Wiley & Sons, New York, 1942.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, Chapter XII. McGraw-Hill Book Company, New York, 1940.

Snedecor, G. W., *Statistical Methods*, Chapters X, XI, XII. Collegiate Press, Inc., Ames, Iowa, 1946.

REVIEW QUESTIONS

1. Define Covariance.
2. What is the difference between variance and covariance?
3. What is the function of regression in covariance?
4. What is the use of the standard error of estimate in covariance?
5. What is the relation of the regression lines of the separate groups to the total regression line?
6. Why is the sum standard error of estimates of the several groups smaller than the standard error of estimate for the total regression line?
7. What is a Latin square and what is its value? How is one constructed?
8. What are Randomized Blocks? What is their use? How are they set up?
9. Why may F for covariance be smaller than F for variance?

EXERCISES

1. Original weight (X) and Gain (Y) During Feeding Experiment of 21 Hogs on 3 Rations.

Hog	Ration 1		Ration 2		Ration 3	
	Weight X	Gain Y	Weight X	Gain Y	Weight X	Gain Y
1	51	1.1	55	1.7	48	1.3
2	62	1.3	69	1.9	58	1.4
3	48	1.2	47	1.5	39	1.2
4	40	1.1	52	1.6	55	1.3
5	39	.9	59	1.6	61	1.4
6	50	1.2	70	1.8	45	1.2
7	58	1.3	63	1.7	53	1.3

2. Capital in millions (X) and Income in millions (Y) of (1) Railways, (2) Public Utilities, and (3) Industrial Corporations. Moody's Railways, 1941.

Company	Railways		Public Utilities		Industrial Corporations	
	Capital X	Income Y	Capital X	Income Y	Capital X	Income Y
1	91	6.7	31	1.1	154	10.3
2	44	4.8	67	2.1	24	2.0
3	10	.6	136	3.5	38	4.8
4	130	5.4	132	3.2	73	3.1
5	60	1.9	192	4.8	143	15.0
6	510	16.8	70	2.4	18	2.5
7	710	44.1	33	.9	11	2.3
8	41	1.3	80	2.5	256	21.8

3. Advertising Expense (X) and Sales (Y) in 5 Sales Territories, 1936–1941.

Year	1		2		3		4		5	
	Advertising Expense X	Sales % Expansion Y	Advertising Expense X	Sales % Expansion Y	Advertising Expense X	Sales % Expansion Y	Advertising Expense X	Sales % Expansion Y	Advertising Expense X	Sales % Expansion Y
1936	210	4.2	560	10 3	1,000	21.4	870	19 3	430	8.7
1937	325	6 1	480	9.2	850	18 7	680	14.7	455	9 2
1938	280	5.4	610	12.4	1,260	24.8	750	17.7	640	12 7
1939	362	7 9	655	14.1	1,420	27.1	880	21 1	580	11.9
1940	480	9.1	790	18 7	1,664	32 8	1,070	23.1	695	15.2
1941	375	7.4	595	11 6	975	19.9	820	18 8	475	10.4

Part Four

Organization of Statistical Problems and Reports

CHAPTER 23

ORGANIZATION OF A STATISTICAL PROBLEM

A statistical study is always set up to answer some question. The query must be one which can be answered by a quantitative measurement. How many or how much is the only reply a statistical study can give. Such questions as, What is courage? How courageous is he? What is virtue? How wise was Solomon? How great a general was Washington? cannot be answered by statistical methods. A question, to be suitable for a statistical problem, should be stated in a form in which it can be answered by absolute or relative numbers. Such questions are, What was the average yield of wheat per acre in Sedgwick County, Kansas, in 1951? What is the cost of producing coal in Harlan County, Kentucky? What was the birth rate in Quebec, Canada, in 1950?

A DEFINITE EXPLICIT QUESTION

Not only must a statistical problem be designed to give a numerical or quantitative answer to a query, but it must also be designed to answer explicitly a specific and delimited question. Such a question as, *What is the price of eggs?* is not a good question

to be the basis of a statistical investigation. It is much too indefinite and general. It gives rise to many other questions, such as, The wholesale or retail price? In Chicago, New York, or on the farm? fresh eggs or storage eggs? American-produced eggs or imported eggs? white or brown? by the dozen or by the pound? in July or December? and a score of other subsidiary queries which must all be decided upon before a statistical answer could be given. The problem to be solved by a statistical study should be definite as to (1) time (2) place or area (3) statistical unit and (4) methods or measures required.

Time. It is always necessary to be explicit as to whether the problem for solution is posed at a point in time or is to cover a specific period of time. The average and range of egg prices might be measured for retail stores as of a point or instant of time, perhaps as of 10 A.M. on Monday. Egg prices might be measured as to change over the period of a year or even several years. The two questions are quite different and require very diverse methods and answers.

Place or Area. Egg prices in Kansas will be lower than in New York City or even in Chicago, because of transportation and storage costs. Wheat yields in Texas may be quite different from those in Dakota or Saskatchewan. The cost of living may be much less in a rural area than in Boston or Washington. In any case the area to be covered by the study must be clearly stated and delimited.

Statistical Unit. One of the most important decisions to be made by the statistician is the unit of data to be secured and measured. At first thought this might seem to be an easy or even unnecessary step.

The unit of measurement applied to the data in any particular problem is the statistical unit. The units to be used in many studies are determined by custom or law and are a matter of general knowledge and are clearly understood. Such statistical units are the common physical units of measurement such as the pound, ton, inch, foot, yard, mile, hour, day, year, hectare, meter, kilowatt, horse power, and light year. In most cases to name the physical unit is sufficient explanation. There are some terms,

however, used with more than one meaning such as ton, bushel, and pound. A ton, for instance, may be a long ton of 2240 pounds or a short ton of 2,000 pounds. A bushel may be measured by cubic inches or by weight in pounds. For corn on the cob a bushel is 70 pounds, but for shelled corn it is 56 pounds. A pound may be avoirdupois, or 16 ounces or 7,000 grains, or it may be troy weight with twelve ounces and only 5760 grains. Whenever there is a diversity of physical units of the same name, the particular one used to measure the data at hand should be fully designated.

In many statistical studies no customary or legal unit is available. In such cases the statistician must arbitrarily create and define the unit to be used. This situation arises most frequently in studies in the social sciences. For instance, in investigating workmen's wages the question arises as to whether we shall measure wages by the year, month, day, or hour. There is much to be said for and against each unit of measurement. The amount of income received in a year sets definite limits to the level of living which may be maintained. A man who gets $5.00 a day but works every day in the year may have an income of $1800.00, but another who is paid $1.50 an hour but is idle much of the time may have an annual income of only $1200.00. On the other hand, hours are of equal length and easy to compare, but work days vary from 5 to 10 or more hours and work weeks from 30 to 66 or more hours, while work months vary from 20 or less to 30 days. Which unit of wage income should be used? The statistician must define the unit of data *before* he begins to collect his data. In each particular problem the unit of measurement must be uniform throughout the study.

Specific Definition of Unit. If a study is to be made of farm families, it is necessary to define clearly and explicitly what a farm family is. Is a farm family any family living on a farm whether it is engaged in agricultural work or not? Many families live on farms but work in cities, and have all the farm work done by tenants or hired managers and laborers. As far as economic production is concerned such a family is as urban as though it lived in a city hotel or apartment. The automobile, modern highways, and high city rents have caused many urban workers

to rent or buy homes in the country. Does living in the country make such a home a farm family? The statistician must decide before he begins collecting his data. If urban workers living in the country are defined as *not* farm families they will be automatically excluded from the data. If, on the other hand, all families living in villages or towns who go out daily to their farms to work and who are engaged in agriculture to the exclusion of urban labor are defined as farm families, they must be included in the data collected. The Federal census actually includes all families living on farms as farm families. The definition could be stated either way or in many other forms, but the point of emphasis is, that whatever the decision is, the definition *must be so crystal clear* that no one can misunderstand its meaning or delimitations. Full understanding of the field of data studied and accurate logical thinking are essential to the clear definition of a statistical unit.

Illustrative Problem Number One

How many inches high should the seats be for first-grade school children? We have all observed little children sitting in chairs made for adults and have noted their evident discomfort. Since physical comfort is essential to maximum mental concentration, it is essential that the seats in the school room be of the correct height. What height is correct? This is a question to be answered by a statistical study. It could be approached in two ways. Physicians and insurance companies have height-age-charts which show the expected average height, or normal height, of children for all ages from 1 year to 16 years. Since it is a matter of knowledge from previous studies that a seat for any person should be approximately one-fourth his height, one could divide the average height of six-year-old children as shown on the height-age-chart by four and quickly ascertain the height at which the first grade seats should be placed. This method is based on secondary data, which is data already in existence and collected for other purposes. The dependability of the results obtained from such studies would, of course, rest on the reliability of the secondary data used. The statistician should check all these points thoroughly, otherwise

his conclusions may be false. The detailed methods used in gathering and testing secondary data are treated in the following chapter. We are interested at this point only in the method to be used in setting up a statistical problem.

Use of Secondary Data. When any statistical question is posed, the logical first step is to discover whether an adequate answer or even a partial answer is already available. If it is, no further time, labor, or expense is required. In most cases, perhaps, some usable information will be discovered, but it may be insufficient for a conclusive answer. It may be out of date, in improper form, insufficient quantity, from the wrong area, indefinite, mixed with extraneous materials, or simply incorrect. Whenever adequate secondary data can be found they should be studied for their answer to the question or at least for their suggestions and implications as to what is lacking or required. It is a foolish waste of time and money to duplicate adequate answers to present questions.

Primary Data. If the data already available are insufficient to answer the question, new data must be procured. Primary data are data obtained first hand by the investigator for a specific problem. The first step in gathering primary data is to determine exactly what data are required for the purpose at hand, and to define the unit of data. In the case of the seats for first-grade children, the data should be the measures of the heights of a sufficient number of six-year-old children, or measurements in inches of the length of their legs from the floor to the knee. If the entire height of the child is the statistical unit, this figure must be divided by some number which will reduce total height to seat height. If such a dependable ratio is known, this method would be adequate, but if there is doubt as to the reliability of such a ratio, a better result would be obtained by measuring the actual length of the children's legs from the floor to the knee.

Selecting the Sample. If knee-height-in-inches is chosen as the statistical unit, the next question to be decided is how many children should be measured and which children. A detailed explanation of sampling is presented in Chapter 13. All that we wish to emphasize at this point is that in laying out a statistical

study, the question of sample size and characteristics must be decided upon *after* the statistical unit has been defined and *before* the data are collected. One might decide to use all the children of only one first-grade class. For several reasons this might be an inadequate sample. A small class of ten or twelve children probably would be insufficient. If this class contained only children who were large or small for their age, or were from a racial group shorter than the average or taller than the average of the total population or contained crippled children, the results would not be valid for general use. For securing a general rule for seat height the sample should cover a larger population and be based on random unbiased sampling. The end to be secured in sampling is to obtain the best possible representation of a large population by a small portion of that population. This might be done by taking at random five children out of ten, or perhaps more, first-grade rooms of a city school system.

Methods of Analysis. Most of this book, from Chapter 5 to Chapter 22, is devoted to theories and methods of analyzing statistical data. At this point it is not desirable to anticipate those detailed analyses, but merely to point out that the methods to be used will necessarily determine to some degree the type of statistical unit to be used and the type and quantity of data to be collected, and, therefore, that the methods of analysis to be employed in a particular study should be largely decided before any data are collected or the study begun. *One point which the student must fix clearly in mind is that all statistical studies should be planned carefully before the data are gathered or the analysis attempted.*

Summary of School Seat Study

1. We might select Cincinnati, Ohio, as a city of typical American population for the location of the study.
2. Our statistical unit would be the length of the legs of six-year-old children measured from the soles of their feet to their knee in inches.
3. Our sample would be five children taken at random from twenty first-grade rooms of six-year-old children by giving each

child in each room a number and then selecting by lot five numbers from each room, or 100 in all.

4. The analysis would consist of (1) computing a simple average for the 100 items and (2) the appropriate measure of the scatter of the items about their mean.

Illustrative Problem Number Two

A large nursery which supplies shade trees for St. Louis desires an answer to the question, What are the five most numerous varieties of shade trees in St. Louis and what percentage of the total of all shade trees does each of these five varieties compose?

The plan for the study is as follows:

1. Include in this study the entire metropolitan area including all adjacent suburbs.

2. Make the study in June when all trees are in full leaf.

3. Mark on a detailed map of the city every tenth block east and west beginning with the blocks bordering on the west side of Kings Highway and its most direct extensions to the north and south ends of the city. Mark on the map every tenth block north and south of Olive Street and its extension to the west in Washington Avenue from the Mississippi River to the west side of the metropolitan area. This would include one marked block in every 100 square blocks of the city.

4. Count all the shade trees on all four sides of each of these selected city blocks, recording the number of trees in each variety.

5. (1) Select from the tally sheets the five most numerous varieties. (2) Reduce the total for each variety to percentages of the grand total. (3) Multiply the grand total by 100 to obtain the approximate total of shade trees in the entire city.

The above organization outline is a rough statistical approach to a problem which does not require results of great accuracy. It does, however, raise several questions. (1) What are to be considered separate varieties of trees? (2) Would the soft maple and the hard maple be designated as two separate varieties? (3) Would all oak trees be considered as one variety, or would red oaks be segregated from white oaks, post oaks, and jack oaks?

(4) What lines, if any, would be drawn among the various types of pine trees, and cedar trees? (5) Would trees of all ages be counted or only trees 10 feet tall and over? (6) Or would it be of more use to the nursery to count only young trees less than 15 feet tall as indicative of present tastes in shade trees? (7) How far can one judge the age of trees by their height, since some are of very slow growth while others attain an equal height in about half the time? These and many other questions would have to be decided by the statistician before his research plan was perfected sufficiently to begin gathering data. One might even raise the question as to whether one square block in a hundred gave a sufficiently large sample to be representative of all parts of the city. Another vital query is whether the large commercial and industrial sections should be included in the study at all, since they would contain fewer trees and would be quite different from the residential areas. To exclude them would raise the difficult question as to when a section or block is residential or industrial or commercial. It is evident from these questions that those who plan such a study would have to give long and careful thought to its proper organization.

Illustrative Problem Number Three

As a means of aiding the Government to control prices more effectively after World War II, it is desired to create a weekly index number that will measure changes in the general price level.

The organization of this problem should include the following points:

1. Statistical Unit — the wholesale prices of commodities and services in dollars.
2. Time — measured each week as of a certain day, as Friday, or the average of the high and low of the week.
3. Area — the United States — a certain number of selected cities of various sizes, and certain central markets.
4. Method — a weighted geometric average of price relatives.

In teaching statistics it seems desirable to put everything first, but that is impossible. The technical terms and methods men-

tioned in this chapter must wait their appropriate place in the body of the text for explanation. All that we wish to emphasize at this point is that the organization of a statistical problem requires:

1. A thorough knowledge of the field in which the problem falls.
2. An adequate knowledge of statistical theory and methods, and
3. A thorough planning of the study as to all essential details and parts *before* the actual collection and analysis of data are begun.

In the case of such an index a large number of preliminary questions would have to be settled. Should one use retail or wholesale prices? How many separate prices should be included? Should the prices be weighted, and if so, with what weights? Should the prices of services be included? Should one take the (1) average price for the six days of the week, (2) the mean of the highest and lowest price, or (3) the price as of a point of time, perhaps 2 P.M. on Friday? What should be the standard or base period? These are merely samples of the many details which would have to be thought out in planning the study. Many times students have come to the author to help them clear up some point in a statistical problem which could not possibly be answered with the data in the student's possession because he had failed to include this particular point in the data when he gathered the information. Such a back-handed procedure is like putting a foundation under a house after the building is up and the plaster is on the walls. All information to be used in the study must be included in the original data.

Illustrative Problem Number Four

A farm experiment station wishes to test the effectiveness of two poisons sold in the market for killing cotton boll weevil. How should the experiment be set up?

1. The statistical unit would be a boll weevil.
2. The area would be a portion of an infected cotton field.

3. The time, preferably, August or a period when boll weevil are quite numerous.

4. The method would be rather complex. It would require a carefully planned series of replications, or blocks of infested cotton bolls, laid out in strips or squares in such an order that both poisons would be applied to blocks or areas that were equally infested with the boll weevil. If the weevil were only one-half as numerous where one poison was applied as they were where the other poison was used, the results would not be comparable. To make the degree of the effectiveness of the poisons clear there should be a check block where no poison was applied, but in which the weevil were equally numerous. The theory and methods of laying out such an experiment were explained in their proper place in Part III and are mentioned at this point only to emphasize the fact that the more exact and scientific are the results desired, the more exact and detailed must be the plans for the statistical study.

The number of illustrations could be continued without limit, but these are sufficient to suggest to the beginning student the great importance of carefully planning all statistical studies. In developing the various statistical methods in succeeding chapters, the particular types of data which should be treated by particular methods are pointed out and emphasized in the discussions. The student should note these points and in planning any statistical problem should select his methods with the same degree of deftness and skill with which the expert craftsman chooses and varies his tools according to the nature of his materials and the purpose of his operations. The scientific statistician is not one who rakes through the garbage pile of numbers and uses whatever he may chance to drag out. He is rather an architect, an engineer, who carefully designs a numerical structure for a specific purpose and selects his materials and methods to that definite end. It is well for the elementary student to begin with simple problems which lie within the range of his own experience.

SUMMARY

1. A statistical problem should be set up to answer a *specific* question.

2. It should be definite as to (1) time, either point or period of time, (2) place or area, (3) statistical unit, and (4) methods or measures required.

3. The *statistical unit* may be a common denominate number such as pounds, dollars, acres, days, etc., or it may be an especially or arbitrarily devised unit for a specific problem, as farm-family, ton-mile, work-day.

4. In setting up a statistical problem, the statistician should review what has already been done in the field to determine (1) whether further study is required, (2) what additional data are required, (3) whether modifications and restatements of the problem should be made or whether it should be further delimited or expanded.

5. Many questions should be asked and every phase of the problem should be explored and every important factor and variable should be clearly defined in its relation to the problem. A complete plan of procedure should be worked out before the data are collected or their analysis begun.

SELECTED REFERENCES

1. Brumbaugh, M. A. and Kellogg, L. S., *Business Statistics*, Chapters III, IV, Irwin, Inc., Chicago, 1941.

2. Croxton, F. E. and Cowden, D. J., *Applied General Statistics*, Chapter II, Prentice-Hall, Inc., New York, 1948.

3. Ezekiel, M., *Methods of Correlation Analysis*, Chapter XXIV, John Wiley & Sons, Inc., New York, 1941.

4. McCormick, T. C., *Elementary Social Statistics*, Chapter IV, McGraw-Hill Book Company, New York, 1941.

REVIEW QUESTIONS

1. What type of questions should a statistical study be expected to answer? Why?

2. How definite and delimited should be the question for a statistical study? Why?

3. What is a *statistical unit?* How is the statistical unit of a study determined? Explain.

4. What steps should be included in the preliminary work for a statistical study? Why? Explain.

5. What work should be done before the data are collected? Why?

EXERCISES

I. Outline a statistical study on one of the following questions:

1. How many pairs of shoes of each size should a shoe store carry in stock?
2. At what age on the average are automobiles junked?
3. To what degree do high school grades indicate college achievement?
4. Was R. E. Lee a greater general than U. S. Grant?
5. Two fertilizers are recommended for yard grass or lawn use. How may one determine whether one is superior to the other?
6. What is the average consumption of cigarettes for a given area, such as a city, county, or state?

CHAPTER 24

COLLECTING AND EDITING STATISTICAL DATA

The nature and complexity of the question to be answered determine the kind and quantity of data to be collected. Data are classified as primary and secondary. *Secondary data are those already in existence and which have been collected for some other purpose than the answering of the question at hand.* Examples of secondary data are the accounting records of a store or factory. These records are collected in the routine of business to enable the accounting department to post the proper charges and credits to the appropriate accounts in order to send out statements, make collections, and prepare profit and loss statements, and balance sheets. But after these data are once in existence, many of them can be used later by other persons or agencies for other purposes, such as estimations of average daily sales, average daily selling expenses, and even budgets for future months or years. When used by others than those who collect them they are secondary data. Other examples of secondary data are the extensive censuses and other reports of the United States Government. These include the decennial census, agricultural census, census of manufactures, treasury reports, reports on foreign and domestic commerce, and hundreds of other publications.

Primary data are those collected first hand by the statistician in order to answer the question at hand. All data are primary to the person or agency which first collects them.

WHAT HAS ALREADY BEEN DONE?

When a statistical question is raised, the first step in answering it is to determine whether an adequate reply is available. If it is, no further time or money need be spent. So much research has been done in so many fields during the past thirty years that many of the questions raised can be answered at once, if a person knows where to turn for the information. If a complete answer is not available at once, in many cases a sufficient amount of data can be discovered to indicate what kind of additional data are required and where they may be found. It is frequently discovered that a beginning has already been made and that all that is necessary is to complete it. Sometimes a new regrouping of the old data is necessary but again a more complete analysis of existing tabulations is all that is required. In any case it is advisable to exhaust the possibilities of using existing studies, secondary data, and earlier analysis before new data are collected. A review of previous studies is valuable in (1) furnishing useful data, (2) suggesting lines of attack on the problem, (3) revealing errors to be avoided and (4) indicating new or other sources of data to be used.

THE USE OF LIBRARIES

The most extensive single aggregate of secondary data is usually to be found in the libraries — city libraries, corporation or plant libraries, and college libraries. Such institutions are permanent depositories of governmental reports and publications as well as of the books and studies made by many other organizations, public and private. Since it is the business of a library staff to obtain all possible sources of data and other information for its patrons, libraries gradually collect vast stores of useful information which is only awaiting the hand of some one to use it. Large city libraries and the libraries of the large universities and land-grant colleges are especially well stocked with all state and Federal reports and publications. They also contain extensive files of metropolitan newspapers and trade magazines. The statistician, therefore, finds the libraries a rich source of data.

The novice in the use of library materials will have to depend much on the aid of library assistants and clerks. While many of the library personnel are efficient in locating book and statistical materials, they usually will not know all the sources in which the obscure but valuable bodies of data are hiding. They may lay before you a series of records or references, but it will still be your responsibility to locate the specific data required, and to pass judgment on their adequacy and validity. It is entirely too much to expect library staffs to do your research for you.

Since the research worker must depend on his own skill and judgment in large measure in locating and interpreting secondary data, he should, as soon as possible, become thoroughly familiar with all the library data possible in his and in related fields. The best way to do this is for him to secure, if possible, a permit to enter the *stacks* or book storage section of the library where he may examine and study at leisure all possible sources of data in his field. In an elementary text there is no space for lengthy references to individual and specific books and sources, but the student should have at his command the principal indexes of sources in all the major fields of study. A principal part of his preparation to do statistical work is a clear knowledge of these indexes and how to use them. They are the dictionaries of statistical studies or research.

UNITED STATES GOVERNMENT INDEXES

The largest collector and publisher of statistical data in the world is the Government of the United States. At various intervals the Government publishes a number of very important and inclusive indexes of all these sources of data. The principal indexes are:

1. *Catalogue of the Public Documents of the Congress and all Departments of the Government of the United States.* (Prepared under the supervision of the Superintendent of Documents)

This is the basic index and the student should become thoroughly familiar with its scope and general contents.

2. *United States Government Publications — A monthly catalogue.* This is the monthly supplement of *Catalogue of the Public Documents* listed above and is consolidated at the end of the year into the larger volume. It is the current list of Federal publications.

3. *Price Lists of Government Publications.* (Especially good for those persons ordering sources for themselves.)

A few of its many fields of listings are:

1	No. 10,	Laws
2	No. 11,	Foods and Cooking
3	No. 15,	Geological Surveys
4	No. 19,	Army and Militia
5	No. 20,	Public Domain
6	No. 21,	Fishes
7	No. 24,	Indians
8	No. 25,	Transportation
9	No. 28,	Finances
10	No. 31,	Education
11	No. 33,	Labor
12	No. 38,	Animal Industries
13	No. 62,	Commerce and Manufacturing
14	No. 71,	Children's Bureau

Dozens of other fields of data are included.

4. *Index to Publications of the United States Department of Agriculture.*

5. *Publications of the Bureau of Standards,* 1901–1925.

6. *Supplementary List of Publications of the Bureau of Standards,* 1925–1931.

7. *List of Selected Publications* (Issued by the Bureau of Foreign and Domestic Commerce)

8. *List of Available Publications — Miscellaneous Publications, No. 60.* (Always carries same number, No. 60)

9. *United States Tariff Commission.*

10. *Monthly Check List of State Publications; U.S. Library of Congress.*

These ten major United States Government indexes do not exhaust the list of Federal indexes or published sources of information. They are only the larger and more inclusive ones. The

student who is interested in securing data to answer a statistical question will find that librarians can secure for him many other minor sources of information as to the Government publications. *All of these publications are only indexes of sources.*

Principal General Federal Publications

In a general elementary statistics text only the most used and general Federal sources should be named specifically. These are:

1. *The Decennial Census* (Beginning in 1790 and available for each succeeding decade up to and including 1950).

Many libraries will have only the last three to five census reports, 1900, 1910, 1920, 1930, 1940, and 1950. Smaller libraries may have only one or two of the more recent reports. Few libraries will go back of 1860. These sources are invaluable for many purposes. The usual present subdivisions of the Decennial Census are:

1. Population (very detailed and complete)
2. Agriculture
3. Distribution
4. Retail Trade
5. Unemployment

2. *The Census of Agriculture* (Taken every ten years as of 1925, 1935, etc). The figures for the agricultural census are also included in the Decennial Census of 1920, 1930, 1940, and 1950.

3. *Census of Manufacutures* (biennial). It reports the number of establishments, employees, value of products and principal costs.

Special Federal Publications

1. *Statistical Abstract of the United States.*

This is a general summary of statistical information in many fields. The data are arranged in table form according to class intervals and time periods.

2. *Foreign Commerce Yearbook of the United States.*

A digest of all data on exports and imports of all commodities from and to the United States, from all foreign counties in terms of quantities and values.

3. *Foreign Commerce and Navigation of the United States.*

Data on exports and imports of all principal products from all states and ports of the United States to all foreign countries.

4. *Industrial Market Data Handbook of the United States.*

Data on all principal manufacturing, processing, and mineral production in the United States by states and counties, giving number of plants, wages, cost of materials, value of products and other related information.

5. *Consumers Market Data Handbook of the United States.*

Data by states, counties, and cities on 82 series ranging all the way from population, types and volume of business, employment, retail distribution, incomes, telephones, automobiles, electric equipment and many other items.

6. *Financial Statistics of State and Local Government.*

Data on state, county and city revenues, debts and expenditures from various sources (an excellent source of information).

7. *Budget of the United States.*

Statement of expenditures for two past fiscal years and estimates for coming fiscal year in detail by departments, bureaus and section for items.

8. *Handbook of Labor Statistics.*

Data on employment, wages, living conditions, housing, health, income, industrial relations, migration of labor, productivity and related questions.

9. *Survey of Current Business.*

Published monthly by Department of Commerce and giving data on industrial production, income, cost of living, prices, construction, retail and wholesale trade, employment, transportation and most other business activities.

10. *Market Research Sources* (biennial).

Published by Department of Commerce, giving quite complete information on all research projects and studies on marketing in the domestic market.

11. *Monthly Labor Review.*

Published monthly by the *Bureau of Labor Statistics*, giving data on employment, unemployment, wages, prices, levels of living and labor legislation.

12. *Agricultural Statistics.*

Published annually by the *Department of Agriculture* containing a wide variety of data on all phases of agricultural production and prices for the United States and foreign countries.

13. *Mineral Yearbook.*

Data on the production of all minerals in the United States and foreign countries in quantities and values in great detail.

14. *Statistics of Income.*

Published annually by the *U.S. Treasury Department*, containing extensive data on all types of income by sources and geographical divisions.

15. *Federal Reserve Bulletin.*

Published monthly by the *Federal Reserve Board*, containing a wide variety of data on all banking and monetary activities for the U.S. and foreign countries, besides many indexes on production, employment, wages, payrolls, and related business activities.

There are scores of other statistical reports, summaries, analyses, and yearbooks published by the various departments, bureaus and commissions of the United States. The number is too great to include specific reference to all of them here. The student will find them all listed in the general indexes given above.

NON-GOVERNMENT INDEXES

All libraries will contain some or all of the following indexes on general statistical sources and studies. They are:

1. *Readers Guide to Periodical Literature* (Published by the H. W. Wilson Co., 950 University Ave., N.Y.)

An index of 120 or more of the principal popular and scientific magazines, by author and title. An excellent service.

2. *Educational Index.*

Published five times a year. A general reference on all educational publications.

3. *Industrial Arts Index* (Chemistry, Physics, etc.)

An index of over 250 leading scientific periodicals. Published annually.

4. *Public Affairs Information Service,* H. W. Wilson Co.

Published annually by *Public Affairs Information Service,* N.Y.

5. *Agricultural Index,* H. W. Wilson Co.

A general index on all periodical literature, books and bulletins on public affairs.

6. *The New York Times Index.*

Published monthly. A complete index to all articles and news in the *New York Times.*

7. *Cumulative Book Index,* H. W. Wilson Co.

A world list of books in the English language.

8. *International Index to Periodicals,* H. W. Wilson Co.

Devoted primarily to the Humanities and Science. A good service.

9. *Trade and Professional Associations of the United States.* Government Printing Office.

SOURCES IN SPECIAL FIELDS

It is desirable even for beginning students to be acquainted with a few of the leading non-governmental sources in the principal fields of statistical studies.

Business

1. Moody's Investment Service, 65 Broadway, N.Y.

Data on banks, financial companies, industrial corporations, railways, public utilities, stocks and bonds.

2. Standard & Poor's Corporation, N.Y.

An extensive service supplying data on all types of business activities, industrial prices, production, wages, transportation, construction and other series.

Economics

1. *The American Economic Review,* Journal of the American Economic Association.

2. *The Quarterly Journal of Economics.*

Education

1. *The 1940 Mental Measurement Yearbook,* Buros, Oscar K., The Mental Measurements Yearbook, Highland Park, New York.

2. *Personnel Bibliography Index,* Cowley, W. H., Ohio State University, Columbus, Ohio.

3. *Educational and Psychological Measurements,* Quarterly Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

4. *Journal of Educational Research,* Barr, A. S., Jr., Editor, Box 21, Old Engineering Building, University of Wisconsin. (A monthly magazine established in 1920)

General References

The Annals of the American Academy of Political and Social Science. Concord, N.H.

Psychology

1. *Journal of Applied Psychology,* Porter, James P., Editor and Publisher, Ohio University, Athens, Ohio. (A bi-monthly journal)

Sociology

1. *The American Journal of Sociology,* University of Chicago, Chicago, Illinois, Herbert Blumer, Editor.

2. *Social Forces,* The University of North Carolina, Chapel Hill, N.C., H. W. Odum, Editor.

3. *The American Sociological Review,* Vassar College, Poughkeepsie, N.Y., J. K. Folsom, Editor.

4. *Sociology and Social Research*, University of Southern California, Los Angeles, California, E. S. Bogardus, Editor.

5. *Rural Sociology*, University of North Carolina, Raleigh, N.C., Carl C. Zimmerman, Editor.

Statistics

Journal of the American Statistical Association.

NON-LIBRARY SOURCES

The student should not think that all secondary data are in libraries. The world is full of them. Secondary data are everywhere. All accounting records of all business firms from the corner grocery store to the United States Steel Corporation, and General Motors are secondary data, to all persons other than the original collectors, and are frequently used by business firms for statistical purposes.

PUBLIC RECORDS

All the records in the many offices of the forty-eight state governments, including the secretary of state, the railway and public service commissions, the highway departments, the state public land commissions, the state treasurer's office, the department of education, the department of labor, the tax commissions, and all the many other state offices, bureaus, departments, and commissions contain large quantities of invaluable secondary data which usually can be made available to the researcher.

The records of all the offices of our thousands of American cities, towns, and villages from New York City and Chicago down to our smallest communities contain many valuable secondary data. Besides these there are the extensive records of the more than three thousand counties in the United States, including data on lands, deeds, taxes, mortgages, wills, births, deaths, marriages, divorces, suits at law, school records, roads, delinquency, local relief, elections, and other economic, social, and political affairs. In addition to all of these there are the data accumulated in the official records of tens of thousands of local townships and hun-

dreds of thousands of school districts, cemetery associations, churches, and newspaper offices.

These are all secondary data when used by other persons than those who originally collected them, and are valuable for statistical studies of many kinds.

CHECKING ACCURACY OF DATA

Since secondary data are second-hand data, collected by someone else in the past for other purposes, it is always necessary for the statistician who uses them to check them as to accuracy and homogeneity. The points to be covered in checking such data are:

1. Under what conditions were they collected and for what purpose? Did any special responsibility attach to the collector or the giver of the data causing him to use ordinary or special care in his work? Were the data gathered by those having first-hand knowledge of them, or were they based on hearsay or mere estimates? How dependable are they?

2. Are the data collected from several areas or over a period of time all expressed in the same statistical unit? Has there been a change from long tons to short tons, or from bushels to hundredweights, from household to family, or in the definition of farm, or business firm, or mile of highway? In other words, are we comparing and adding units of the same size, quality, and measurement throughout the study?

3. Are the data from the same area? This is of prime importance for all spatial data. Frequently a county is divided, two or more villages consolidated, and the boundaries of wards, school districts, precincts, and even states altered. Farms, plantations, and business corporations are often consolidated or divided. Are the data for the entire study taken for the same or for altered areas?

4. Are there omissions or duplications in the data? Were some of the records lost or were some of the facts never included in the records? Are the incomplete data available reasonably representative of the entire area or period, or do the gaps invalidate the measurements? In every case in which secondary data are

used, the statistician must satisfy himself as to their accuracy and if he uses them must make notations as to the breaks and discrepancies in them. The best way to check data for accuracy is to *compare several sources* if such are available. For instance, district school population figures may be compared with city directories and with the Decennial Census. Crop yields reported by various authorities may be checked against each other. Branch sales records may be checked against central office records. The parts may be compared with the whole.

TESTING THE VALIDITY OF DATA

Checking the accuracy of data means determining whether the facts are as represented. Testing the validity of data means determining whether the data are suitable for answering the questions for which they are being used. This point may be illustrated with a few cases.

Are the tax assessment records valid data for measuring the wealth of a country? No, for the following reasons:

1. Not all private property is listed on tax records. Some portions of it for one reason or another escape tabulation.
2. Not all property is listed at full value. An uncertain ratio or estimate would have to be used.
3. Not all property is listed at the same ratio of its true market value. For instance, it is not uncommon for one piece of property to remain for years at the same value on the tax records although its market value may have doubled or quadrupled while another piece has declined in value with no adjustment on the records.
4. In some states public utilities and railways are assessed by the state and do not appear on local tax records at all.
5. Much of the wealth of any community is publicly owned, such as, highways, streets, bridges, public buildings, libraries, schools, hospitals, etc.

These and other conditions make the tax data invalid as an accurate measurement of the country's wealth. The best that could be obtained from such data would be a very rough estimate.

Can the number of families in a community be determined from the number of school children? No, for the following reasons:

1. Not all families have children in school. Such families would be missed.
2. Not all families with children have the same number in school. Of two families with four children, one family might have only one child in school while the other might have all four children in school.
3. No valid comparison could be made between communities with such data because some communities are composed more largely than others of old people who have no children in school.

Are the charge account sales slips of a store valid data for the measurement of total sales? No, unless cash sales are added to them and the value of goods returned is subtracted.

Are the number of deaths occurring in a city during a year valid data on which to compute the city's death rate? No, for the following reasons:

1. Many of the deaths may be of persons dying in city hospitals who come from outside the city and are no part of the local population.
2. Some persons living in the city may die or be killed in highway or other accidents outside the city.

Only the deaths of persons living in the city at the beginning of the year would compose valid data for measuring the death rate of the area.

The student may think of many other illustrations of the validity of data or the lack of it. He should keep clearly in mind that in every case in which secondary data are used, the statistician must determine carefully their validity for the use at hand. Second-hand data are often no better than second-hand clothing or automobiles.

PRIMARY DATA

Although secondary data are usually cheaper and more immediately available, they are of little value if they are not adequate,

accurate, or valid for the question at hand, and primary data must be obtained to solve the problem. All that has been said in the previous chapter on the organization of a statistical problem applies with full force to the collection of primary data. The entire study should be carefully planned in detail, and accurate decisions made as to the statistical unit to be used, the time and the area to be covered, and the methods of analysis to be employed. Only after these plans have been perfected is it time to collect the data.

THE SCHEDULE

An essential part of the planning of the study is the organization of the schedule to be used in obtaining the data. *A statistical schedule is an organized logical sequence of questions for securing specific data for a given problem.*

Some Examples of Schedules

The typical modern schedule is a compact set of questions arranged in blocks so that they may be answered by (1) a check, $\surd$, or $\times$ (2) a letter, as M, F, (3) a word, as yes, no, in, out, all, part, none, (4) a number, as 1, 2, \$125, 5%, .06, $\frac{1}{2}$, (5) a name, as John Smith, Mary Brown, (6) a date, as December 7, 1941, January 1, 1942, or (7) at most by a very brief statement.

The questions are designed to require answers which may be stated as quantities, or specific classifications, and may be tabulated in frequency distributions or exclusive classes.

The actual size of Schedule I shown opposite is $8 \times 10\frac{1}{2}$ inches. It contains eleven blocks of questions including in all 196 possible answers. Of this number 2 are dates, 8 are names, 10 are letters, 26 are checks, 27 are words or brief phrases, and 123 are numbers. It is a good example of a brief schedule which exhibits a maximum of brevity, exactness, clearness, and objectivity.

Nature of Questions

The framing of the right kind of questions for a schedule requires much careful thought and considerable experience. Frequently the cooperation of two or more persons is desirable in

SCHEDULE I

VII. MONEY EARNINGS OF FAMILY FROM EMPLOYMENT OR BUSINESS OUTSIDE OF HOME OR AT HOME (during schedule year)

A	B	C	D	E	F	G	H	I
Members of family gainfully employed			OCCUPATIONS DURING YEAR					
Relationship	Line number from IV-A	Age	Kind of work (such as shoebinder, bookkeeper, merchant, author)	Nature of industry (such as cotton mill, bank, shoe store, independent)	Status of worker (E, S, O)	Rate of earnings per unit of time	Time employed (use same time unit as in G)	Total money earnings from employment or business
1						$ ..... per		$
2						..... per		
3						..... per		
4						..... per		
5						..... per		
6						..... per		
7						..... per		
8						..... per		
9						..... per		
10						TOTAL (1-9)		$ .....

VIII. OTHER MONEY INCOME (during schedule year)	Total for year
1 Roomers and boarders (gross)	$
2 From work in home not entered in VIII above (specify)	
3. Interest and dividends from bonds, stocks, bank accounts, trust funds, etc.	
4. Profits not included in VIII above, less expenses	
5. Rents from property, less expenses	
6. Pensions, annuities, benefits	

VIII. OTHER MONEY INCOME (during schedule year)—Continued	Total for year
7 Gifts in cash for current use from persons not members of economic family	$
8. Other money income	
9 TOTAL (1-8)	$
10. Losses from business not deducted above	
11 DIFFERENCE (9 minus 10)	$

X. VALUE OF FOOD HOME PRODUCED FOR FAMILY USE	Total for year
1 Fruits and vegetables	$
2 Eggs ..... doz Milk ..... qts.	
3 Poultry for meat No	
4. Other	
5. TOTAL	$
XI. TOTAL INCOME (VIII-10 plus IX-11 plus X-5)	$

12. Has family received direct relief (in cash or kind) during schedule year? a. ☐ Yes. b. ☐ No.
13. Has any member of family had work relief during schedule year? a. ☐ Cash b ☐ Kind c. ☐ None.

S. P. GOVERNMENT PRINTING OFFICE 0—2020

Form 101

CONFIDENTIAL.—The information requested in this schedule is strictly confidential. Giving it is voluntary. It will not be seen by any except sworn agents of the cooperating agencies and will not be available for taxation purposes.

I. YEAR COVERED BY SCHEDULE

Twelve months beginning ....., 1935
and ending ....., 193..

II. FAMILY COMPOSITION DURING SCHEDULE YEAR

A	B	C	D	E
Members of economic family (all persons sharing family income, including those temporarily away from home)	Sex	Age at last birthday	Number of weeks during year	
			At home	Away from home
1 Husband	M			
2. Wife	F			
Other members of family (give relationship)				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

If any member of family died during schedule year, check number in front of name.

U. S. DEPARTMENT OF AGRICULTURE
BUREAU OF HOME ECONOMICS
IN COOPERATION WITH
NATIONAL RESOURCES COMMITTEE
WORKS PROGRESS ADMINISTRATION
AND
DEPARTMENT OF LABOR
WASHINGTON

STUDY OF CONSUMER PURCHASES

A FEDERAL WORKS PROJECT

FAMILY SCHEDULE—TOWN OR VILLAGE

III. OTHER MEMBERS OF HOUSEHOLD
(Persons living in home for one week or longer during schedule year)

A	B	C
STATUS (paid help, roomers, guests, children boarding)	Number of persons	Total number of weeks in household
1 Sons and daughters boarding and rooming at home (specify age and sex)		
2 Other roomers with board		
3. Roomers without board		
4. Boarders without room		
5. Tourists and transients		
6 Guests (overnight or longer)		
7 Paid help living in		

IV HOME OWNERSHIP

1 Number of months during schedule year living.
a As renter .....
b As owner
1st home ..... 2d home .....

	1st home	2d home
If as owner: 2. Monthly rental value	$	$
3. Was home mortgaged (or being purchased on land contract)	a. ☐ Yes. b ☐ No.	a. ☐ Yes. b. ☐ No.
4. If mortgaged, interest on mortgage (or land contract) for months occupied	$	$

Number in economic family .....
Ins. .....
Cir .....
Occ. .....
Agent .....
Date of interview .....
Code No .....
Family schedule No .....
Expenditure schedule No. .....
Town or village .....
E D. .....
State .....

V RESIDENCE IN THIS TOWN OR VILLAGE

For how many months of schedule year did the family live in this town or village? ..... mos.

VI. LIVING QUARTERS OCCUPIED at date of interview

1 Did family occupy these living quarters at end of schedule year?
a. ☐ Yes. b ☐ No
2. Does family ☐ own or ☐ rent these living quarters?
3 Monthly rent, $..... ....., if renter.
4. Type of living quarters (check one):

One family house.
a. ☐ Detached.
b. ☐ Attached.

Two family house
c. ☐ Side by side.
d. ☐ Two decker.

Apartment in building for:
e. ☐ Three families.
f. ☐ Four families.
g. ☐ Five or more families.

Dwelling unit, business building:
h ☐

Room or rooms.
i ☐ With another family.
j ☐ In rooming house.

Other
k. ☐

VII. COLOR

a ☐ White b ☐ Negro.

0—2020

framing questions. If a question is so stated that it has the same meaning to several persons and elicits the same answer from all of them, it is likely to be a good question, definite, clear and explicit.

Question 2 of block VI in Schedule I (Family Schedule — Town or Village) above is an excellent question. It reads, "Does family .... own or .... rent these living quarters?" It may be answered by one check mark. *Own* or *rent* are perfectly clear and alternate concepts to even persons of less than ordinary intelligence. This question is not likely to be misunderstood or to be given an incorrect or confusing answer. If, however, it were stated, "How does the family obtain its living quarters?," the evident alternative would not be so clear. Some might answer, "We own them," others might reply, "We rent them," but someone might give the indefinite reply, "The place belongs to my father," or "It belongs to the company and is included in our salary," or "We have an option on it." The statistician might have difficulty with the last three answers and, perhaps, would have to go back for more exact information.

This question would be still worse if it were stated, "Explain how you get your living quarters." The answer might be, "A real estate company really owns the house, but we are paying it out like rent." This answer is equivalent to "own," but he has cluttered up the problem with a lot of useless and extraneous information which has no bearing on the main issue of "own" or "rent."

Many of the other questions in this schedule could be so stated that several sentences would be required to answer them. This inexactness of question and answer is the thing the makers of schedules must avoid.

Schedule II is $13\frac{1}{2} \times 19\frac{1}{2}$ inches and contains 16 blocks of questions with a possible more than 400 answers. Only one-fourth of the schedule is shown below. Many others, both shorter and longer and better and worse, are available in corporation, research, and government offices. These are sufficient to stress the types of questions and kind of organization which should be followed.

SCHEDULE II

6

16-826-6

SAMPLE

CONFIDENTIAL CENSUS REPORT.—Your report is required by Act of Congress. This Act also makes it unlawful for the Bureau to disclose any facts, including [illegible] population, resources, and business activities. Year [illegible]

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
WASHINGTON
SIXTEENTH CENSUS OF THE UNITED STATES: 1940

FARM AND RANCH SCHEDULE

(Including Special Agricultural Operations)

Inventory Items, April 1, 1940 — Production Items, Calendar Year 1939

I.—FARM OPERATOR, APRIL 1, 1940

SAMPLE COPY

Code

1. Name of person
2. Address (Street or Route No.) (Post office) (State)
3. Age — Note.—Report age and color or race as shown on Population Schedule. — A 1
4. Color or race [Place a check (√) in proper block]

11. White, including Mexican	51. Negro	52. Indian	53. Chinese	54. Japanese	55. Other (specify)

— A 2

5. Do you reside on this farm? (Yes or No) — 3

II.—FARM TENURE, APRIL 1, 1940

6. If you rent any farm land from others or manage any farm land for others, give names and addresses of the owners of the land and indicate for each owner whether a corporation:

Name Incorporated? (Yes or No)
Address (Street or Route No.) (Post office) (State)
Name Incorporated? (Yes or No)
Address (Street or Route No.) (Post office) (State)

7. What does the landlord furnish as his share in the operation of this farm?

(a) Work animals (All, Part, or None)
(b) Tractor power (All, Part, or None)
(c) Fertilizer (All, ½, ¼, etc., or None)
(d) Seed (All, ½, ¼, etc., or None)
(e) Other (Name and give share)

8. What did you agree to pay as rent for the year?

If cash, give total amount, explaining for what paid, as $200 for entire farm, $20 for 10 acres pasture, $75 for 50 acres hay land, etc.
If share of crops or of animal increase, give kind and report share in fractions, as ½ corn, ⅓ young animals, ¼ cotton, etc.
If definite quantity, give kind and report in bushels, pounds, etc., as 600 bushels wheat, 4 bales cotton, etc.
If other rent, specify, as upkeep of farm, lump of hundred, etc.

— B 1

9. Do you operate this farm for others as hired manager? (Yes or No)

10. How many acres in this farm do you own? Acres — A 4
11. How many acres in this farm do you rent from others? Acres
The sum of Questions 10 and 11 must equal Question 12, unless operated by a hired manager.

III.—FARM ACREAGE, APRIL 1, 1940

The sum of Questions 13 to 16, inclusive, must equal Question 12.

12. Total number of acres in this farm Acres — 5 Q 1

Omit from this report any land now rented to or cropped by others. Include all cutlying or separate fields, meadows, pastures, woodland, or vacant lands operated by the person whose name is given under Question 1, whether owned, rented from others, or managed for others. Borrow a separate schedule for each tenant, renter, or cropper

Suggestion To Enumerator.—Before proceeding with the Farm Acreage, it may be helpful to fill in the information for Columns 1 and 2 on the reverse side of schedule.

13. Land from which crops were harvested in 1939 Acres — A 6

Include all field crops, all hay and wild hay cut, and gardens, orchards, and vineyards. Do not count the same land twice, even though two crops were harvested from it.

14. Land from which no crop was harvested in 1939 because of crop failure or destruction Acres — 7

List "planted crops which failed" under Supplemental Information.

15. Cropland lying idle all of 1939 or land in summer fallow in 1939 (omit crop failure and land pastured) Acres — 8
16. Land used ONLY for pasture or grazing in 1939 which could be plowed and used for crops without additional clearing, drainage, or irrigation (omit cropland harvested and hay cut) Acres — 9

17. Woodland in this farm Acres — 10

Include as woodland all farm wood lots or timber tracts, natural or planted, and cutover land with young growth, which has or will have value as wood or timber. Omit shelterbelts and woody shrubs.

18. All other land now in this farm Acres — 11

Include rough pasture land (not woodland pastured), all waste land, also house yards, barnyards, land lots, lanes, roads, etc.

IV.—VALUES, APRIL 1, 1940

Code A

19. Total value of this farm (land and buildings), including farm land and buildings rented from others $ (Omit cents) — 12

Give the amount for which this farm (Question 12) would sell, including the land actually operated by you as owner, renter, or manager, and all farm buildings and improvements. Omit farm implements and machinery, livestock, and land rented by you to tenants or croppers.

If you own a part, but not all, of this farm—
(a) How much of the total amount under Question 19 represents the value of the land and buildings owned by you? $ (Omit cents) — 13

20. Value of all buildings on this farm $ (Omit cents) — 14

Included in answer to Question 19

21. Value of farm implements and machinery used in operating this farm, including automobiles, tractors, motortrucks, and trailers (present market value) $ (Omit cents) — 15

Include all farm implements; tools; wagons, harnesses, dairy equipment, cotton gins; threshing machines; combines, apparatus for making cider, grape juice, and syrup and for drying fruits; and all other farm machinery. Omit residential units and machines also permanently installed irrigation and drainage equipment.

V.—FARM MORTGAGE DEBT AND FARM TAXES

If you own all or part of this farm—

B

22. Was there any mortgage debt on the land and buildings so owned on April 1, 1940? (Yes or No) — 2

23. Total amount of outstanding mortgage debt on such land and buildings $ (Omit cents) — 3

24. What was the annual rate (contract rate) of interest on the first mortgage debt? (Report fractions) (Percent) — 4

If you own all or part of this farm—

25. Give amount of taxes levied in 1939 on the REAL ESTATE of this farm owned by you on April 1, 1940 (include buildings and other improvements, but do not include taxes levied by drainage or irrigation districts) $ (Omit cents) — 5

26. Give amount of taxes levied in 1939 on PERSONAL property owned by you on this farm (include livestock, machinery, etc., but do not include automobile taxes, fees, or licenses) $ (Omit cents) — 6

VI.—OTHER LAND OWNED, APRIL 1, 1940

27. Do you own any land in addition to that shown under Question 10? (Do not include residential and nonfarm building sites) (Yes or No) — 7

28. How many acres of such land are rented or are to be rented this year to others (including that rented to croppers)? Acres — 8

VII.—WORK OFF FARM AND TIMES ON THIS FARM

C

29. How many days in 1939 did you work for pay or income off the farm you operated? Days (If no days, write "None") — 1

OF THESE, HOW MANY WERE SPENT—
(a) At farm work, not connected with your farm? Days — 2
(b) At nonfarm jobs (including road work and relief or "made" work), businesses, or professions? Days — 3

Note.—Total of (a) and (b) must equal total days worked off the farm. — 4

For (b), specify principal nonfarm occupation and industry
Nonfarm occupation (Miner, salesman, proprietor, farmhand, mechanic, hired hand, etc.) Industry (Coal mine, cotton mill, grocery store, public school, road work, railroad, W.P.A., C.C.C., N.Y.A., etc.)

30. Years when you began to operate THIS farm (Year) — 5

VIII.—IRRIGATION

A

31. Land from which irrigated crops were harvested in 1939 Acres — 16

Do not count land twice, even though two crops were harvested from it.

32. Land irrigated in 1939 and used ONLY for grazing or pasture Acres — 17

33. Irrigation enterprise supplying water

If water is not supplied by an irrigation enterprise, explain, as pumped from own well, stream diversion, or city water, etc.

NAME
ADDRESS (Street or Route No.) (Post office) (State)

Testing Schedules

After the schedule has been carefully organized and checked for omissions in the office and before large quantities are printed it is well to send an interviewer out into the field with a dozen or so to test it out under actual field conditions. It is very difficult to think out beforehand every point that should be covered. If anything has been overlooked this field test will likely bring it to light in the filling out of a few schedules. Actual field conditions frequently bring to view questions which it is difficult or impossible to anticipate. Such revisions as these practical tests reveal can then be made before the final schedules are printed.

Instructions for Enumerators

A complete set of detailed instructions should be prepared for the enumerators. It is always well to train schedule takers before they begin their work, but in the field questions sometimes arise for which they have forgotten the answers, if they ever knew them. If they have a complete set of instructions covering every detail of data to be obtained and the various circumstances under which they may have to get the information, they can do a much better job. These instructions should usually be revised after the questionnaire or schedule has been tested under field conditions.

THE QUESTIONNAIRE

Schedules are usually filled in by trained interviewers who visit the persons or firms from whom information is desired and who are, therefore, able to explain questions which are not clear to the persons giving the information. This use of trained workers eliminates much confusion and most inaccuracies from the data. Many times, however, the expense of using trained interviewers to secure answers to the schedule questions through personal visits is too large to be justified by the probable value of the study. In such cases a less expensive method must be used. The cheaper method is usually the questionnaire. Sometimes, however, a large number of schedules may be collected at a small cost. This is usually true if the traveling costs are not large. The costs can

be computed only after the problem has been planned and the location and the quantity of the data required are known.

A questionnaire is a related series of questions, usually prepared on a paper with intervening spaces for answers, which is mailed to a number of persons from whom information is desired. Return postage is provided for the reply. A trained interviewer usually can get answers to the questions on a schedule, but *a questionnaire is its own and only advocate.* It must be so simple and crystal clear that the person who is supposed to answer it can understand it with a minumum of effort and time. Most of the persons to whom questionnaires are sent are busy people who will throw any questionnaire into a waste basket if it takes more than a few minutes to answer or is not perfectly clear, or pries into their private business. Questionnaires should, therefore, always be as brief as possible, clear and reasonable.

Statisticians who contemplate using questionnaires should consider well their relationship to the persons who are supposed to answer the questions. These relationships will usually fall in one of five classes as follows:

1. *Those who are under your authority*, as salesmen under a sales manager, soldiers under a commanding officer, branches or subsidiaries under a central business office, teachers under a dean or principal, or even employees under a foreman.
2. *Those who are under some obligation*, as debtors to a creditor, tenants to a landlord, clients of an insurance firm, etc.
3. *Those who expect to make a profit out of the relationship*, as patrons of a store, a factory, a dispenser of services, an advertising agency, or an exporter.
4. *Those who expect to or wish to render some service*, such as schools, churches, lodges, civic organizations, citizens, or patrons.
5. *Those whom you expect merely to accommodate you as a free favor*, such as unknown persons or organizations which are in no way under obligation to you or whom you do not expect to pay or reward for their favors.

These five classes are listed in the diminishing likelihood of your receiving a reply, or an adequate reply, from your questionnaire.

Statisticians should usually refrain from sending questionnaires to persons or firms in the fifth class.

Sometimes a questionnaire has an advantage over a schedule in that it is more secret and impersonal, and, therefore, the subject is more willing to give the desired information.

EDITING THE SCHEDULES

After the filled-in schedules or questionnaires have been received they should be edited before the data are used. Editing is a preliminary critical evaluation and checking process. It may be done by one or more persons, depending on the size of the project and the amount of technical knowledge and skill required. In small studies one person usually does it all. The points to be covered in editing usually are as follows:

1. *Check the schedule for any omissions.* Sometimes the person filling in the schedule will accidentally overlook some question. At other times he may not know the answer and merely leaves the space vacant. In either case the data should be secured by a second effort, if possible, or the schedule, or that portion of it, thrown out.

2. *Checking for conflicts and contradictions.* Frequently two or more questions are so related and interdependent that one is a check on the accuracy of the other. In making schedules and questionnaires such internal automatic checks should be carefully designed. This may be done by checking parts against a total; two dates or a date and age against each other, or a class against a quantity, as in a schedule dated January 1, 1943:

(1) *Age* — 40 years, 7 month.
(2) Date of Birth, 4 *day* June *Mo.* 1912 *year.*

In this case it is evident that an error has been made in either (1) Age, or (2) Date of Birth. The person is either 30 years old or was born in 1902. The schedule or questionnaire should be sent back for correction or discarded.

3. *Verify answers marked with check marks or letters* as √, X, F, M, in, out, and the like. For instance, if the name of the per-

son is John Smith and *Sex* is marked *F* a check or correction should be made.

4. *Look for errors in computations or subject matter.* 65 + 12 may be given as 87. Chicago may be listed as in *Lake* County instead of *Cook* County, and Kansas City, Missouri, listed as Kansas City, Kansas.

5. *Check for conformity among schedules.* If the question of "Amount of arable land on farm?" is answered for most schedules as including plowable pastures but excluding woodlands, it should be so answered on all schedules as to reach comparable results in the study. If the question, "Number of minor children?" is answered to include both boys and girls under 21 years in some of the schedules it should not be answered to exclude girls over 18 in other schedules.

After the preliminary checking is completed all schedules answered properly should be so marked and segregated and the incomplete schedules sent back for correction or eliminated from the study.

CODING DATA

In modern statistical work in which machines are largely used, much of the data is frequently coded to shorten and facilitate tabulation. For instance, a family of (1) husband (2) wife, and (3) several children may be tabulated as 1, 2, 3, 4, 5, 6, etc. or 11, 12, 13, 14, 15, etc. Dates such as Jan. 14, 1943, Dec. 24, 1944 may be coded as 01, 14, 43 and 12, 24, 44, etc. Large numbers may be shortened by cutting off (.000) or (.000,000). Any coding that will change written statements or words to numbers, or shorten the processes of tabulation and computation should be employed. Numbers expressed as decimal fractions, such as .0122, .0097, .0107, etc., may be coded by multiplying by 10,000, or by moving the decimal point four places to the right, making the numbers, in this case, whole numbers, 122, 97, and 107. They can be further coded by subtracting a common number from each item as

$$\begin{aligned} 122 - 100 &= 22 \\ 97 - 100 &= -3 \\ 107 - 100 &= 7 \end{aligned}$$

By the two methods of coding the original number, decimal fractions have been reduced to the small numbers, 22, − 3, and 7 which are easy to manipulate. This type of coding required *decoding* after the statistic or statistics desired are computed. In this case any average would have to be divided by 10,000 to *counteract* original *multiplication* by 10,000, and 100 would have to be *added* to counteract the subtraction of 100 earlier.

Sometimes when the tabulations are to be made without machines, the data are transferred from the original schedules to smaller cards. This has the double advantage (1) of saving the original schedules from being quickly worn out by too frequent handling and (2) of reducing the required figures and other information to smaller space which materially facilitates their tabulation. In transferring the original data to the cards, they can usually be condensed or abbreviated.

SUMMARY

1. Secondary data are those already in existence and which have been collected for some other purpose than the answering of the question at hand.

2. Primary data are those collected first hand by the statistician in order to answer the question at hand.

3. The most effective use of libraries for statistical work is made by employing the fullest and most accurate indexes of statistical sources.

4. The statistician should become thoroughly familiar with all principal sources and subject matter in his special field of work. Access to the *stacks* is an excellent means of doing this.

5. All data should be checked for accuracy. This is especially true of secondary data.

6. Secondary data must be tested for validity. Do these data apply to this problem? is the question.

7. A statistical schedule is an organized logical sequence of questions for securing specific data for a given problem.

8. A questionnaire is a related series of questions, usually prepared on a paper with intervening spaces for answers which is mailed to a number of persons from whom information is desired.

9. Questionnaires should be sent only to those from whom answers may reasonably be expected.

10. Coding data is reducing them to briefer and more simple form for purposes of tabulation.

SELECTED REFERENCES

1. Brown, Lyndon O., *Market Research and Analysis*, The Ronald Press Co., New York, 1937.
2. Brumbaugh, M. A. and Kellogg, L. S., *Business Statistics*, Chapters VI, VII, Irwin, Inc., Chicago, 1941.
3. Croxton, F. E. and Cowden, D. J., *Applied General Statistics*, Chapter II, Prentice-Hall, Inc., New York, 1948.

REVIEW QUESTIONS

1. What determines the kind of data sought for a statistical problem? Why?
2. Define *secondary data.* Indicate sources of such data not mentioned in the text.
3. What are the functions of libraries as related to secondary data and research?
4. What are the best ways in which to use libraries for statistical research?
5. Name the ten most important United States indexes and indicate the general nature of the data each covers.
6. What Federal censuses are taken and when?
7. Name and give a general idea of the contents of seven special U.S. statistical publications.
8. Name and give a general idea of the contents of seven non-government indexes.
9. What are some of the best sources of information in each of the following fields:

    1. Business
    2. Economics
    3. Education
    4. Psychology
    5. Sociology
    6. Theoretical Statistics

10. What are the various public records which may be useful to the statistician?
11. How is the accuracy of data checked? Explain in detail.
12. What is *testing the validity* of data? Explain in detail.
13. What are *primary data?*
14. What is a schedule? Explain in detail.
15. How should the questions be drawn up? Why?
16. How should schedules be checked and tested before they are put into final form?

17. What instructions should accompany the set of schedules for the schedule taker and how should he be trained for his work?

18. What is a questionnaire? How does it differ from a schedule?

19. What are the relative advantages of the schedule as compared with the questionnaire?

20. Explain the purpose and methods of coding data.

EXERCISES

1. Prepare a schedule and instruction sheet for the schedule takers to secure data to answer the question, "What percentage of the total automobiles in your community are Fords, Chevrolets, Plymouths, Nashes, or Pontiacs?"

2. Prepare a schedule and instruction sheet for the schedule takers to secure data to answer the question, "What percentage of the boys in your community who smoke cigarettes learned to smoke in the 9th, 10th, 11th, or 12th grades of high school?"

3. Prepare a questionnaire to secure data to answer the question, ' In what grade in school do boys and girls (1) learn to dance, (2) have their first dates, and (3) join their first social cliques or societies?"

4. Prepare a questionnaire to secure data to answer the question, "What percentage of the children by sex, age, and distance from school would prefer to change from 'daylight-saving-time' to regular time during part of the school year, and in what month would they prefer to make the change?"

5. Set up a plan for a statistical study to determine the average assessed value of single family residences in the poorest one-third, the middle one-third, and the richest one-third of your city.

CHAPTER 25

PREPARATION OF STATISTICAL REPORTS

A statistical study is usually undertaken to answer some specific question. When the study is completed it is necessary to state the results in a form that is appropriate to the subject matter and to the persons who are to use it. It has often been taken for granted (1) either that anyone could prepare an adequate statistical report (2) or that this was a matter beyond the province of the statistician. Neither assumption is correct. It is difficult to prepare a good and adequate statistical write-up of a research problem as many students and statisticians who have attempted it have discovered. One large business firm offered a large salary to anyone who could train their personnel to prepare good reports. With the greatly increased use of statistics in the affairs of daily life the preparation of clear, concise, logical, and effective statistical reports is a necessary practical accomplishment for many persons. The student should be trained in this skill early in his college course. It will pay him large dividends in his research work and in his business career. Teachers should not neglect it.

TYPES OF STATISTICAL REPORTS

The type of report in which a statistical study terminates depends upon (1) the subject matter and (2) the persons for whom it is prepared. Three general types cover the field fairly well. They are (1) the scientific report, (2) the business report, and (3) the popular report. There are, of course, wide variations

within each of these classes and many degrees of combinations of two or more of them.

Form of Scientific Report

A strictly scientific report is designed primarily for scientists. It takes for granted on the part of the reader a sufficient knowledge of the technical terms employed. It assumes on the part of the reader an adequate understanding of the scientific methods and mathematical computations involved. The scientific report, therefore, is brief and strikes directly to the heart of the problem. It omits all unnecessary detailed and routine computations and states the results of the study in summary form. Only new or unusual techniques are explained in detail.

Outline of Scientific Study. Although there is some variety in the form of scientific reports, in the main they follow rather closely the following outline:

1. General title or statement of the problem
2. Author or authors
3. Institution or auspices under which the study is conducted
4. Introduction
5. Statement of methods
6. Body of report
7. Summary
8. Bibliography
9. Footnotes

1. The *title* of a scientific study should be an exact and full statement of the question studied. It should be framed to reveal information and not as a catch phrase. The following are fair examples:

> "The Quantitative Estimation of Relative Concentrations of the Viruses of Ordinary and Yellow Tobacco Mosaics and Tomato Spotted Wilt by the Primary Lesion Method."[1]

[1] By Rupert J. Best, from the Wait Agricultural Research Institute, University of Adelaide, South Australia.

or

"Land Utilization and Classification in New York and Its Relation to Roads, Electricity and Reforestation." [1]

The title should make clear at once the nature and scope of the study.

2. The *names of authors* are usually stated without title or comment.

3. The *university, foundation, society, or other institution* under whose authority, aid, or auspices the study is made should be concisely stated after the author's name at the beginning of the study, as is indicated in footnotes on this and the previous pages.

4. The *introduction* is a brief statement of from one to three pages setting forth (1) the origin of the problem, (2) its development up to the present time and (3) the proposed advances or additions to be made in the present study. The length of the introduction will depend on the age and nature of the problem but should be no longer than is needed to orient the reader in the subject.

5. The *statement of methods* should be as brief as possible to cover the essential attacks on the problem. These should include (1) methods of collecting data, (2) experimental design, (3) sampling, (4) coding, (5) statistical measures used, and (6) all extraordinary or unusual procedures or methods. It should make perfectly clear the nature and limitations of the data on which the study is based.

6. The *body of the report* is the logical systematic statement and explanation of the several points or principles covered in the study and of the relationships which have been discovered to obtain among them. It will usually contain:

a. Graphic presentation of data and the relationships which exist among them.

b. Tables of ratios, coefficients, equations, indexes and errors which express in mathematical terms the relations among the data and variables of the study.

[1] By T. E. LaMont, of the New York State Land Survey, *Bulletin 372*, March, 1937, New York State College of Agriculture.

c. Explanation and discussion of the several relationships discovered as revealed in the graphs, coefficients, ratios and equations.

The body of the report will be developed in some logical sequence. Usually it will begin with the several details of specific data and move toward a summary conclusion through the inductive method. The strength, weakness, force, and limitations of each point and step should be fully stated and its bearing on the entire problem pointed out. Statistical methods are merely devices for making logical correct thinking more exact and dependable.

7. The *summary* is a recapitulation in numerical order of the more significant points and conclusions of the study. It is a final brief statement of the results of the research. It should be clear, to the point, and free from all overstatement. It is unethical for a statistician to claim in his conclusions more than his data and study substantiate.

8. *A complete bibliography* or roll of sources used or referred to in the study should appear at the end of the report in alphabetical order. This part of the report is very important. Those who examine your work may wish to pursue the study further and to go back to the sources of the material. The bibliography at the end of this text is an example.

9. *Adequate footnotes* should appear at the bottom of every page or at the end of the study giving full recognition to every source used.

Perhaps the best practice for the student who wishes to prepare a scientific report is to examine a number of such reports in the document room of the library. All college libraries and most of the larger city libraries will be well supplied with such reports. These documents will vary slightly from one field to another. The student should study the best reports in the field in which he is interested.

Business Reports

Statistical reports in business may be roughly divided into three classes: (1) research problems in business, (2) special managerial reports, and (3) routine reports. Although each of these

three classes is a distinct type, there is a wide variation within each class.

Research Reports in Business. Such reports are the most pretentious and formal in this field of statistics. They are, in fact, closely related to the scientific reports discussed above and may be quite long and represent a large amount of careful study. They range all the way from monographs and small books to reports containing ten or a dozen pages. Among the most common fields touched by such studies are (1) location of plants or industries, (2) sources of raw materials, (3) power, (4) labor problems, (5) working conditions, (6) population problems, (7) consumer purchasing power, (8) costs and standards of living, (9) trade territories, (10) market analysis, (11) transportation, (12) banking and credit problems, (13) foreign trade, and (14) international exchange. These are, of course, only general fields and thousands of research problems are always in process on special delimited problems in these general areas.

Outline of Business Report

1. Title
2. Letter of submittal
3. Index of general contents
4. Index of graphs and tables
5. Body of report
6. Summary
7. Bibliography
8. Appendixes

1. The *title* of the report with the name of the author should appear on the cover page. The title should convey a clear idea of the general contents of the report. It is well for it to answer the three questions: What? Where? and When? as

Report on the Industrial and Economic Situation in Chile, Nov., 1927, by W. F. Vaughn Scott

or

Relation of Central Market Prices of Strawberries to Production Planning, 1932, by Orville J. Hall, Fayetteville, Arkansas.

2. A *letter of submittal* is not necessary to all business reports but is proper in reports which have been authorized by some authority such as a government body or official, a board of directors, corporation president, or department head.

The following letter of submittal by the National Resources Board is an example:

National Resources Board
Interior Building
Washington
November 28, 1934

The President
The White House
Washington, D.C.

My Dear Mr. President:

We have the honor to transmit herewith the report of the National Resources Board with the supporting documents used in its preparation.

This report carries this significance: That it is the first attempt in our national history to make an inventory of our national assets and of the problems related thereto. Moreover, for the first time it draws together the foresight of various planning agencies of the Federal Government and suggests a method for future cooperation.

The members of the Board have not all had an opportunity to give full consideration to all of the points involved. However, they unanimously agree in principle and desire to indicate to you their belief in the great importance of this study and of initiating steps toward the accomplishment of the broad program herein outlined.

Very respectfully yours,

(s) Harold L. Ickes

Harold L. Ickes
Secretary of the Interior,
Chairman

This is the submittal letter of a pretentious 455 page study which involved the entire United States and the labor of many persons. The submittal letter of many reports might be much shorter and less formal.

3. The *index of the general contents* of a business report should

be quite complete. If the report is worthwhile it may be used a great deal. Ready reference to its contents is an economy of time.

4. The *index to the graphs and tables* should be complete. Since much of its value will consist of its tables and illustrative graphs as summary material, one should be able to locate any desired material with the least delay.

5. The *body of the report* should be organized in a logical, coherent, progressive form. Beginning with the original data and basic principle each step should lead logically and naturally to the next up to the final summary.

Such outlines may fall in the form of (1) demand, (2) supply, (3) price, and (4) costs; or (1) fixed capital costs, (2) other fixed costs, taxes, etc., (3) direct material costs, (4) direct labor costs, (5) service costs, and (6) transportation costs; or (1) population by classes, (2) income by classes, (3) sales outlets; or (1) labor turnover, (2) working conditions, (3) wage rates, (4) union activities, and (5) management and labor cooperation; or any other set of logical and coherent relationships which the problem and data suggest or determine.

The body of the report *is* the report and it must stand or fall upon the soundness of its logical relations and statistical analysis. All this should be taken into careful consideration in the preliminary planning and organization of the project before the data are collected. A good business report cannot be made from a conglomerate of irrelevant data and conflicting logical relationships. The objective of the study must be clearly in view from the beginning.

6. The *summary* of a business report is as necessary as that of any scientific study. It should state in brief numerical order the points and principles established or suggested by the study. It should be unbiased. It should point out the weak as well as the strong points revealed in the research and should lay the foundation for a wise and dependable judgment.

7. The *bibliography* should be complete and given in alphabetical form by either authors or subjects or both for purposes of easy reference.

8. The *appendixes* should include the original data, the prin-

cipal worksheets and computations, and other material on which the tables, graphs, coefficients and percentages in the body of the report are based. In this respect a business report should be somewhat different from the report prepared for scientists. The report prepared for scientists usually omits most or all of those detailed materials which it is proper and necessary to include in the appendix of the business report. Since the business report is frequently used as a basis for making far-reaching managerial decisions, it should make available the full evidence on which such decisions can be made with the best measure of success.

Special Managerial Reports. Frequently a board of directors, a business executive, a staff member, or committee will require a relatively small amount of statistical information and analysis for some immediate managerial problem. In such cases some statistician or statistical clerk, or perhaps some accountant, bookkeeper, filing clerk, or other capable employee will be requested to get the data and make the computations.

The usual characteristics of such reports are that they are brief and specific. They are sometimes, however, difficult to make. The very fact that they must be so exact and specific makes it necessary to discover the exact and detailed information, which is often not easy to do. The fact that such requests are usually made on short notice to meet some immediate need makes the problem doubly difficult. The board of directors may require some information for their meeting tomorrow or even for today. The statistical department of a business should anticipate such needs as far as possible and have their materials organized in such form that such necessary requests can be met with a minimum of delay and effort.

Such special reports are usually quite informal and simply drawn. They depend on secondary data which may be recombined or reanalyzed for some immediate specific purpose. A firm with a modern mechanized statistical department may only need to resort and tabulate a portion of their punched cards to supply the information. Sometimes, however, outside and additional data must be obtained and analyzed. Such hurried reports are necessarily as brief and informal as possible. This requirement does

not eliminate the necessity for source references, bibliography, indexes of subject matter and graphs and tables. The essential parts of the report are present but are usually briefer and less formal.

ROUTINE BUSINESS REPORTS

Such reports consist of filling in prepared blanks. They may be made daily, weekly, or occasionally. The information may be that obtained by a traveling salesman during his day's visits on the firm's customers as to orders, stocks, business conditions, customer demands, credits, and such matters. The routine report may be that of the branch manager to the central office, or of a department head to the manager.

When the printed form for such reports is made, the nature and limits of the reports are definitely fixed. The making of the original forms is, therefore, the principal concern of the statistician in such cases. The forms for all routine reports should be most carefully designed by the business executives who require the information with the aid of the firm's statistician. The basic principle to follow is to reduce such forms and reports to the minimum required for adequate management decisions and to make one report supply as many needs as can be done efficiently. Since the requirements of such reports range through an infinite variety, they must be evolved by each firm through tests and experience and revisions.

Popular Statistical Reports

There has been for many years a continual increase in the use of popular statistical reports. The reason for this is the general desire to economize time and space. Statistical tables, graphs, ratios, and coefficients are the short-hand of presenting ideas. Editors, writers, teachers, advertisers, lecturers, and others who appeal to the popular audience have discovered that they can *get over* to readers a larger amount of information more effectively and more quickly by statistical devices than in any other way. It is probable that this trend will increase.

It is much more difficult to advise students how to prepare a

popular statistical report than to make a scientific or business report, because the former has a much wider range and variety. A popular report is less subject to exact and formal methods. It partakes of the nature of art and journalism. It is a device for presenting simple general ideas instead of making nice and exact distinctions. For this reason some statisticians abhor it and attempt to discourage it. Such efforts are fruitless and a wiser plan is to educate and direct statisticians to more sane and ethical standards. Besides, much of the popular presentation of ideas by statistical devices is already highly constructive and beneficial.

GENERAL RULES FOR PREPARING POPULAR STATISTICAL REPORTS

1. Such reports range all the way from (1) literary articles with a flavor of statistics to (2) a statistical report in a literary dress. The first type is frequently found in newspaper articles or even in editorials dealing with social, economic, or scientific materials. The second type is found in such popular business magazines as *Business Week, The Nation's Business, Bradstreet's Review, The Santa Fé Magazine.*

2. The amount of statistical analysis and the amount of literary dress combined in any given case should depend on the (1) purpose of the article and (2) the type of reader to whom it is directed. If it is an economic subject and the readers have been accustomed to thinking in terms of prices, production, and profits, more statistics may be included than would be suitable for political, social, or historical subjects for readers who think in terms of general ideas. But even general ideas can be enlivened and colored by well chosen data and analysis.

3. *Scientific and business reports should always be written in terms of the third person only,* but sometimes popular statistical reports are more effective in terms of the editorial, *we,* or even occasionally the first person singular, especially if the author is well known and is recognized as an authority. It is, however, usually more effective for most popular reports to use the third person as the general rule, with an occasional lapse to the first

person for special appeal or emphasis. Popular reports cannot entirely escape the functions of the editor or orator. It is a popular appeal based on data.

4. Popular statistical reports of good quality are more difficult to prepare than rigid and exact scientific and business reports and should not be attempted by beginners until they have made a special study of such reports prepared by specialists in that field and have acquired some literary training and effectiveness in expression.

The most effective training for the preparation of any type of statistical report is the careful examination and detailed study of a number of the best existing reports in that particular field. The list of reports at the end of this chapter is designed to suggest to the teacher and student specific examples for such study.

It is suggested that at least one semester of a student's work in statistics should close with a term problem in which the student makes a statistical analysis of a degree of difficulty suitable to his ability and writes it up in a formal adequate report. Such training under the conscientious direction of the teacher is a most valuable part of a student's training in statistics.

SELECTED REFERENCES

SCIENTIFIC REPORTS

1. Agricultural Experiment Station Bulletins [1]
   1) California Agricultural Experiment Station, Technical Bulletins
   2) Illinois Agricultural Experiment Station, Technical Bulletins
   3) Iowa Agricultural Experiment Station, Technical Bulletins
   4) New York Agricultural Experiment Station, Technical Bulletins

2. *Smithsonian Miscellaneous Collections*, published by the Smithsonian Institution, Washington, D.C. (These are usually excellent modes of scientific reports and may be found in the document rooms of college libraries and large city libraries. They range from Volume 1, published in 1862, to Volume 102, published September 1, 1942.)

[1] The technical bulletins of all the State College Agricultural Experiment Stations are similar and any of them will serve approximately as well as those of the four large states named above. Some of them can be found in any college library or large city library. Most of them are fair examples of a scientific report.

BUSINESS REPORTS

1. *American Economic Review,* published by The American Economic Association, Washington, D.C.
2. *Canadian Journal of Economics and Political Science,* The University of Toronto Press, Toronto, Canada.
3. *Harvard Business Review,* published by McGraw-Hill Book Co., New York.
4. *Journal of Accountancy,* published by American Institute Publishing Co., New York.
5. *Journal of Marketing,* published by American Marketing Association, Menasha, Wisconsin.
6. *Review of Economic Statistics,* published by Harvard University Press, Cambridge, Mass.

POPULAR REPORTS

1. *Advertising and Selling,* published by Robbins Publishing Co., Inc., Philadelphia, Pa.
2. *Dun's Review,* published by Dun & Bradstreet, Inc., New York.
3. *The Santa Fé Magazine,* published by Railway Exchange, Chicago, Ill.

SCIENTIFIC TITLE

The Angular Distribution of 1 to 3.5 Mev Deuterons Scattered by Deuterons

The differential cross section for the elastic scattering of deuterons by deuterons has been obtained in the energy range from 1 to 3.5 Mev. Observed yields were corrected for the counted particles produced in the $D - D$ reactions and for those scattered by contaminants by using the results of other experiments. The resultant elastic cross sections are probably accurate to about 2.5 percent.*

* By J. Morris Blair, George Freier, Eugene Lampi, William Sleator, Jr., and J. H. Williams, University of Minnesota, Minneapolis, Minnesota.

APPENDIX I

GLOSSARY OF SYMBOLS AND INDEX OF IMPORTANT FORMULAS USED IN THIS TEXT

Latin Symbols

a, b, c, d = constants.
b_{yx} = simple linear regression coefficient.
$b_{12.34\ldots n}$ = net regression coefficient.
$d_{12.34\ldots n}$ = coefficient of separate determination.
d = a deviation.
D = decile.
e = natural logarithm base, 2.71828.
f = frequency of a class interval.
G = geometric mean.
H = harmonic mean.
i = class interval, as 4–7.9 or 4 to 8.
I = index.
k = a constant.
L = lower limit of a class interval, as in 4–7.9, 8–11.9, 4 and 8 are the lower limits of the classes.
m = midpoint of class intervals, as in 4–7.9, 8–11.9, 6 and 10 are the midpoints of these two classes.
Me = median.
Mo = mode.
n = degrees of freedom.
N = total number of items in a sample, total of frequencies in a frequency distribution.

p = percentage.
p_0 = price of base period.
p_1 = price of given period.
p_1' = price of commodity at time 1.
p_1'' = price of commodity at time 2.
p_1^n = price of commodity at time n.
$\frac{p_1}{p_0}$ = price relative.
q = quantity weight.
Q = quartile.
r = coefficient of linear correlation.
r^2 = coefficient of linear determination.
$r_{12.34\ldots n}$ = coefficient of partial correlation.
$R_{1.234\ldots n}$ = coefficient of linear multiple correlation.
$R^2_{1\,234\ldots n}$ = coefficient of linear multiple determination.
s_1 = standard deviation of small sample.
S_k = skewness of a frequency distribution.
S_y = standard error of estimate for y.
t = deviation of a given statistical measure from a hypothetical value divided by its standard error.
T = deviation of a given statistic from the mean of a normal distribution divided by its standard deviation.
u = a variable.
v = a variable.
V = coefficient of variation.
W = weight.
w = a variable.
X = a variable, an item of data.
$\bar{X}$ = arithmetic mean
x = deviation from true mean.
x' = deviation from an assumed mean or origin.
X_1 = dependent variable.
X_2, X_3, X_4, X_n = independent variables.
Y = a variable, an item of data.
$\bar{Y}$ = arithmetic mean.
y = deviation from mean.
y' = deviation from assumed mean or origin.

z = ratio of r to z, Appendix Table III.

z = a residual, deviation between an observed item and its estimated value on the regression line.

Greek Symbols

β_1 = beta coefficient.

π = pi, ratio of circumference of circle to its diameter, 3.1416.

ρ_1 = rho, coefficient of simple curvilinear correlation.

ρ_1^2 = rho, coefficient of simple curvilinear determination.

$P_{1.234\ldots n}$ = rho_1, coefficient of multiple curvilinear correlation.

$P^2{}_{1\,234\ldots n}$ = rho_1, coefficient of multiple curvilinear determination.

σ = sigma, standard deviation.

σ^2 = variance.

$\sigma_{\bar{x}}$ = standard error of arithmetic mean.

σ_σ = standard error of standard deviation.

σ_r = standard error of coefficient of linear correlation.

σ_ρ = standard error of index of correlation.

$\sigma_{\bar{x}_1 - x_2}$ = standard error of the difference of two means.

$\sigma_{p_1 - p_2}$ = standard error of the difference of two percentages.

Σ, sigma, = sum of, or summation.

χ^2 = Chi-square.

Important Formulas

1. Formula for locating midpoint of class intervals

$$m = \frac{L_1 + L_2}{2}$$

2. Formulas for arithmetic mean

(1) $X = \frac{\Sigma X}{N}$, $\bar{Y} = \frac{\Sigma Y}{N}$, for individual items.

(2) $\bar{X} = \frac{\Sigma fm}{N}$, for class intervals, long method.

(3) $\bar{X} = A + \frac{\Sigma f(x')}{N} i$, for class intervals, short method.

(4) $\bar{X}w = \frac{\Sigma WX}{\Sigma W}$, for weighted mean.

3. Formulas for geometric mean

(1) $G = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots X_n}$, for individual items.

(2) $\text{Log } G = \dfrac{\Sigma(\text{Log } X_1 + \text{Log } X_2 + \text{Log } X_3 + \cdots + \text{Log } X_n)}{N}$

or $\text{Log } G = \dfrac{\Sigma(\text{Log } X)}{N}$, for individual items.

(3) $\text{Log } Gw = \dfrac{\Sigma(W_1 \text{ Log } X_1 + W_2 \text{ Log } X_2 + \cdots W_n \text{ Log } X_n)}{\Sigma W}$,

or $\text{Log } G_w = \dfrac{\Sigma(W \text{ Log } X)}{N}$, weighted geometric mean.

(4) $\text{Log } G = \dfrac{\Sigma(f \text{ Log } m)}{N}$ when (class intervals are used).

4. Formulas for harmonic mean

(1) $\dfrac{1}{H} = \dfrac{\Sigma\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)}{N}$ for individual items.

(2) $\dfrac{1}{H} = \dfrac{\Sigma\left(\frac{1}{m_1} + f\frac{1}{m_2} + \cdots + f\frac{1}{m_n}\right)}{N}$ (when class intervals are used).

(3) $\dfrac{1}{H_w} = \dfrac{\Sigma(W_1\frac{1}{x_1} + W_2\frac{1}{x_2} + \cdots + W_n\frac{1}{x_n}}{\Sigma W}$, weighted harmonic mean.

5. Formula for quadratic mean

$$M_q = \sqrt{\frac{\Sigma[(X_1)^2 + (X_2)^2 + (X_3)^2 + \cdots + (X_n)^2]}{N}}$$

6. Formulas for median, quartiles, deciles, etc.

(1) $Me = L + \dfrac{p}{f} i, \dfrac{n}{2}$

(2) $Q_1 = L + \dfrac{p}{f} i, \dfrac{n}{4}$

(3) $Q_3 = L + \dfrac{p}{f} i, \dfrac{3n}{4}$

(4) $D_3 = L + \dfrac{p}{f} i, \dfrac{3n}{10}$

7. Formulas for mode

(1) $Mo = L + \frac{f_2}{f_1 + f_2} i$, interpolation formula.

(2) $Mo = \bar{X} - 3(\bar{X} - Me)$, locational formula.

8. Formula for range

$(X_n - X_1) + 1$

9. Formula for quartile deviation

$QD = \frac{Q_3 - Q_1}{2}$

10. Formulas for mean deviation

(1) $Md = \frac{\Sigma(X_1 - Me) + (X_2 - Me) + \cdots + (X_n - Me)}{N}$, for individual items.

(2) $Md = \frac{\Sigma f(m - Me)}{N}$, for class intervals.

11. Formulas for standard deviation

(1) $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$

(2) $\sigma = i\sqrt{\frac{\Sigma f x'^2}{N} - \left(\frac{\Sigma f x'}{N}\right)^2}$, for class intervals.

12. Formulas for skewness

(1) $S_k = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$

(2) $S_k = \frac{3(\bar{X} - Me)}{\sigma}$

13. Formula for coefficient of variation

$V = \frac{\sigma}{\bar{X}} 100$

14. Formulas for linear regression (least squares)

(1) $Y = a + b_{yx}X$, or $X = a + b_{xy}Y$

(2) $\begin{cases} \Sigma Y = na + b\Sigma X \\ \Sigma XY = a\Sigma X + b\Sigma X^2 \end{cases}$ or $\begin{cases} \Sigma X = na + b\Sigma Y \\ \Sigma XY = a\Sigma Y + b\Sigma Y^2 \end{cases}$

(3) $a = \bar{Y} - b\bar{X}$

(4) $b_{yx} = \dfrac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma X^2 - N\bar{X}^2}$, in terms of the original data.

(5) $b_{yx} = \dfrac{\Sigma xy}{\Sigma x^2}$, in terms of deviations from the means.

15. Formulas for curvilinear regression

(1) Simple parabola

$Y = a + bX + cX^2$

(2) When the original data are used

$$\begin{cases} \Sigma Y = na + b\Sigma X + c\Sigma X^2 \\ \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 \\ \Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 \end{cases}$$

or

$$\begin{cases} na + b\Sigma X + c\Sigma X^2 = \Sigma Y \\ a\Sigma X + b\Sigma X^2 + c\Sigma X^3 = \Sigma XY \\ a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 = \Sigma X^2Y \end{cases}$$

(3) When the deviations from the means are used

$a = \bar{Y} - b\bar{X} - c\bar{U}$ (and U is substituted for X^2)

$\Sigma(x)^2b + \Sigma(xu)c = \Sigma xy$

$\Sigma(xu)b + \Sigma(u)^2c = \Sigma uy$

(4) Cubic parabola

(i) $Y = a + bX + cX^2 + dX^3$

(ii)
$$\left.\begin{cases} na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 = \Sigma Y \\ a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 = \Sigma XY \\ a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 = \Sigma X^2Y \\ a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 = \Sigma X^3Y \end{cases}\right\} \text{when the original data are used}$$

(iii)
$$\left.\begin{cases} (\Sigma x^2)b + (\Sigma xu)c + (\Sigma xv)d = \Sigma xy \\ (\Sigma xu)b + (\Sigma u^2)c + (\Sigma uw)d = \Sigma uy \\ (\Sigma xv)b + (\Sigma uv)c + (\Sigma v^2)d = \Sigma vy \end{cases}\right\} \text{when the deviations from the mean are used and } U \text{ is substituted for } X^2 \text{ and } V \text{ for } X^3$$

(5) a. Hyperbolic curve $y = ax^b$

$\log Y = \log a + b(\log X)$,

$$\begin{cases} \Sigma(\log Y) = N \log a + b\Sigma(\log X) \\ \Sigma(\log X \cdot \log Y) = \log a\Sigma(\log) + b\Sigma(\log X)^2 \end{cases}$$

b. Gompertz curve, $y = ab^{c^x}$

$\log Y = \log a + (\log b)c^x$

$$c^n = \frac{d_2}{d_1}$$

$$\log b = \frac{d_1(c-1)}{(c^n-1)^2}$$

$$\log = \frac{1}{N}\left\{ S_1 - \frac{d_1}{(c^n-1)} \right\}$$

(6) (i) Exponential curve (1) $Y = ab^x$, or

$\log y = \log a + (\log b)x$

$$\begin{cases} \Sigma(\log y) = n \log a + \log b\Sigma(x) \\ \Sigma(x \cdot \log y) = \log a\Sigma(x) + \log \cdot b\Sigma(x^2) \end{cases}$$

(ii) Compound interest curve, or a special case of exponential curves

$y = P(1+r)^x$, or

$\log y = \log P + \log (1+r)x$

(7) Logistic (1) $\frac{1}{y} = a + bc^x$

$$c^n = \frac{d_2}{d_1}$$

$$d = \frac{d_1(c-1)}{(c^n-1)^2}$$

$$a = \frac{1}{N}\left(S_1 - \frac{d_1}{c^{n-1}}\right)$$

See Worksheet No. 47.

16. Formulas for standard error of estimate

(1) $S_y = \sqrt{\frac{\Sigma(Y-Y')^2}{N}}$, $(Y - Y') = z$

(2) $S_y = \sqrt{\frac{\Sigma z^2}{N}}$

(3) $S_y = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N}}$

(4) $S_y = \sqrt{\dfrac{\Sigma y^2 - b_{yx} \cdot \Sigma xy}{N}}$

(5) $S_y{}^2 = \sigma_y{}^2(1 - r^2)^2$

(6) $S_y = \sigma_x\sqrt{1 - r^2}$

17. Formulas for computing coefficient of linear determination and correlation

(1) $r^2 = \dfrac{\sigma_{y'}{}^2}{\sigma_y{}^2}$

(2) $r^2 = 1 - \dfrac{S_y{}^2}{\sigma_y{}^2}$

(3) $r^2 = \dfrac{(\Sigma xy)^2}{\Sigma x^2 \cdot \Sigma y^2}$

(4) $r = \dfrac{\sigma_{y'}}{\sigma_y}$

(5) $r = \sqrt{1 - \dfrac{S_y{}^2}{\sigma_y{}^2}}$

(6) $r = \dfrac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$

(7) $r = \dfrac{\Sigma xy}{N\sigma_x\sigma_y}$

(8) $r = \dfrac{N\Sigma XY - (\Sigma X \cdot \Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$

(9) $r^2 = b_{yx} \cdot b_{xy}$

(10) $r = \sqrt{b_{yx} \cdot b_{xy}}$

18. Formulas for the normal curve

(1) $y_c = \dfrac{Ni}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}}$ for any ordinate in the normal curve

(2) $y_c = \dfrac{Ni}{2.5066\sigma}$ for the maximum ordinate at zero (0) deviation from the mean

19. Formulas for standard errors

(1) The mean $\sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{N}}$

(2) The standard deviation $\sigma_\sigma = \frac{\sigma_x}{\sqrt{2N}}$

(3) The coefficient of correlation $\sigma_r = \frac{1 - r^2}{\sqrt{N}}$

(4) The standard error of the difference between two means

$\sigma_D = \sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}}$ when computed from the two standard deviations

(5) $\sigma_D = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}$ when computed from the standard error of the two means

20. The t-test formulas

(1) Difference between two means

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_D}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{1\cdot 2}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

(2) Difference between two regression coefficients

$$t = \frac{b_1 - b_2}{\sigma_{b_1 + b_2}} \quad \text{and} \quad \sigma_{b_1+b_2} = \frac{S_{y_1}^2}{\Sigma x_1^2} + \frac{S_{y_2}^2}{\Sigma x_2^2}$$

(3) Test of significance of coefficients of correlation

$$t = r \frac{\sqrt{N - 2}}{\sqrt{1 - r^2}}$$

(4) Difference between two percentages

$$t = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} \quad \text{and} \quad \sigma_{p_1+p_2} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}}$$

in which p = favorable occurrences and q = unfavorable occurrences

21. Equations for Chi-square

(1) $\chi^2 = \frac{(X_1 - m)^2}{m} + \frac{(X_2 - m)^2}{m}$

(2) $\chi^2 = \Sigma \left\{ \frac{(f_1 - f_2)^2}{f_2} \right\}$ for a series of frequencies

22. Formulas for secular trend in time series

(1) $Y = a + b_{yx}X$, for linear trends, short cut

(2) $a = \dfrac{\Sigma Y}{N}$

(3) $b_{yx} = \dfrac{\Sigma XY}{\Sigma X^2}$

(4) $Y = a + bX + cX^2$, for simple parabola, short cut

(5) $b = \dfrac{\Sigma XY}{\Sigma X^2}$

(6) $\left.\begin{aligned} Na + c\Sigma X^2 &= \Sigma Y \\ \Sigma X^2 a + c\Sigma X^4 &= \Sigma X^2 Y \end{aligned}\right\}$ for solution of a and c-values

(7) Log $Y = a + bX$ for simple log line

(8) $Y = a + bX + cX^2 + dX^3$ for cubic parabola, short cut

(9) $\left.\begin{aligned} \Sigma Y &= Na + c\Sigma X^2 \\ \Sigma X^2 Y &= a\Sigma X^2 + c\Sigma X^4 \end{aligned}\right\}$ for solution of a and c-values

$\left.\begin{aligned} \Sigma XY &= b\Sigma X^2 + d\Sigma X^4 \\ \Sigma X^3 Y &= b\Sigma X^4 + d\Sigma X^6 \end{aligned}\right\}$ for solution of b and d-values

(10) $\log Y = a + bX$

(11) $a = \dfrac{\Sigma \log Y}{N}$

$b = \dfrac{\Sigma X \log Y}{\Sigma X^2}$

23. Residual method of computing cycles

(1) $(C \times I) = \dfrac{T \times C \times S \times I}{T \times C}$, or

(2) $(C \times I) = \dfrac{S \times C \times I}{S}$, or

(3) $(C \times I) = \dfrac{T \times C \times I}{T}$, or

(4) $(C + I) = \dfrac{T(S + C + I)}{T} - S$

C = Cycle
I = Irregular change
S = Seasonal variation
T = Secular trend

24. Formulas for index numbers

(1) Unweighted aggregative index of prices

$$I = \frac{\Sigma(p_1' + p_1'' + p_1''' + \cdots + p_1^n)}{\Sigma(p_0' + p_0'' + p_0''' + \cdots + p_0^n)} \text{ or } I = \frac{\Sigma(p_1)}{\Sigma(p_0)}$$

(2) Weighted aggregative index of prices

$$I = \frac{\Sigma[(p_1q_1)' + (p_1q_1)'' + (p_1q_1)''' + \cdots + (p_1q_1)^n]}{\Sigma[(p_0q_0)' + (p_0q_0)'' + (p_0q_0)''' + \cdots + (p_0q_0)^n]}$$

or $\dfrac{\Sigma(p_1q_1)}{\Sigma(p_0q_0)}$

(3) Unweighted average of price relatives index

$$I = \frac{\Sigma\left(\frac{p_1}{p_0}\right)' + \left(\frac{p_1}{p_0}\right)'' + \left(\frac{p_1}{p_0}\right)''' + \cdots + \left(\frac{p_1}{p_0}\right)^n}{N} \quad \text{or} \quad \frac{\Sigma\left(\frac{p_1}{p_0}\right)}{N}$$

(4) Weighted average of price relatives index in which the weights (p_0q_0), or price x quantity, or value are used

$$I = \frac{\Sigma\left[\left(\frac{p_1}{p_0'}\right)' p_0'q_0' + \left(\frac{p_1}{p_0''}\right)'' p_0''q_0'' + \cdots + \left(\frac{p_1^n}{p_0^n}\right) p_0^n q_0^n\right]}{\Sigma[p_0'q_0' + p_0''q_0'' + \cdots + p_0^n q_0^n}$$

or $\dfrac{\Sigma\left[\left(\frac{p_1}{p_0}\right)p_0q_0\right]}{\Sigma(p_0q_0)}$

(5) Weighted geometric average of price relatives index using logarithms

$$I = \Sigma\left[p_0'q_0' \log\left(\frac{p_1'}{p_0'}\right) + p_0''q_0'' \log\left(\frac{p_1''}{p_0''}\right) + \cdots + p_0^n q_0^n \log\left(\frac{p_1^n}{p_0^n}\right)\right]$$

or $$I = \Sigma\left[p_0q_0 \log\left(\frac{p_1}{p_0}\right)\right]$$

(6) Quantity indexes weighted by prices

$$I = \frac{\Sigma[(q_1'p_0') + (q_1''p_0'') + (q_1'''p_0''') + \cdots + (q_1^n p_0^n)]}{\Sigma[(q_0'p_0') + (q_0''p_0'') + (q_0'''p_0''') + \cdots + (q_0^n p_0^n)]}$$

or $\dfrac{\Sigma(q_1p_0)}{\Sigma(q_0p_0)}$

25. Formulas for analysis of variance

(1) $\sigma_x^2 = \dfrac{\Sigma x^2}{N-1}$, variance is square of σ_x

(2) $\Sigma x^2 = \Sigma X^2 - \dfrac{(\Sigma X)^2}{N}$, total squared deviations from mean of all classes

(3) $\dfrac{(\Sigma X)^2}{N}$ = correction value for total of all classes

(4) $\dfrac{\Sigma[(\Sigma X_1)^2 + (\Sigma X_2)^2 + \cdots + (\Sigma X_n)^2]}{f} - \dfrac{(\Sigma X)^2}{N} =$ total sum of squares between class means

APPENDIX II

APPENDIX OF TECHNICAL TERMS

A

ABSCISSA = the distance measured from Y-axis to a point along a parallel line to the x-axis. If the point is to the right of the Y-axis on the graph, the abscissa is positive. If the point is to the left of the Y-axis, the abscissa is negative.

ADDITIVE RELATIONSHIP or ADDITIVE FUNCTION = the joining of two or more constants or values so that the result is their sum.

AGGREGATE = the sum or total of a group of values or numbers.

AGGREGATIVE INDEX NUMBER = an index number computed by dividing the sum of values for a given period or area by the sum of values for a base period or area for the same series. The given total is divided by the base total and the quotient expressed as a percentage.

ALGEBRAIC SUM = the sum of a group of numbers with signs + and/or − taken so that the sum is the difference between the sum of the + numbers and the sum of the − numbers.

AMPLITUDE = the maximum distance from the lowest to the highest points in a curve.

ANALYSIS OF COVARIANCE = a combination of the analysis of variance with the analysis of regression and correlation in which the significance of the variation between classes is computed after the influence of one or more other variables has been removed.

ANALYSIS OF VARIANCE = a method of computing the degree of significance in the variations of a particular characteristic among comparable groups or classes of data.

ARBITRARY ORIGIN = any value which is used as a zero point from which to compute deviations on a scale such as the guessed average or assumed mean used in computing the arithmetic mean by the short method.

ARITHMETIC MEAN OR AVERAGE = the sum of the items, numbers, or magnitudes in a series divided by their number, $\bar{X} = \frac{\Sigma X}{N}$.

ARITHMETIC SCALE = a scale on which equal absolute numbers are represented by equal spaces on the scale regardless of where they appear on the scale.

ARRAY = an orderly arrangement of numbers or values according to magnitude usually from the smallest to the largest.

AVERAGE = any measure of central tendency. Frequently used to indicate the arithmetic mean.

AVERAGE DEVIATION OR MEAN DEVIATION = the arithmetic mean of the absolute deviations of the items of a series of data from a measure of central tendency, usually the median, but sometimes the arithmetic mean is used. It is a minimum from a median and equals $.7979\sigma$.

AXIS, X-AXIS, Y-AXIS = the horizontal and the vertical scales of a coordinate graph or lines.

B

b_{yx} = symbol for total regression coefficient of Y and X when both are expressed as original data or deviations from any origin but not in standard deviations. Also expressed as b_{12} for X_1 and X_2.

$b_{12.34\ldots n}$ = symbol of partial or net regression and expressing the relationship between X_1 and X_2 after including $X_3 \cdot X_4 \cdots X_n$ in the computations and removing their influence on X_1 when the data are not expressed in terms of their standard deviations.

BAND CHART = a graph composed of several irregular zones extending across the scale from left to right indicating several series of varying magnitudes.

BAR CHART = a graph composed of a series of bars of varying length each of which represents comparable magnitudes. The bars may be either vertical or horizontal and are all measured from the same base.

BASE = a value, price, quantity, or other magnitude in space or time from which a series of items are measured. In graphs and cycles the base is usually zero (0). In seasonal variation and in index numbers it is usually 100. In ratios the base is the denominator of the fraction of which the ratio is the quotient.

BETA REGRESSION COEFFICIENT = the value of a regression coefficient when it expresses the relationship between two variables in terms of their standard deviations. $\beta_{12.34\ldots n} = b_{12.34\ldots n}\frac{\sigma_2}{\sigma_1}$

BIASED SAMPLE = a sample not taken at random or one which does not represent the population from which it is drawn.

BIMODAL = a frequency distribution which has two maximum frequencies of equal size is said to be bimodal; having two modes.

BINOMIAL = an algebraic expression consisting of two terms, $(a + b)$, $(x - y)$.

BINOMIAL EXPANSION = the expansion of a binomial to a given power, as $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$.

BUSINESS CYCLE = the variation of business activity which remains after secular trend and seasonal variation have been removed. The more or less regular movement of business activity through successive periods of prosperity and depression. It may be a cycle for a single series or an average of a group of series.

C

CAPTION = the heading of a column or vertical space in a table.

CELL = the space enclosed in a table by the intersection of two parallel vertical and two parallel horizontal lines.

CENTERING = the process of shifting a moving average or trend from the beginning to the middle of a time period.

CHECK SUM = the sum of a group of data which may be used to check the accuracy of computations in a worksheet. It is a grand total which checks against sub-totals or group totals.

CHI-SQUARE, χ^2 or SQUARE CONTINGENCY = the sum of the quotients obtained by dividing the square of the difference between an actual and an assumed or theoretical frequency by the assumed frequency, $\chi^2 = \frac{(X_1 - m)^2}{m} + \frac{(X_2 - m)^2}{m}$.

CLASS INTERVAL, i, = the width of a class as 10–14.99, 15–19.99, or 9.50–14.49, 14.50–19.49, or 10 to 15, 15 to 20.

COEFFICIENT = a statistical constant or ratio which is independent of the units in which the data are measured. b_{yx}, or b_{12}, is the regression coefficient between X and Y, or X_1 and X_2. r_{yx} is the coefficient of correlation between X and Y.

COEFFICIENT OF CORRELATION = a measure of the amount of variation in a dependent variable which is associated with variation in one or more independent variables expressed as the square root of a percentage. Complete or perfect correlation is designated as 1.00.

COEFFICIENT OF TOTAL DETERMINATION = a measure of the amount of variation in a dependent variable which is associated with one or more independent variables expressed as a percentage.

COEFFICIENT OF PARTIAL CORRELATION = a measure of the net relationship of a dependent variable to one independent variable after the effect of other independent variables has been computed out or removed; $r_{12.34 \ldots n}$.

COEFFICIENT OF SKEWNESS = a measure of the onesidedness or asymmetry of a frequency distribution. If the left side of the distribution is longer and thinner than the right, the distribution is said to be skewed to the left. If the right side is the longer, it is skewed to the right. In any distribution the mean falls nearest the skewed end and the mode falls farthest from the skewed

end, while the median always falls between the mean and the mode.

COEFFICIENT OF VARIATION = the measure of the relative variability of a frequency distribution based on the ratio of its standard deviation to its mean expressed as a percentage, $V = \frac{\sigma}{\overline{X}} 100$.

COLUMN = a series of spaces or numbers in a vertical space in a table.

CONSTANT = a quantity which retains the same value throughout a problem or series of computations; in $Y = a + bX$, a and b are constants.

CONTINUOUS DATA = values which may be measured in infinitesimally small fractions of a unit of measure, as tons of coal, acres of land, or ounces of gold.

CORRECTION = the quantity which must be added to or subtracted from one value to give another desired value. (1) The sum of the deviations from an assumed mean divided by their number. (2) N times the product of the means which must be subtracted from the sum of the products of original data to give the sum of the product of the deviations from the means.

CORRELATION TABLE = a coordinate table of two frequency distributions designed to facilitate the computations of the sums of the squared deviations and products of the two variables required for the coefficient of linear correlation.

CRITICAL RATIO = the difference between two comparable statistics divided by the standard error of that difference, also called the t-test; $t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{D_{1-2}}}$.

CUBIC PARABOLA = the curve fitted by the equation

$$Y = a + bX + cX^2 + dX^3,$$

in which a, b, c, and d are constants based upon the average relationship in the data.

CUMULATIVE FREQUENCY DISTRIBUTION = a tabulation of frequencies formed by (*a*) adding to the frequency of each succeeding class the total of the frequencies of all preceding classes, or (*b*) adding to the frequency of each preceding class the total of all succeeding classes. A frequency distribution obtained by the (*a*) method is cumulated on the "less than" basis and those obtained by the (*b*) method are cumulated on the "more than" basis.

CURVILINEAR CORRELATION = any correlation based on regression lines which are non-linear.

CURVILINEAR REGRESSION = any regression between two variables which is expressed by a line that is other than straight.

D

DATA = the raw materials for statistical studies. Any group of facts, figures, measurements, values, or numbers.

DECILE = any one of the nine points which divide a frequency distribution into ten equal parts.

DEPENDENT VARIABLE = the variable whose estimates depend on another variable or variables. A causal relationship is not essential. All that is necessary is that the relationship between the variables is set up so that values for the dependent variable are estimated from the other variable on the basis of some function or relationship.

DEVIATION = the difference between two quantities or values, or between two statistics or between a statistic and an item of data, as deviation from the mean; absolute error.

DISCRETE DATA = values of a variable which may be taken at only certain limited points, usually in whole units, as men, wives, or soldiers.

DOOLITTLE METHOD = a systematic method for the rapid solution of simultaneous equations. Its details are explained in Chapter 23 of this book. If a statistician has many long equations to solve the Doolittle method should be mastered.

E

ERROR = (1) the difference between an observed and an estimated value. (2) Differences between observed values and their mean. (3) Differences between a statistic of a sample and the parameter of the population from which the sample is taken.

ERROR OF ESTIMATE = (1) the deviation of an observed item of a dependent variable from the regression line, $(Y - Y') = z$. (2) The difference between any estimated value and an observed value.

EXPONENTIAL CURVE = a curve determined by the equation $y = ab^x$ in which a and b are constants.

F

FACTOR = (1) a variable considered in a statistical problem. (2) a characteristic or quality common to two or more variables.

FREE-HAND CURVE = a line drawn through a graph of plotted data to represent their relationship without the computation of a mathematical equation.

FREQUENCY CURVE = any curve or graph which represents the size of the frequencies of a frequency distribution or the number of items in successive class intervals.

FREQUENCY DISTRIBUTION = a table showing the number of items in a sample which fall in each one of a series of succeeding classes.

FUNCTION = any variable whose value depends on another variable, $Y = a + bX$. The value of Y depends on the value of X or Y is a function of X.

G

GEOMETRIC MEAN = the nth root of the product of n factors. It may be computed by finding the antilogarithm of the mean logarithms of the sample items.

GUESSED MEAN = estimated average, assumed mean, or guessed average, an arbitrary point from which an arithmetic mean may be computed by the short method.

H

HARMONIC MEAN = the reciprocal of the arithmetic mean of the reciprocals of the sample items.

HISTOGRAM = a graph which represents the magnitude of the frequencies of a frequency distribution by adjoining vertical bars.

I

INDEPENDENT VARIABLE = any variable which may be given any desired value which will in turn determine the value of an associated dependent variable, $Y = a + bX$. The assigning of any value to X determines the associated value of Y.

INDEX NUMBER = a computed value which in relation to a base measures the relative change of a quantity from time to time or area to area. An index is usually an average of some type expressed as a percentage.

INTERCEPT = the distance on either axis, X or Y, from the point of origin to the point at which a regression line crosses that axis.

INTERPOLATION = any process of estimating values between two or more known points.

INTERQUARTILE RANGE = the distance on the range scale of a frequency distribution between the Q_1 and Q_3.

J

J-SHAPED CURVE = a frequency distribution which is high at one end and low at the other instead of low at the ends and high in the middle.

JOINT FUNCTION = the joining of two or more functions so that their gross result is other than their sum. Usually they are multiplied or divided in their effect by being joined.

L

LAG = the length of time by which change in one variable or series follows change in another.

LEAD = the length of time by which change in one variable or series precedes change in another.

LINEAR = a relationship which may be expressed by a straight line.

LINEAR CORRELATION = correlation assuming a straight-line relationship between or among the variables.

LOGARITHM = an exponent indicating the power to which a base must be raised to equal a number. There are two systems of logarithms, (1) the natural logarithms based on e, or (2.7182828) which is not used much in elementary statistics, (2) the common logarithms based on 10 which are very useful in statistical computations and a table of which is included in Appendix Table V. A logarithm has two parts, the characteristic or whole number, and the mantissa or fractional part. Only the mantissas appear in tables of logarithms.

LOGARITHMIC SCALE = a scale for charts on which equal spaces are given to equal percentages.

LORENZE CURVE = a graph which measures cumulative values of two variables as portions of 100 percent for the total on the two axes. A straight line from the lower left-hand corner to the upper right-hand corner of the graph indicates an even or equal distribution of the variables throughout their range. As the concentration of either quantity becomes more marked in either the upper or lower limits of the 100 range, the line becomes more curved.

M

MEASURE OF CENTRAL TENDENCY = any one of the averages, mean, median, mode, etc.

MEASURE OF DISPERSION = a measure of the scatter of the items around some measure of central tendency, as range, standard deviation, average deviation, etc.

MEDIAN = that point on the range of frequency distribution which divides the observations into two equal parts.

MEDIAN CLASS = the class in which the median falls.

METHOD OF LEAST SQUARES = that method of computing a statistic which reduces the squared deviations from the point or line to a minimum.

MID-POINT = the value at the middle of a class interval.

MODE = that frequency or value of a distribution that is most numerous.

MOVING AVERAGE = a series of averages obtained by averaging a given number of items in a series successively so that for each successive group the first item of the previous group is dropped and the next item below that group is added.

MULTIPLE CORRELATION = the correlation between one dependent variable and two or more independent variables weighted by the net regression coefficients between the dependent and each of the independent variables.

N

N = the number of items in a sample.

NET REGRESSION COEFFICIENTS = the relation between a dependent variable and an independent variable after the effect of one or more other independent variables has been brought into the problem and its effect held constant, or computed out of the result.

NORMAL DISTRIBUTION = a distribution which falls in a bell-shape so that the two variations or slopes from the middle are uniform and in the shape of the normal probability curve.

NORMAL PROBABILITY CURVE = the curve fitted or expressed by the normal probability equation, $y = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$

O

ORDINATE = the distance from the X-axis to a point measured parallel to the Y-axis.

ORGANIZATION TABLE = a table designed and used to reveal and measure the relationship between two or more related variables expressed as class intervals. The prime tool of tabular analysis.

P

PARABOLA = a curve determined by the equation,

$$Y = a + bX + cX^2.$$

$y = ax^2$ in which a, b, and c are constants.

PARTIAL CORRELATION = same as coefficient of partial correlation.

π, PI = the ratio of the circumference of a circle to its diameter, 3.1416 or $\frac{22}{7}$, roughly.

PICTOGRAM = a statistical chart composed of pictures arranged to show magnitudes.

PIE CHART or PIE DIAGRAMS = a graph in the form of a circle, usually with its circumference divided into 100 equal parts to measure percentages.

POINT BINOMIAL = the term coefficients obtained by the expansion of the binomial expression $(a + b)^n$ when $(a + b) = 1$ and n is a positive finite integer, as $(\frac{1}{2} + \frac{1}{2})^{20}$.

POPULATION = the entire field or group of items from which a sample is taken. Same as UNIVERSE.

PRICE RELATIVE = the price of any single commodity at any given time divided by the price of the same commodity for another time, called the base, expressed as a percentage.

PROBABILITY = "the probability of an event is the relative frequency with which this event recurs in an indefinitely prolonged sequence or series of observations." [1] The likelihood of the occurrence of an event.

Q

QUADRATIC MEAN or ROOT-SQUARE-MEAN = the square root of the mean of the sum of the items squared.

[1] Mises, Richard von, *Probability, Statistics, and Truths*, The Macmillan Co., N.Y., 1939.

QUARTILE, Q_1 MD. Q_3 = one of the three points on the range scale of a frequency distribution which divides the total of the frequencies into four equal parts.

QUESTIONNAIRE = an orderly arrangement of questions related to a statistical problem submitted to a number of persons for answers. Questionnaires are usually submitted by mail.

QUINTILE = one of the four points on the range scale of a frequency distribution which divides the total of the frequencies into five equal parts.

R

$R_{1.234\ldots n}$ = symbol of the coefficient of linear multiple correlation. $R^2_{1.234\ldots n}$ = the symbol of coefficient of linear multiple determination.

RANDOM SAMPLING = the process of selecting a sample from a population so that every item in the population has an equal and independent chance of being included in the sample.

RANGE = the total distance between the smallest and the largest items in a sample plus one unit.

RECIPROCAL = the reciprocal of a number is 1 divided by that number. The reciprocal of 5 is $\frac{1}{5}$ or .2.

REGRESSION COEFFICIENT = the average measure of the number of units of change in a dependent variable for each unit of change in the independent variable.

REGRESSION LINE = the line which describes the relationship between two variables based on the average relationship between them.

RELATIVE = the quotient representing the relationship of the values of two times or two places expressed as a percentage.

REPLICATION = the division of an experiment into subdivisions or cells to secure greater cross-section uniformity of data among variables or to reduce the effects of exterior variables.

RESIDUAL = the difference between an actual and an estimated value, as $(Y - Y') = z$.

ROUNDING = expressing a number in a more general form by changing one or more of its digits at the extreme right to zeros and either raising or lowering the next digit to the left as the digits altered are less than or more than 5, as 473 may be rounded to 470 if only one place is involved or to 500 if two places are involved.

ROW = the series of numbers extending crosswise or parallel to the X-axis in a table.

S

SAMPLE = a limited or finite number of items or data selected from a population or universe.

SAMPLING RELIABILITY = the degree of accuracy with which a sample represents its population.

SCATTER DIAGRAM or CHART = a graph of coordinate scales on which the paired values of the data of X and Y are located by dots or marks opposite the appropriate points on the two axes.

SCHEDULE = an orderly arrangement of questions related to a statistical problem which is to be filled out by an interviewer from information secured during an interview.

SEASONAL INDEX = an index usually in the form of a percentage with the median of the year as a base of 100 which measures the changes in an activity from month to month.

SEASONAL VARIATION = that change in a time series which is due to the season or time of year.

SECULAR TREND = a line describing the long time expansion or contraction of values in a time series.

SPATIAL SERIES = data based on location or space instead of time or duration.

STANDARD DEVIATION = σ, the square root of the mean of the squared deviations of the items of a sample from its arithmetic mean. In a normal distribution one σ plus and minus from the arithmetic mean includes 68.27% of the items in the sample if the sample is large.

STANDARD ERROR OF ESTIMATE = the square root of the mean of the squared deviations of the values of the dependent variable from the regression line. $S_y = \sqrt{\frac{\Sigma z^2}{N}}$, or $S_y = \sqrt{\frac{(Y - Y')^2}{N}}$.

STANDARD SCORE = a deviation from the mean divided by the standard deviation of distribution, as $\frac{x}{\sigma x} = T$.

STATISTIC = a value computed from a sample, such as a mean, mode, range, standard deviation, etc.

T

t = the ratio of a statistic to its standard error.
T = a statistic in terms of its standard deviation.

TABULATION = the orderly arrangement of data (1) in tables of columns and rows, or (2) in frequency distributions, arrays, or other systematic forms.

TALLY SHEET = a sheet or table with captions and stubs for class intervals and frequencies into which data are recorded by a tally mark for each item.

TERTILE = either one of the two points on the range of class intervals which divide the frequency distribution into three equal parts.

TIME SERIES = a sequence of quantities corresponding to successive points of time such as seconds, minutes, hours, days, weeks, months, years, or decades.

V

VARIABLE = a value or measure which varies in magnitude for separate points of time or location or other characteristic or base.

VARIANCE = the square of the standard deviation, σ^2.

W

W = symbol of weight.

WEIGHTED ARITHMETIC MEAN = a mean in the computation of which the several items are included as many times as there are

units in another number called the weight. The weights indicate the relative importance of the items.

X

X = an item of data.
$\bar{X}$ = symbol of arithmetic mean.
x = symbol of deviation from the mean.
x' = symbol of deviation from an arbitrary origin.
x-AXIS = the horizontal base of a chart or graph.

Y

Y = symbol of an item of data.
$\bar{Y}$ = symbol of mean of data.
y = symbol of deviation from arithmetic mean.
y' = symbol of deviation from arbitrary origin.
y-AXIS = the vertical base of a chart or graph.

APPENDIX III

MATHEMATICAL APPENDIX, SQUARE ROOTS, SIMULTANEOUS EQUATIONS, AND LOGARITHMS

SQUARE ROOTS

Most texts and workbooks in statistics contain tables of squares and square roots of all numbers up to 1,000. Appendix Table IV in this text is such a table. If a student requires the square root of any whole number up to one thousand, he should read it from the table in order to economize time. By this method he can locate and record the square roots of a dozen or more numbers in the time that would be required to extract one root. If, however, a root must be extracted, the following model problems will clearly indicate the necessary steps.

1. Beginning with the decimal point, divide the number into periods of two digits each, as

$$\overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}$$

2. Take the largest possible square out of the first period on the left, as

$$\begin{array}{l} \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\,|\,\underline{2} \\ 4 \end{array}$$

In this case it is 4, the square of 2.

3. Subtract the square from the period and bring down the next period, as

$$\begin{array}{l} \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\,|\,\underline{2} \\ 4 \\ \overline{2\ 78} \end{array}$$

4. Double the root already found for a trial divisor and divide it into the digits brought down excluding the last one, as

$$\begin{array}{r} \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\,|\,2 \\ 4 \\ \hline 4\ )2\ 78 \end{array}$$

4 will go into 27 six times.

5. Write the 6 in the quotient and also after the 4 in the divisor, as

$$\begin{array}{r} \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\,|\,26 \\ 4 \\ \hline 46\ )2\ 78 \\ 2\ 76 \\ \hline 2 \end{array}$$

and multiply the completed divisor 46 by 6 = 276 and subtract from the 278, which leaves the remainder 2.

6. Bring down the next period, 48, and proceed as before, as

$$\begin{array}{r} \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\,|\,260 \\ 4 \\ \hline 46\ )2\ 78 \\ 2\ 76 \\ \hline 52\ )248 \end{array}$$

Double the root already found, 26. Divide it into the figures brought down excepting the last. 52 into 24 will not go. Record a zero (0) in the quotient as 260.

7. Bring down the next period and proceed as before, as

$$\begin{array}{r} \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\,|\,260 \\ 4 \\ \hline 46\ )2\ 78 \\ 2\ 76 \\ \hline 520\ )2\ 48\ 25 \end{array}$$

Double the root already found, 260, and divide the 520 into the figures brought down, 24825, excepting the last one. 520 goes into 2482 four times. Write the 4 in the quotient as 2604 and also

in the divisor as 5204 and multiply the completed divisor, 5204, by 4 which equals 20816 and subtract,

$$\begin{array}{rl} & \overline{6}\ \overline{78}.\ \overline{48}\ \overline{25}\ \big|\ 26.04 \quad \text{(or } 26.05\text{)} \\ & 4 \\ 46\ \big) & 2\ 78 \\ & 2\ 76 \\ 5204\ \big) & 2\ 48\ 25 \\ & 2\ 08\ 16 \\ & \overline{\quad 40\ 09} \end{array}$$

Since the remainder, 4009 is more than one-half of the divisor, 5204, the quotient may be written 2605 instead of 2604, or another period of $\overline{00}$ may be brought down and the root carried to another decimal place.

8. Beginning at the right point off as many decimal places in the root as there are periods to the right of the decimal point in the original number. This is 2. The root correct to two decimal places is, therefore, 26.05.

A second example is:

$$\begin{array}{rl} & \overline{85}.\ \overline{69}\ \overline{70}\ \overline{00}\ \big|\ 9.257+ \\ & 81 \\ 182\ \big) & 4\ 69 \\ & 3\ 64 \\ 1845\ \big) & 1\ 05\ 70 \\ & 92\ 25 \\ 18507\ \big) & 13\ 45\ 00 \\ & 12\ 95\ 49 \\ & \overline{\quad 49\ 51} \end{array}$$

SIMPLE ALGEBRAIC EQUATIONS

Elementary statistics occasionally require the solution of simple simultaneous equations. The equations,

$$\begin{aligned} (1)\quad & 2X + 4Y = 20 \\ (2)\quad & 3X - 2Y = 12 \end{aligned}$$

may be solved by substituting Y values for X values as follows:

First Method: $2X = 20 - 4Y$ $\qquad$ $3X = 12 + 2Y$

$$X = \frac{20 - 4Y}{2} \qquad X = \frac{12 + 2Y}{3}$$

Simplifying this equation gives

$$\frac{20 - 4Y}{2} = \frac{12 + 2Y}{3}$$

$$60 - 12Y = 24 + 4Y$$

$$-12Y - 4Y = -60 + 24$$

$$-16Y = -36$$

$$Y = \frac{36}{16} = 2\tfrac{1}{4}$$

Substituting in equation (1)

$$2X = 20 - 4(2\tfrac{1}{4})$$

$$2X = 20 - 9$$

$$2X = 11$$

$$X = 5.5$$

Second Method: The same equations may be solved by multiplication and subtraction, as follows:

(1) $2X + 4Y = 20$

(2) $3X - 2Y = 12$

Multiply each equation by numbers that will make the coefficients of X equal, as

$$3(2X + 4Y = 20) = 6X + 12Y = 60$$

$$2(3X - 2Y = 12) = 6X - 4Y = 24$$

Subtracting $\qquad 16Y = 36$

$$Y = \frac{36}{16} = 2\tfrac{1}{4}$$

Substituting the value of $Y(2\frac{1}{4})$ in equation (1) to obtain the value of X, we have,

$$2X + 4(2\tfrac{1}{4}) = 20$$

$$2X + 9 = 20$$

$$2X = 20 - 9$$

$$2X = 11$$

$$X = 5.5$$

Proof: $2(5.5) + 4(2\tfrac{1}{4}) = 20$

$$3(5.5) - 2(2\tfrac{1}{4}) = 12$$

$$11 + 9 = 20$$

$$16.5 - 4.5 = 12$$

In the above example the figures are small and easy of solution, but in the chapters on Regression and Correlation the student will find that this method gives numbers so large that they are cumbersome to manipulate on a calculating machine. To overcome this difficulty a shorter method has been developed.

Third Method: The equations are solved by division and subtraction, as

$$\begin{aligned} (1)\quad & 2X + 4Y = 20 \\ (2)\quad & 3X - 2Y = 12 \end{aligned}$$

Divide each equation by the coefficient of X in that equation, as

$$\begin{aligned} (2X + 4Y = 20) &= (1X + 2Y = 10) \\ (3X - 2Y = 12) &= (1X - .6667Y = 4) \end{aligned}$$

$$\begin{array}{rr} \text{Subtract} & (1')X + 2.0000Y = 10 \\ & (2')X - 0.6667Y = 4 \\ \hline & 2.6667Y = 6 \end{array}$$

$$Y = \frac{6}{2.6667} = 2\tfrac{1}{4}$$

Substituting Y in equation (1') we have

$$\begin{aligned} X + 2.0000 \times 2\tfrac{1}{4} &= 10 \\ X &= 10 - 4\tfrac{1}{2} \\ X &= 5.5 \end{aligned}$$

By (1) division and (2) subtraction the size of the numbers are decreased and the arithmetic solution of the problem made much easier. The student will find it to his great advantage to use this method exclusively.

POWERS AND EXPONENTS

An exponent indicates how many times a number has been multiplied by itself to produce a product, as

$$\begin{aligned} 10 \times 10 &= 100 = 10^2, \text{ or} \\ 10 \times 10 \times 10 &= 1000 = 10^3, \text{ or} \\ 6 \times 6 \times 6 &= 216 = 6^3, \text{ or} \\ 12 \times 12 \times 12 \times 12 &= 20{,}736 = 12^4, \text{ or} \\ a \times a \times a \times a \times a &= a^5, \text{ or} \\ f \times x \times x &= fx^2 \end{aligned}$$

An exponent indicates the degree to which a number is to be dissolved or broken down into equal factors, as $\sqrt[2]{16} = 4$. This means then when 16 is dissolved into two equal factors each one of them is 4. $\sqrt[3]{27}$ means that 27 broken down into 3 equal factors equals 3, or one of the three equal factors of 27 is 3. $\sqrt[6]{64} = 2$. 2 is one of the 6 equal factors of 64. $\sqrt{\ }$ is called the radical sign and indicates that a root is to be extracted. One of the most common uses of an exponent in statistics is the binomial *expansion*, $(a + b)^n$.

N Exponent	
2	$(a + b)^2 = a^2 + 2ab + b^2$
3	$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$
4	$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$
5	$(a + b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$
6	$(a + b)^6 = a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6$
7	$(a + b)^7 = a^7 + 7a^6b + 21a^5b^2 + 35a^4b^3 + 35a^3b^4 + 21a^2b^5 + 7ab^6 + b^7$
8	$(a + b)^8 = a^8 + 8a^7b + 28a^6b^2 + 56a^5b^3 + 70a^4b^4 + 56a^3b^5 + 28a^2b^6 + 8ab^7 + b^8$
9	$(a + b)^9 = a^9 + 9a^8b + 36a^7b^2 + 84a^6b^3 + 126a^5b^4 + 126a^4b^5 + 84a^3b^6 + 36a^2b^7 + 9ab^8 + b^9$
10	$(a + b)^{10} = a^{10} + 10a^9b + 45a^8b^2 + 120a^7b^3 + 210a^6b^4 + 252a^5b^5 + 210a^4b^6 + 120a^3b^7 + 45a^2b^8 + 10ab^9 + b^{10}$

This formula $(a + b)^n$ is used to explain the development of such equations, as

$$b = \frac{\Sigma XY - NXY}{\Sigma X^2 - N(\bar{X})^2}, \quad \text{and} \quad \Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

It is also useful in explaining the shape of frequency distributions.

LOGARITHMS

Logarithms are in a sense the shorthand, at least the short cut, of mathematics and as such are of great value in statistical computations. In multiplying and dividing large numbers and especially in extracting high roots, they not only make otherwise most difficult operations quite easy, but enable us to use superior methods that would be impossible without them.

The common system of logarithms is based on 10. The log of a number is the *exponent* to which 10 must be raised to be equal to that number.

$$10 \times 10 = 100 = 10^2$$
$$10 \times 10 \times 10 = 1{,}000 = 10^3$$
$$10 \times 10 \times 10 \times 10 = 10{,}000 = 10^4$$

A log is an *exponent.* The log of 100 is 2, the log of 1,000 is 3, the log of 10,000 is 4, and the log of 1,000,000 is 6. The exponent or log of 10, 100, 1,000, 1,000,000 and all numbers consisting of 1 plus ciphers only are whole numbers, as 1, 2, 3, 6, etc. The logarithms of all other numbers, such as 12, 75, 120, 400, 7,000, etc. are mixed numbers, or contain a whole number plus a fraction. If the logarithm of 100 is 2 and of 1,000 is 3, the logarithms of all the numbers between 100 and 1,000 must be more than 2 and less than 3, or must be mixed numbers containing fractions.

Number	*Logarithm*
100	2.00000
105	2.02119
150	2.17609
500	2.69897
800	2.90309
1000	3.00000

Every logarithm is composed of two parts, the whole number which is called the *characteristic* and the fraction which is called the *mantissa.* In the log of 105 the characteristic is 2. and the mantissa is .02119. In the log of 800 the characteristic is 2. and the mantissa is .90309. In the case of 100, the characteristic is 2 and the mantissa is .00000. The mantissa, or fractional part of the logarithm, is all that is ever found in Log-Tables, or Tables of Logarithms. Although all decimal points are omitted in such tables it is understood that there is a decimal point before each mantissa in the table. They must be so used.

The characteristic is determined by the number of digits to the left of the decimal point in the original number or *anti-log* or *anti-logarithm,* as it is called. 100 is the anti-logarithm of the logarithm, 2. 150 is the anti-logarithm of the logarithm, 2.17609.

These relationships are illustrated by the following examples:

	Number	*is*	*Logarithm*
log of	732,600.		5.86487
log of	73,260.		4.86487
log of	7,326.		3.86487
log of	732.6		2.86487
log of	73.26		1.86487
log of	7.326		0.86487
log of	.7326		9.86487 − 10
log of	.07326		8.86487 − 10
log of	.007326		7.86487 − 10

The characteristic of the logarithm of the number 1 or of any number larger than 1, is *one unit less* than the number of digits to the left of the decimal point in that number, or anti-logarithm. The characteristic for 732,600 is 5. because there are six digits in this number to the left of the decimal point. The characteristic for 73.26 is 1 because there are two digits, 7 and 3, to the left of the decimal point in this number.

The characteristic of the logarithm of any number less than 1, or a fraction, is *one more than the number of zeros* between the decimal point and the first significant figure and has a minus sign. The characteristic for .7326, a fraction, is − 1, and is written 9 − 10 to facilitate addition and subtraction with other logarithms. For .07326 the characteristic is − 2, or 8 − 10, and for .007326 it is − 3, or 7 − 10.

To multiply numbers add their logarithms.

$$\begin{array}{rl} 100 = 10^2 & 25. = 10^{1.39794} \\ 1000 = 10^3 & 3. = 10^{0.47712} \\ \hline 100{,}000 = 100^5 & 75. = 10^{1.87506} \end{array}$$

The logarithm of 7,896. is 3.897407, and the logarithm of 95 is 1.977724; therefore, their product is the anti-logarithm of

$$\begin{array}{l} 3.897407 \\ 1.977724 \\ \hline 5.875131, \quad \text{or} \quad 750{,}120. \end{array}$$

To divide one number by another subtract the logarithm of the divisor from the logarithm of the dividend.

$$75 \div 3 = 25$$
$$\text{Log of } 75 = 1{,}87506$$
$$\text{Log of } 3 = \underline{.47712}$$
$$1.39794$$

The anti-logarithm of 1.39794 is 25.

1,000,000 has a log of 6
100 has a log of 2
$\overline{4}$

$1{,}000{,}000 \div 100 = 10{,}000$
The logarithm of 10,000 is 4.

To raise any number to a high power multiply its logarithm by the desired power.

$$25^2 = 625$$

The logarithm of 25 is 1.39794. $1.39794 \times 2 = 2.79588$. The anti-logarithm of 2.79588 is 625. To solve the problem, what is the cube of 77? One locates the logarithm of 77 in the table. It is 1.886491. Multiply this logarithm by three.

$$3 \times 1.886491 = 5.659473.$$

Locate the anti-logarithm in the table. The number is 456,533.

To extract any root of a number, divide the logarithm of the number by the index of the root. What is the square root of one million? The logarithm of 1,000,000 is 6. $6 \div 2 = 3$. The anti-logarithm of 3 is 1000, the square root of 1,000,000. What is the 4th root of 1,234? Its logarithm is 3.091315.

$$3.091315 \div 4 = .772829.$$

The anti-logarithm of .772829 is 5.9269.

Such operations can be performed by the use of a table of logarithms in a small fraction of the time that would otherwise be required. In fact, the extraction of roots beyond the third degree is impossible for the person with no mathematics beyond elementary algebra without logarithms. Appendix Table V in this text is a Table of Logarithms. The student should learn to use it effectively early in the course.

APPENDIX TABLE I

I. Ordinates of the Normal Probability Curve

Expressed as fractional parts of the mean ordinate y_0. Each ordinate is erected at a given distance from the mean. The height of the ordinate erected at the mean can be computed from,

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}} = \frac{N}{2.5066\sigma}$$

The corresponding height of any other ordinate can be read from the table by assigning the distance that the ordinate is from the mean (x). Distances on x are measured as fractional parts of σ. Thus the height of an ordinate at a distance from the mean of $.7\sigma$ will be $.78270y_0$; the height of an ordinate at 2.15σ from the mean will be $.09914y_0$, etc.

x/σ	0	1	2	3	4	5	6	7	8	9
0.0	100000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0.4	92312	91939	91558	91169	90774	90371	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0.8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0.9	66689	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1.1	54607	54007	53409	52812	52214	51620	51027	50437	49848	49260
1.2	48675	48092	47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38569	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11981	11737	11496	11259
2.1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2.2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07100	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01288	01252	01215	01179	01145
3.0	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.0	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.0	00000									

[1] From Rugg's *Statistical Methods Applied to Education.* Reprinted by arrangement with the publishers, Houghton Mifflin Company.

APPENDIX TABLE II

II. Areas of the Normal Curve of Error in Terms of Abscissa

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.00000	.00399	.00798	.01197	.01595	.01994	.02392	.02790	.03188	.03586
0.1	.03983	.04380	.04776	.05172	.05567	.05962	.06356	.06749	.07142	.07535
0.2	.07926	.08317	.08706	.09095	.09483	.09871	.10257	.10642	.11026	.11409
0.3	.11791	.12172	.12552	.12930	.13307	.13683	.14058	.14431	.14803	.15173
0.4	.15542	.15910	.16276	.16640	.17003	.17364	.17724	.18082	.18439	.18793
0.5	.19146	.19497	.19847	.20194	.20540	.20884	.21226	.21566	.21904	.22240
0.6	.22575	.22907	.23237	.23565	.23891	.24215	.24537	.24857	.25175	.25490
0.7	.25804	.26115	.26424	.26730	.27035	.27337	.27637	.27935	.28230	.28524
0.8	.28814	.29103	.29389	.29673	.29955	.30234	.30511	.30785	.31057	.31327
0.9	.31594	.31859	.32121	.32381	.32639	.32894	.33147	.33398	.33646	.33891
1.0	.34134	.34375	.34614	.34850	.35083	.35314	.35543	.35769	.35993	.36214
1.1	.36433	.36650	.36864	.37076	.37286	.37493	.37698	.37900	.38100	.38298
1.2	.38493	.38686	.38877	.39065	.39251	.39435	.39617	.39796	.39973	.40147
1.3	.40320	.40490	.40658	.40824	.40988	.41149	.41309	.41466	.41621	.41774
1.4	.41924	.42073	.42220	.42364	.42507	.42647	.42786	.42922	.43056	.43189
1.5	.43319	.43448	.43574	.43699	.43822	.43943	.44062	.44179	.44295	.44408
1.6	.44520	.44630	.44738	.44845	.44950	.45053	.45154	.45254	.45352	.45449
1.7	.45543	.45637	.45728	.45818	.45907	.45994	.46080	.46164	.46246	.46327
1.8	.46407	.46485	.46562	.46638	.46712	.46784	.46856	.46926	.46995	.47062
1.9	.47128	.47193	.47257	.47320	.47381	.47441	.47500	.47558	.47615	.47670
2.0	.47725	.47778	.47831	.47882	.47932	.47982	.48030	.48077	.48124	.48169
2.1	.48214	.48257	.48300	.48341	.48382	.48422	.48461	.48500	.48537	.48574
2.2	.48610	.48645	.48679	.48713	.48745	.48778	.48809	.48840	.48870	.48899
2.3	.48928	.48956	.48983	.49010	.49036	.49061	.49086	.49111	.49134	.49158
2.4	.49180	.49202	.49224	.49245	.49266	.49286	.49305	.49324	.49343	.49361
2.5	.49379	.49396	.49413	.49430	.49446	.49461	.49477	.49492	.49506	.49520
2.6	.49534	.49547	.49560	.49573	.49585	.49598	.49609	.49621	.49632	.49643
2.7	.49653	.49664	.49674	.49683	.49693	.49702	.49711	.49720	.49728	.49736
2.8	.49744	.49752	.49760	.49767	.49774	.49781	.49788	.49795	.49801	.49807
2.9	.49813	.49819	.49825	.49831	.49836	.49841	.49846	.49851	.49856	.49861
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49897	.49900
3.1	.49903	.49906	.49910	.49913	.49916	.49918	.49921	.49924	.49926	.49929
3.2	.49931	.49934	.49936	.49938	.49940	.49942	.49944	.49946	.49948	.49950
3.3	.49952	.49953	.49955	.49957	.49958	.49960	.49961	.49962	.49964	.49965
3.4	.49966	.49968	.49969	.49970	.49971	.49972	.49973	.49974	.49975	.49976

APPENDIX TABLE III

III. Showing the Relations between r and z for Values of z from 0 to 5[1]

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0100	.0200	.0300	.0400	.0500	.0599	.0699	.0798	.0898
.1	.0997	.1096	.1194	.1293	.1391	.1489	.1587	.1684	.1781	.1878
.2	.1974	.2070	.2165	.2260	.2355	.2449	.2543	.2636	.2729	.2821
.3	.2913	.3004	.3095	.3185	.3275	.3364	.3452	.3540	.3627	.3714
.4	.3800	.3885	.3969	.4053	.4136	.4219	.4301	.4382	.4462	.4542
.5	.4621	.4700	.4777	.4854	.4930	.5005	.5080	.5154	.5227	.5299
.6	.5370	.5441	.5511	.5581	.5649	.5717	.5784	.5850	.5915	.5980
.7	.6044	.6107	.6169	.6231	.6291	.6352	.6411	.6469	.6527	.6584
.8	.6640	.6696	.6751	.6805	.6858	.6911	.6963	.7014	.7064	.7114
.9	.7163	.7211	.7259	.7306	.7352	.7398	.7443	.7487	.7531	.7574
1.0	.7616	.7658	.7699	.7739	.7779	.7818	.7857	.7895	.7932	.7969
1.1	.8005	.8041	.8076	.8110	.8144	.8178	.8210	.8243	.8275	.8306
1.2	.8337	.8367	.8397	.8426	.8455	.8483	.8511	.8538	.8565	.8591
1.3	.8617	.8643	.8668	.8693	.8717	.8741	.8764	.8787	.8810	.8832
1.4	.8854	.8875	.8896	.8917	.8937	.8957	.8977	.8996	.9015	.9033
1.5	.9052	.9069	.9087	.9104	.9121	.9138	.9154	.9170	.9186	.9202
1.6	.9217	.9232	.9246	.9261	.9275	.9289	.9302	.9316	.9329	.9342
1.7	.9354	.9367	.9379	.9391	.9402	.9414	.9425	.9436	.9447	.9458
1.8	.9468	.9478	.9498	.9488	.9508	.9518	.9527	.9536	.9545	.9554
1.9	.9562	.9571	.9579	.9587	.9595	.9603	.9611	.9619	.9626	.9633
2.0	.9640	.9647	.9654	.9661	.9668	.9674	.9680	.9687	.9693	.9699
2.1	.9705	.9710	.9716	.9722	.9727	.9732	.9738	.9743	.9748	.9753
2.2	.9757	.9762	.9767	.9771	.9776	.9780	.9785	.9789	.9793	.9797
2.3	.9801	.9805	.9809	.9812	.9816	.9820	.9823	.9827	.9830	.9834
2.4	.9837	.9840	.9843	.9846	.9849	.9852	.9855	.9858	.9861	.9863
2.5	.9866	.9869	.9871	.9874	.9876	.9879	.9881	.9884	.9886	.9888
2.6	.9890	.9892	.9895	.9897	.9899	.9901	.9903	.9905	.9906	.9908
2.7	.9910	.9912	.9914	.9915	.9917	.9919	.9920	.9922	.9923	.9925
2.8	.9926	.9928	.9929	.9931	.9932	.9933	.9935	.9936	.9937	.9938
2.9	.9940	.9941	.9942	.9943	.9944	.9945	.9946	.9947	.9949	.9950
3.0	.9951									
4.0	.9993									
5.0	.9999									

[1] The figures in the body of the table are values of r corresponding to z-values read from the scales on the left and top of the table.

APPENDIX TABLE IV

IV. Squares, Square Roots, and Reciprocals to 1000[1]

No.	Square	Square Root	Reciprocal ×100	No.	Square	Square Root	Reciprocal ×100
1	1	1.0000000	100.0000000	51	26 01	7.1414284	1.9607843
2	4	1.4142136	50.0000000	52	27 04	7.2111026	1.9230769
3	9	1.7320508	33.3333333	53	28 09	7.2801099	1.8867925
4	16	2.0000000	25.0000000	54	29 16	7.3484692	1.8518519
5	25	2.2360680	20.0000000	55	30 25	7.4161985	1.8181818
6	36	2.4494897	16.6666667	56	31 36	7.4833148	1.7857143
7	49	2.6457513	14.2857143	57	32 49	7.5498344	1.7543860
8	64	2.8284271	12.5000000	58	33 64	7.6157731	1.7241379
9	81	3.0000000	11.1111111	59	34 81	7.6811457	1.6949153
10	1 00	3.1622777	10.0000000	60	36 00	7.7459667	1.6666667
11	1 21	3.3166248	9.0909091	61	37 21	7.8102497	1.6393443
12	1 44	3.4641016	8.3333333	62	38 44	7.8740079	1.6129032
13	1 69	3.6055513	7.6923077	63	39 69	7.9372539	1.5873016
14	1 96	3.7416574	7.1428571	64	40 96	8.0000000	1.5625000
15	2 25	3.8729833	6.6666667	65	42 25	8.0622577	1.5384615
16	2 56	4.0000000	6.2500000	66	43 56	8.1240384	1.5151515
17	2 89	4.1231056	5.8823529	67	44 89	8.1853528	1.4925373
18	3 24	4.2426407	5.5555556	68	46 24	8.2462113	1.4705882
19	3 61	4.3588989	5.2631579	69	47 61	8.3066239	1.4492754
20	4 00	4.4721360	5.0000000	70	49 00	8.3666003	1.4285714
21	4 41	4.5825757	4.7619048	71	50 41	8.4261498	1.4084507
22	4 84	4.6904158	4.5454545	72	51 84	8.4852814	1.3888889
23	5 29	4.7958315	4.3478261	73	53 29	8.5440037	1.3698630
24	5 76	4.8989795	4.1666667	74	54 76	8.6023253	1.3513514
25	6 25	5.0000000	4.0000000	75	56 25	8.6602540	1.3333333
26	6 76	5.0990195	3.8461538	76	57 76	8.7177979	1.3157895
27	7 29	5.1961524	3.7037037	77	59 29	8.7749644	1.2987013
28	7 84	5.2915026	3.5714286	78	60 84	8.8317609	1.2820513
29	8 41	5.3851648	3.4482759	79	62 41	8.8881944	1.2658228
30	9 00	5.4772256	3.3333333	80	64 00	8.9442719	1.2500000
31	9 61	5.5677644	3.2258065	81	65 61	9.0000000	1.2345679
32	10 24	5.6568542	3.1250000	82	67 24	9.0553851	1.2195122
33	10 89	5.7445626	3.0303030	83	68 89	9.1104336	1.2048193
34	11 56	5.8309519	2.9411765	84	70 56	9.1651514	1.1904762
35	12 25	5.9160798	2.8571429	85	72 25	9.2195445	1.1764706
36	12 96	6.0000000	2.7777778	86	73 96	9.2736185	1.1627907
37	13 69	6.0827625	2.7027027	87	75 69	9.3273791	1.1494253
38	14 44	6.1644140	2.6315789	88	77 44	9.3808315	1.1363636
39	15 21	6.2449980	2.5641026	89	79 21	9.4339811	1.1235955
40	16 00	6.3245553	2.5000000	90	81 00	9.4868330	1.1111111
41	16 81	6.4031242	2.4390244	91	82 81	9.5393920	1.0989011
42	17 64	6.4807407	2.3809524	92	84 64	9.5916630	1.0869565
43	18 49	6.5574385	2.3255814	93	86 49	9.6436508	1.0752688
44	19 36	6.6332496	2.2727273	94	88 36	9.6953597	1.0638298
45	20 25	6.7082039	2.2222222	95	90 25	9.7467943	1.0526316
46	21 16	6.7823300	2.1739130	96	92 16	9.7979590	1.0416667
47	22 09	6.8556546	2.1276596	97	94 09	9.8488578	1.0309278
48	23 04	6.9282032	2.0833333	98	96 04	9.8994949	1.0204082
49	24 01	7.0000000	2.0408163	99	98 01	9.9498744	1.0101010
50	25 00	7.0710678	2.0000000	100	1 00 00	10.0000000	1.0000000

[1] Reprinted by permission from *Business Statistics* by Davies and Yoder, published by John Wiley & Sons, Inc.

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal × 10^9	No.	Square	Square Root	Reciprocal × 10^9
101	1 02 01	10.0498756	9900990	151	2 28 01	12.2882057	6622517
102	1 04 04	10.0995049	9803922	152	2 31 04	12.3288280	6578947
103	1 06 09	10.1488916	9708738	153	2 34 09	12.3693169	6535948
104	1 08 16	10.1980390	9615385	154	2 37 16	12.4096736	6493506
105	1 10 25	10.2469508	9523810	155	2 40 25	12.4498996	6451613
106	1 12 36	10.2956301	9433962	156	2 43 36	12.4899960	6410256
107	1 14 49	10.3440804	9345794	157	2 46 49	12.5299641	6369427
108	1 16 64	10.3923048	9259259	158	2 49 64	12.5698051	6329114
109	1 18 81	10.4403065	9174312	159	2 52 81	12.6095202	6289308
110	1 21 00	10.4880885	9090909	160	2 56 00	12.6491106	6250000
111	1 23 21	10.5356538	9009009	161	2 59 21	12.6885775	6211180
112	1 25 44	10.5830052	8928571	162	2 62 44	12.7279221	6172840
113	1 27 69	10.6301458	8849558	163	2 65 69	12.7671453	6134969
114	1 29 96	10.6770783	8771930	164	2 68 96	12.8062485	6097561
115	1 32 25	10.7238053	8695652	165	2 72 25	12.8452326	6060606
116	1 34 56	10.7703296	8620690	166	2 75 56	12.8840987	6024096
117	1 36 89	10.8166538	8547009	167	2 78 89	12.9228480	5988024
118	1 39 24	10.8627805	8474576	168	2 82 24	12.9614814	5952381
119	1 41 61	10.9087121	8403361	169	2 85 61	13.0000000	5917160
120	1 44 00	10.9544512	8333333	170	2 89 00	13.0384048	5882353
121	1 46 41	11.0000000	8264463	171	2 92 41	13.0766968	5847953
122	1 48 84	11.0453610	8196721	172	2 95 84	13.1148770	5813953
123	1 51 29	11.0905365	8130081	173	2 99 29	13.1529464	5780347
124	1 53 76	11.1355287	8064516	174	3 02 76	13.1909060	5747126
125	1 56 25	11.1803399	8000000	175	3 06 25	13.2287566	5714286
126	1 58 76	11.2249722	7936508	176	3 09 76	13.2664992	5681818
127	1 61 29	11.2694277	7874016	177	3 13 29	13.3041347	5649718
128	1 63 84	11.3137085	7812500	178	3 16 84	13.3416641	5617978
129	1 66 41	11.3578167	7751938	179	3 20 41	13.3790882	5586592
130	1 69 00	11.4017543	7692308	180	3 24 00	13.4164079	5555556
131	1 71 61	11.4455231	7633588	181	3 27 61	13.4536240	5524862
132	1 74 24	11.4891253	7575758	182	3 31 24	13.4907376	5494505
133	1 76 89	11.5325626	7518797	183	3 34 89	13.5277493	5464481
134	1 79 56	11.5758369	7462687	184	3 38 56	13.5646600	5434783
135	1 82 25	11.6189500	7407407	185	3 42 25	13.6014705	5405405
136	1 84 96	11.6619038	7352941	186	3 45 96	13.6381817	5376344
137	1 87 69	11.7046999	7299270	187	3 49 69	13.6747943	5347594
138	1 90 44	11.7473401	7246377	188	3 53 44	13.7113092	5319149
139	1 93 21	11.7898261	7194245	189	3 57 21	13.7477271	5291005
140	1 96 00	11.8321596	7142857	190	3 61 00	13.7840488	5263158
141	1 98 81	11.8743421	7092199	191	3 64 81	13.8202750	5235602
142	2 01 64	11.9163753	7042254	192	3 68 64	13.8564065	5208333
143	2 04 49	11.9582607	6993007	193	3 72 49	13.8924440	5181347
144	2 07 36	12.0000000	6944444	194	3 76 36	13.9283883	5154639
145	2 10 25	12.0415946	6896552	195	3 80 25	13.9642400	5128205
146	2 13 16	12.0830460	6849315	196	3 84 16	14.0000000	5102041
147	2 16 09	12.1243557	6802721	197	3 88 09	14.0356688	5076142
148	2 19 04	12.1655251	6756757	198	3 92 04	14.0712473	5050505
149	2 22 01	12.2065556	6711409	199	3 96 01	14.1067360	5025126
150	2 25 00	12.2474487	6666667	200	4 00 00	14.1421356	5000000

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
201	4 04 01	14.1774469	4975124	251	6 30 01	15.8429795	3984064
202	4 08 04	14.2126704	4950495	252	6 35 04	15.8745079	3968254
203	4 12 09	14.2478068	4926108	253	6 40 09	15.9059737	3952569
204	4 16 16	14.2828569	4901961	254	6 45 16	15.9373775	3937008
205	4 20 25	14.3178211	4878049	255	6 50 25	15.9687194	3921569
206	4 24 36	14.3527001	4854369	256	6 55 36	16.0000000	3906250
207	4 28 49	14.3874946	4830918	257	6 60 49	16.0312195	3891051
208	4 32 64	14.4222051	4807692	258	6 65 64	16.0623784	3875969
209	4 36 81	14.4568323	4784689	259	6 70 81	16.0934769	3861004
210	4 41 00	14.4913767	4761905	260	6 76 00	16.1245155	3846154
211	4 45 21	14.5258390	4739336	261	6 81 21	16.1554944	3831418
212	4 49 44	14.5602198	4716981	262	6 86 44	16.1864141	3816794
213	4 53 69	14.5945195	4694836	263	6 91 69	16.2172747	3802281
214	4 57 96	14.6287388	4672897	264	6 96 96	16.2480768	3787879
215	4 62 25	14.6628783	4651163	265	7 02 25	16.2788206	3773585
216	4 66 56	14.6969385	4629630	266	7 07 56	16.3095064	3759398
217	4 70 89	14.7309199	4608295	267	7 12 89	16.3401346	3745318
218	4 75 24	14.7648231	4587156	268	7 18 24	16.3707055	3731343
219	4 79 61	14.7986486	4566210	269	7 23 61	16.4012195	3717472
220	4 84 00	14.8323970	4545455	270	7 29 00	16.4316767	3703704
221	4 88 41	14.8660687	4524887	271	7 34 41	16.4620776	3690037
222	4 92 84	14.8996644	4504505	272	7 39 84	16.4924225	3676471
223	4 97 29	14.9331845	4484305	273	7 45 29	16.5227116	3663004
224	5 01 76	14.9666295	4464286	274	7 50 76	16.5529454	3649635
225	5 06 25	15.0000000	4444444	275	7 56 25	16.5831240	3636364
226	5 10 76	15.0332964	4424779	276	7 61 76	16.6132477	3623188
227	5 15 29	15.0665192	4405286	277	7 67 29	16.6433170	3610108
228	5 19 84	15.0996689	4385965	278	7 72 84	16.6733320	3597122
229	5 24 41	15.1327460	4366812	279	7 78 41	16.7032931	3584229
230	5 29 00	15.1657509	4347826	280	7 84 00	16.7332005	3571429
231	5 33 61	15.1986842	4329004	281	7 89 61	16.7630546	3558719
232	5 38 24	15.2315462	4310345	282	7 95 24	16.7928556	3546099
233	5 42 89	15.2643375	4291845	283	8 00 89	16.8226038	3533569
234	5 47 56	15.2970585	4273504	284	8 06 56	16.8522995	3521127
235	5 52 25	15.3297097	4255319	285	8 12 25	16.8819430	3508772
236	5 56 96	15.3622915	4237288	286	8 17 96	16.9115345	3496503
237	5 61 69	15.3948043	4219409	287	8 23 69	16.9410743	3484321
238	5 66 44	15.4272486	4201681	288	8 29 44	16.9705627	3472222
239	5 71 21	15.4596248	4184100	289	8 35 21	17.0000000	3460208
240	5 76 00	15.4919334	4166667	290	8 41 00	17.0293864	3448276
241	5 80 81	15.5241747	4149378	291	8 46 81	17.0587221	3436426
242	5 85 64	15.5563492	4132231	292	8 52 64	17.0880075	3424658
243	5 90 49	15.5884573	4115226	293	8 58 49	17.1172428	3412969
244	5 95 36	15.6204994	4098361	294	8 64 36	17.1464282	3401361
245	6 00 25	15.6524758	4081633	295	8 70 25	17.1755640	3389831
246	6 05 16	15.6843871	4065041	296	8 76 16	17.2046505	3378378
247	6 10 09	15.7162336	4048583	297	8 82 09	17.2336879	3367003
248	6 15 04	15.7480157	4032258	298	8 88 04	17.2626765	3355705
249	6 20 01	15.7797338	4016064	299	8 94 01	17.2916165	3344482
250	6 25 00	15.8113883	4000000	300	9 00 00	17.3205081	3333333

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
301	9 06 01	17.3493516	3322259	351	12 32 01	18.7349940	2849003
302	9 12 04	17.3781472	3311258	352	12 39 04	18.7616630	2840909
303	9 18 09	17.4068952	3300330	353	12 46 09	18.7882942	2832861
304	9 24 16	17.4355958	3289474	354	12 53 16	18.8148877	2824859
305	9 30 25	17.4642492	3278689	355	12 60 25	18.8414437	2816901
306	9 36 36	17.4928557	3267974	356	12 67 36	18.8679623	2808989
307	9 42 49	17.5214155	3257329	357	12 74 49	18.8944436	2801120
308	9 48 64	17.5499288	3246753	358	12 81 64	18.9208879	2793296
309	9 54 81	17.5783958	3236246	359	12 88 81	18.9472953	2785515
310	9 61 00	17.6068169	3225806	360	12 96 00	18.9736660	2777778
311	9 67 21	17.6351921	3215434	361	13 03 21	19.0000000	2770083
312	9 73 44	17.6635217	3205128	362	13 10 44	19.0262976	2762431
313	9 79 69	17.6918060	3194888	363	13 17 69	19.0525589	2754821
314	9 85 96	17.7200451	3184713	364	13 24 96	19.0787840	2747253
315	9 92 25	17.7482393	3174603	365	13 32 25	19.1049732	2739726
316	9 98 56	17.7763888	3164557	366	13 39 56	19.1311265	2732240
317	10 04 89	17.8044938	3154574	367	13 46 89	19.1572441	2724796
318	10 11 24	17.8325545	3144654	368	13 54 24	19.1833261	2717391
319	10 17 61	17.8605711	3134796	369	13 61 61	19.2093727	2710027
320	10 24 00	17.8885438	3125000	370	13 69 00	19.2353841	2702703
321	10 30 41	17.9164729	3115265	371	13 76 41	19.2613603	2695418
322	10 36 84	17.9443584	3105590	372	13 83 84	19.2873015	2688172
323	10 43 29	17.9722008	3095975	373	13 91 29	19.3132079	2680965
324	10 49 76	18.0000000	3086420	374	13 98 76	19.3390796	2673797
325	10 56 25	18.0277564	3076923	375	14 06 25	19.3649167	2666667
326	10 62 76	18.0554701	3067485	376	14 13 76	19.3907194	2659574
327	10 69 29	18.0831413	3058104	377	14 21 29	19.4164878	2652520
328	10 75 84	18.1107703	3048780	378	14 28 84	19.4422221	2645503
329	10 82 41	18.1383571	3039514	379	14 36 41	19.4679223	2638522
330	10 89 00	18.1659021	3030303	380	14 44 00	19.4935887	2631579
331	10 95 61	18.1934054	3021148	381	14 51 61	19.5192213	2624672
332	11 02 24	18.2208672	3012048	382	14 59 24	19.5448203	2617801
333	11 08 89	18.2482876	3003003	383	14 66 89	19.5703858	2610966
334	11 15 56	18.2756669	2994012	384	14 74 56	19.5959179	2604167
335	11 22 25	18.3030052	2985075	385	14 82 25	19.6214169	2597403
336	11 28 96	18.3303028	2976190	386	14 89 96	19.6468827	2590674
337	11 35 69	18.3575598	2967359	387	14 97 69	19.6723156	2583979
338	11 42 44	18.3847763	2958580	388	15 05 44	19.6977156	2577320
339	11 49 21	18.4119526	2949853	389	15 13 21	19.7230829	2570694
340	11 56 00	18.4390889	2941176	390	15 21 00	19.7484177	2564103
341	11 62 81	18.4661853	2932551	391	15 28 81	19.7737199	2557545
342	11 69 64	18.4932420	2923977	392	15 36 64	19.7989899	2551020
343	11 76 49	18.5202592	2915452	393	15 44 49	19.8242276	2544529
344	11 83 36	18.5472370	2906977	394	15 52 36	19.8494332	2538071
345	11 90 25	18.5741756	2898551	395	15 60 25	19.8746069	2531646
346	11 97 16	18.6010752	2890173	396	15 68 16	19.8997487	2525253
347	12 04 09	18.6279360	2881844	397	15 76 09	19.9248588	2518892
348	12 11 04	18.6547581	2873563	398	15 84 04	19.9499373	2512563
349	12 18 01	18.6815417	2865330	399	15 92 01	19.9749844	2506266
350	12 25 00	18.7082869	2857143	400	16 00 00	20.0000000	2500000

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
401	16 08 01	20.0249844	2493766	451	20 34 01	21.2367606	2217295
402	16 16 04	20.0499377	2487562	452	20 43 04	21.2602916	2212389
403	16 24 09	20.0748599	2481390	453	20 52 09	21.2837967	2207506
404	16 32 16	20.0997512	2475248	454	20 61 16	21.3072758	2202643
405	16 40 25	20.1246118	2469136	455	20 70 25	21.3307290	2197802
406	16 48 36	20.1494417	2463054	456	20 79 36	21.3541565	2192982
407	16 56 49	20.1742410	2457002	457	20 88 49	21.3775583	2188184
408	16 64 64	20.1990099	2450980	458	20 97 64	21.4009346	2183406
409	16 72 81	20.2237484	2444988	459	21 06 81	21.4242853	2178649
410	16 81 00	20.2484567	2439024	460	21 16 00	21.4476106	2173913
411	16 89 21	20.2731349	2433090	461	21 25 21	21.4709106	2169197
412	16 97 44	20.2977831	2427184	462	21 34 44	21.4941853	2164502
413	17 05 69	20.3224014	2421308	463	21 43 69	21.5174348	2159827
414	17 13 96	20.3469899	2415459	464	21 52 96	21.5406592	2155172
415	17 22 25	20.3715488	2409639	465	21 62 25	21.5638587	2150538
416	17 30 56	20.3960781	2403846	466	21 71 56	21.5870331	2145923
417	17 38 89	20.4205779	2398082	467	21 80 89	21.6101828	2141328
418	17 47 24	20.4450483	2392344	468	21 90 24	21.6333077	2136752
419	17 55 61	20.4694895	2386635	469	21 99 61	21.6564078	2132196
420	17 64 00	20.4939015	2380952	470	22 09 00	21.6794834	2127660
421	17 72 41	20.5182845	2375297	471	22 18 41	21.7025344	2123142
422	17 80 84	20.5426386	2369668	472	22 27 84	21.7255610	2118644
423	17 89 29	20.5669638	2364066	473	22 37 29	21.7485632	2114165
424	17 97 76	20.5912603	2358491	474	22 46 76	21.7715411	2109705
425	18 06 25	20.6155281	2352941	475	22 56 25	21.7944947	2105263
426	18 14 76	20.6397674	2347418	476	22 65 76	21.8174242	2100840
427	18 23 29	20.6639783	2341920	477	22 75 29	21.8403297	2096436
428	18 31 84	20.6881609	2336449	478	22 84 84	21.8632111	2092050
429	18 40 41	20.7123152	2331002	479	22 94 41	21.8860686	2087683
430	18 49 00	20.7364414	2325581	480	23 04 00	21.9089023	2083333
431	18 57 61	20.7605395	2320186	481	23 13 61	21.9317122	2079002
432	18 66 24	20.7846097	2314815	482	23 23 24	21.9544984	2074689
433	18 74 89	20.8086520	2309469	483	23 32 89	21.9772610	2070393
434	18 83 56	20.8326667	2304147	484	23 42 56	22.0000000	2066116
435	18 92 25	20.8566536	2298851	485	23 52 25	22.0227155	2061856
436	19 00 96	20.8806130	2293578	486	23 61 96	22.0454077	2057613
437	19 09 69	20.9045450	2288330	487	23 71 69	22.0680765	2053388
438	19 18 44	20.9284495	2283105	488	23 81 44	22.0907220	2049180
439	19 27 21	20.9523268	2277904	489	23 91 21	22.1133444	2044990
440	19 36 00	20.9761770	2272727	490	24 01 00	22.1359436	2040816
441	19 44 81	21.0000000	2267574	491	24 10 81	22.1585198	2036660
442	19 53 64	21.0237960	2262443	492	24 20 64	22.1810730	2032520
443	19 62 49	21.0475652	2257336	493	24 30 49	22.2036033	2028398
444	19 71 36	21.0713075	2252252	494	24 40 36	22.2261108	2024291
445	19 80 25	21.0950231	2247191	495	24 50 25	22.2485955	2020202
446	19 89 16	21.1187121	2242152	496	24 60 16	22.2710575	2016129
447	19 98 09	21.1423745	2237136	497	24 70 09	22.2934968	2012072
448	20 07 04	21.1660105	2232143	498	24 80 04	22.3159136	2008032
449	20 16 01	21.1896201	2227171	499	24 90 01	22.3383079	2004008
450	20 25 00	21.2132034	2222222	500	25 00 00	22.3606798	2000000

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal × 10^9	No.	Square	Square Root	Reciprocal × 10^9
501	25 10 01	22.3830293	1996008	551	30 36 01	23.4733892	1814882
502	25 20 04	22.4053565	1992032	552	30 47 04	23.4946802	1811594
503	25 30 09	22.4276615	1988072	553	30 58 09	23.5159520	1808318
504	25 40 16	22.4499443	1984127	554	30 69 16	23.5372046	1805054
505	25 50 25	22.4722051	1980198	555	30 80 25	23.5584380	1801802
506	25 60 36	22.4944438	1976285	556	30 91 36	23.5796522	1798561
507	25 70 49	22.5166605	1972387	557	31 02 49	23.6008474	1795332
508	25 80 64	22.5388553	1968504	558	31 13 64	23.6220236	1792115
509	25 90 81	22.5610283	1964637	559	31 24 81	23.6431808	1788909
510	26 01 00	22.5831796	1960784	560	31 36 00	23.6643191	1785714
511	26 11 21	22.6053091	1956947	561	31 47 21	23.6854386	1782531
512	26 21 44	22.6274170	1953125	562	31 58 44	23.7065392	1779359
513	26 31 69	22.6495033	1949318	563	31 69 69	23.7276210	1776199
514	26 41 96	22.6715681	1945525	564	31 80 96	23.7486842	1773050
515	26 52 25	22.6936114	1941748	565	31 92 25	23.7697286	1769912
516	26 62 56	22.7156334	1937984	566	32 03 56	23.7907545	1766784
517	26 72 89	22.7376340	1934236	567	32 14 89	23.8117618	1763668
518	26 83 24	22.7596134	1930502	568	32 26 24	23.8327506	1760563
519	26 93 61	22.7815715	1926782	569	32 37 61	23.8537209	1757469
520	27 04 00	22.8035085	1923077	570	32 49 00	23.8746728	1754386
521	27 14 41	22.8254244	1919386	571	32 60 41	23.8956063	1751313
522	27 24 84	22.8473193	1915709	572	32 71 84	23.9165215	1748252
523	27 35 29	22.8691933	1912046	573	32 83 29	23.9374184	1745201
524	27 45 76	22.8910463	1908397	574	32 94 76	23.9582971	1742160
525	27 56 25	22.9128785	1904762	575	33 06 25	23.9791576	1739130
526	27 66 76	22.9346899	1901141	576	33 17 76	24.0000000	1736111
527	27 77 29	22.9564806	1897533	577	33 29 29	24.0208243	1733102
528	27 87 84	22.9782506	1893939	578	33 40 84	24.0416306	1730104
529	27 98 41	23.0000000	1890359	579	33 52 41	24.0624188	1727116
530	28 09 00	23.0217289	1886792	580	33 64 00	24.0831892	1724138
531	28 19 61	23.0434372	1883239	581	33 75 61	24.1039416	1721170
532	28 30 24	23.0651252	1879699	582	33 87 24	24.1246762	1718213
533	28 40 89	23.0867928	1876173	583	33 98 89	24.1453929	1715266
534	28 51 56	23.1084400	1872659	584	34 10 56	24.1660919	1712329
535	28 62 25	23.1300670	1869159	585	34 22 25	24.1867732	1709402
536	28 72 96	23.1516738	1865672	586	34 33 96	24.2074369	1706485
537	28 83 69	23.1732605	1862197	587	34 45 69	24.2280829	1703578
538	28 94 44	23.1948270	1858736	588	34 57 44	24.2487113	1700680
539	29 05 21	23.2163735	1855288	589	34 69 21	24.2693222	1697793
540	29 16 00	23.2379001	1851852	590	34 81 00	24.2899156	1694915
541	29 26 81	23.2594067	1848429	591	34 92 81	24.3104916	1692047
542	29 37 64	23.2808935	1845018	592	35 04 64	24.3310501	1689189
543	29 48 49	23.3023604	1841621	593	35 16 49	24.3515913	1686341
544	29 59 36	23.3238076	1838235	594	35 28 36	24.3721152	1683502
545	29 70 25	23.3452351	1834862	595	35 40 25	24.3926218	1680672
546	29 81 16	23.3666429	1831502	596	35 52 16	24.4131112	1677852
547	29 92 09	23.3880311	1828154	597	35 64 09	24.4335834	1675042
548	30 03 04	23.4093998	1824818	598	35 76 04	24.4540385	1672241
549	30 14 01	23.4307490	1821494	599	35 88 01	24.4744765	1669449
550	30 25 00	23.4520788	1818182	600	36 00 00	24.4948974	1666667

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
601	36 12 01	24.5153013	1663894	651	42 38 01	25.5147016	1536098
602	36 24 04	24.5356883	1661130	652	42 51 04	25.5342907	1533742
603	36 36 09	24.5560583	1658375	653	42 64 09	25.5538647	1531394
604	36 48 16	24.5764115	1655629	654	42 77 16	25.5734237	1529052
605	36 60 25	24.5967478	1652893	655	42 90 25	25.5929678	1526718
606	36 72 36	24.6170673	1650165	656	43 03 36	25.6124969	1524390
607	36 84 49	24.6373700	1647446	657	43 16 49	25.6320112	1522070
608	36 96 64	24.6576560	1644737	658	43 29 64	25.6515107	1519757
609	37 08 81	24.6779254	1642036	659	43 42 81	25.6709953	1517451
610	37 21 00	24.6981781	1639344	660	43 56 00	25.6904652	1515152
611	37 33 21	24.7184142	1636661	661	43 69 21	25.7099203	1512859
612	37 45 44	24.7386338	1633987	662	43 82 44	25.7293607	1510574
613	37 57 69	24.7588368	1631321	663	43 95 69	25.7487864	1508296
614	37 69 96	24.7790234	1628664	664	44 08 96	25.7681975	1506024
615	37 82 25	24.7991935	1626016	665	44 22 25	25.7875939	1503759
616	37 94 56	24.8193473	1623377	666	44 35 56	25.8069758	1501502
617	38 06 89	24.8394847	1620746	667	44 48 89	25.8263431	1499250
618	38 19 24	24.8596058	1618123	668	44 62 24	25.8456960	1497006
619	38 31 61	24.8797106	1615509	669	44 75 61	25.8650343	1494768
620	38 44 00	24.8997992	1612903	670	44 89 00	25.8843582	1492537
621	38 56 41	24.9198716	1610306	671	45 02 41	25.9036677	1490313
622	38 68 84	24.9399278	1607717	672	45 15 84	25.9229628	1488095
623	38 81 29	24.9599679	1605136	673	45 29 29	25.9422435	1485884
624	38 93 76	24.9799920	1602564	674	45 42 76	25.9615100	1483680
625	39 06 25	25.0000000	1600000	675	45 56 25	25.9807621	1481481
626	39 18 76	25.0199920	1597444	676	45 69 76	26.0000000	1479290
627	39 31 29	25.0399681	1594896	677	45 83 29	26.0192237	1477105
628	39 43 84	25.0599282	1592357	678	45 96 84	26.0384331	1474926
629	39 56 41	25.0798724	1589825	679	46 10 41	26.0576284	1472754
630	39 69 00	25.0998008	1587302	680	46 24 00	26.0768096	1470588
631	39 81 61	25.1197134	1584786	681	46 37 61	26.0959767	1468429
632	39 94 24	25.1396102	1582278	682	46 51 24	26.1151297	1466276
633	40 06 89	25.1594913	1579779	683	46 64 89	26.1342687	1464129
634	40 19 56	25.1793566	1577287	684	46 78 56	26.1533937	1461988
635	40 32 25	25.1992063	1574803	685	46 92 25	26.1725047	1459854
636	40 44 96	25.2190404	1572327	686	47 05 96	26.1916017	1457726
637	40 57 69	25.2388589	1569859	687	47 19 69	26.2106848	1455604
638	40 70 44	25.2586619	1567398	688	47 33 44	26.2297541	1453488
639	40 83 21	25.2784493	1564945	689	47 47 21	26.2488095	1451379
640	40 96 00	25.2982213	1562500	690	47 61 00	26.2678511	1449275
641	41 08 81	25.3179778	1560062	691	47 74 81	26.2868789	1447178
642	41 21 64	25.3377189	1557632	692	47 88 64	26.3058929	1445087
643	41 34 49	25.3574447	1555210	693	48 02 49	26.3248932	1443001
644	41 47 36	25.3771551	1552795	694	48 16 36	26.3438797	1440922
645	41 60 25	25.3968502	1550388	695	48 30 25	26.3628527	1438849
646	41 73 16	25.4165301	1547988	696	48 44 16	26.3818119	1436782
647	41 86 09	25.4361947	1545595	697	48 58 09	26.4007576	1434720
648	41 99 04	25.4558441	1543210	698	48 72 04	26.4196896	1432665
649	42 12 01	25.4754784	1540832	699	48 86 01	26.4386081	1430615
650	42 25 00	25.4950976	1538462	700	49 00 00	26.4575131	1428571

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
701	49 14 01	26.4764046	1426534	751	56 40 01	27.4043792	1331558
702	49 28 04	26.4952826	1424501	752	56 55 04	27.4226184	1329787
703	49 42 09	26.5141472	1422475	753	56 70 09	27.4408455	1328021
704	49 56 16	26.5329983	1420455	754	56 85 16	27.4590604	1326260
705	49 70 25	26.5518361	1418440	755	57 00 25	27.4772633	1324503
706	49 84 36	26.5706605	1416431	756	57 15 36	27.4954542	1322751
707	49 98 49	26.5894716	1414427	757	57 30 49	27.5136330	1321004
708	50 12 64	26.6082694	1412429	758	57 45 64	27.5317998	1319261
709	50 26 81	26.6270539	1410437	759	57 60 81	27.5499546	1317523
710	50 41 00	26.6458252	1408451	760	57 76 00	27.5680975	1315789
711	50 55 21	26.6645833	1406470	761	57 91 21	27.5862284	1314060
712	50 69 44	26.6833281	1404494	762	58 06 44	27.6043475	1312336
713	50 83 69	26.7020598	1402525	763	58 21 69	27.6224546	1310616
714	50 97 96	26.7207784	1400560	764	58 36 96	27.6405499	1308901
715	51 12 25	26.7394839	1398601	765	58 52 25	27.6586334	1307190
716	51 26 56	26.7581763	1396648	766	58 67 56	27.6767050	1305483
717	51 40 89	26.7768557	1394700	767	58 82 89	27.6947648	1303781
718	51 55 24	26.7955220	1392758	768	58 98 24	27.7128129	1302083
719	51 69 61	26.8141754	1390821	769	59 13 61	27.7308492	1300390
720	51 84 00	26.8328157	1388889	770	59 29 00	27.7488739	1298701
721	51 98 41	26.8514432	1386963	771	59 44 41	27.7668868	1297017
722	52 12 84	26.8700577	1385042	772	59 59 84	27.7848880	1295337
723	52 27 29	26.8886593	1383126	773	59 75 29	27.8028775	1293661
724	52 41 76	26.9072481	1381215	774	59 90 76	27.8208555	1291990
725	52 56 25	26.9258240	1379310	775	60 06 25	27.8388218	1290323
726	52 70 76	26.9443872	1377410	776	60 21 76	27.8567766	1288660
727	52 85 29	26.9629375	1375516	777	60 37 29	27.8747197	1287001
728	52 99 84	26.9814751	1373626	778	60 52 84	27.8926514	1285347
729	53 14 41	27.0000000	1371742	779	60 68 41	27.9105715	1283697
730	53 29 00	27.0185122	1369863	780	60 84 00	27.9284801	1282051
731	53 43 61	27.0370117	1367989	781	60 99 61	27.9463772	1280410
732	53 58 24	27.0554985	1366120	782	61 15 24	27.9642629	1278772
733	53 72 89	27.0739727	1364256	783	61 30 89	27.9821372	1277139
734	53 87 56	27.0924344	1362398	784	61 46 56	28.0000000	1275510
735	54 02 25	27.1108834	1360544	785	61 62 25	28.0178515	1273885
736	54 16 96	27.1293199	1358696	786	61 77 96	28.0356915	1272265
737	54 31 69	27.1477439	1356852	787	61 93 69	28.0535203	1270648
738	54 46 44	27.1661554	1355014	788	62 09 44	28.0713377	1269036
739	54 61 21	27.1845544	1353180	789	62 25 21	28.0891438	1267427
740	54 76 00	27.2029410	1351351	790	62 41 00	28.1069386	1265823
741	54 90 81	27.2213152	1349528	791	62 56 81	28.1247222	1264223
742	55 05 64	27.2396769	1347709	792	62 72 64	28.1424946	1262626
743	55 20 49	27.2580263	1345895	793	62 88 49	28.1602557	1261034
744	55 35 36	27.2763634	1344086	794	63 04 36	28.1780056	1259446
745	55 50 25	27.2946881	1342282	795	63 20 25	28.1957444	1257862
746	55 65 16	27.3130006	1340483	796	63 36 16	28.2134720	1256281
747	55 80 09	27.3313007	1338688	797	63 52 09	28.2311884	1254705
748	55 95 04	27.3495887	1336898	798	63 68 04	28.2488938	1253133
749	56 10 01	27.3678644	1335113	799	63 84 01	28.2665881	1251564
750	56 25 00	27.3861279	1333333	800	64 00 00	28.2842712	1250000

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
801	64 16 01	28.3019434	1248439	851	72 42 01	29.1719043	1175088
802	64 32 04	28.3196045	1246883	852	72 59 04	29.1890390	1173709
803	64 48 09	28.3372546	1245330	853	72 76 09	29.2061637	1172333
804	64 64 16	28.3548938	1243781	854	72 93 16	29.2232784	1170960
805	64 80 25	28.3725219	1242236	855	73 10 25	29.2403830	1169591
806	64 96 36	28.3901391	1240695	856	73 27 36	29.2574777	1168224
807	65 12 49	28.4077454	1239157	857	73 44 49	29.2745623	1166861
808	65 28 64	28.4253408	1237624	858	73 61 64	29.2916370	1165501
809	65 44 81	28.4429253	1236094	859	73 78 81	29.3087018	1164144
810	65 61 00	28.4604989	1234568	860	73 96 00	29.3257566	1162791
811	65 77 21	28.4780617	1233046	861	74 13 21	29.3428015	1161440
812	65 93 44	28.4956137	1231527	862	74 30 44	29.3598365	1160093
813	66 09 69	28.5131549	1230012	863	74 47 69	29.3768616	1158749
814	66 25 96	28.5306852	1228501	864	74 64 96	29.3938769	1157407
815	66 42 25	28.5482048	1226994	865	74 82 25	29.4108823	1156069
816	66 58 56	28.5657137	1225490	866	74 99 56	29.4278779	1154734
817	66 74 89	28.5832119	1223990	867	75 16 89	29.4448637	1153403
818	66 91 24	28.6006993	1222494	868	75 34 24	29.4618397	1152074
819	67 07 61	28.6181760	1221001	869	75 51 61	29.4788059	1150748
820	67 24 00	28.6356421	1219512	870	75 69 00	29.4957624	1149425
821	67 40 41	28.6530976	1218027	871	75 86 41	29.5127091	1148106
822	67 56 84	28.6705424	1216545	872	76 03 84	29.5296461	1146789
823	67 73 29	28.6879766	1215067	873	76 21 29	29.5465734	1145475
824	67 89 76	28.7054002	1213592	874	76 38 76	29.5634910	1144165
825	68 06 25	28.7228132	1212121	875	76 56 25	29.5803989	1142857
826	68 22 76	28.7402157	1210654	876	76 73 76	29.5972972	1141553
827	68 39 29	28.7576077	1209190	877	76 91 29	29.6141858	1140251
828	68 55 84	28.7749891	1207729	878	77 08 84	29.6310648	1138952
829	68 72 41	28.7923601	1206273	879	77 26 41	29.6479342	1137656
830	68 89 00	28.8097206	1204819	880	77 44 00	29.6647939	1136364
831	69 05 61	28.8270706	1203369	881	77 61 61	29.6816442	1135074
832	69 22 24	28.8444102	1201923	882	77 79 24	29.6984848	1133787
833	69 38 89	28.8617394	1200480	883	77 96 89	29.7153159	1132503
834	69 55 56	28.8790582	1199041	884	78 14 56	29.7321375	1131222
835	69 72 25	28.8963666	1197605	885	78 32 25	29.7489496	1129944
836	69 88 96	28.9136646	1196172	886	78 49 96	29.7657521	1128668
837	70 05 69	28.9309523	1194743	887	78 67 69	29.7825452	1127396
838	70 22 44	28.9482297	1193317	888	78 85 44	29.7993289	1126126
839	70 39 21	28.9654967	1191895	889	79 03 21	29.8161030	1124859
840	70 56 00	28.9827535	1190476	890	79 21 00	29.8328678	1123596
841	70 72 81	29.0000000	1189061	891	79 38 81	29.8496231	1122334
842	70 89 64	29.0172363	1187648	892	79 56 64	29.8663690	1121076
843	71 06 49	29.0344623	1186240	893	79 74 49	29.8831056	1119821
844	71 23 36	29.0516781	1184834	894	79 92 36	29.8998328	1118568
845	71 40 25	29.0688837	1183432	895	80 10 25	29.9165506	1117318
846	71 57 16	29.0860791	1182033	896	80 28 16	29.9332591	1116071
847	71 74 09	29.1032644	1180638	897	80 46 09	29.9499583	1114827
848	71 91 04	29.1204396	1179245	898	80 64 04	29.9666481	1113586
849	72 08 01	29.1376046	1177856	899	80 82 01	29.9833287	1112347
850	72 25 00	29.1547595	1176471	900	81 00 00	30.0000000	1111111

IV. Squares, Square Roots, and Reciprocals to 1000 (*Cont.*)

No.	Square	Square Root	Reciprocal $\times 10^9$	No.	Square	Square Root	Reciprocal $\times 10^9$
901	81 18 01	30.0166620	1109878	951	90 44 01	30.8382879	1051525
902	81 36 04	30.0333148	1108647	952	90 63 04	30.8544972	1050420
903	81 54 09	30.0499584	1107420	953	90 82 09	30.8706981	1049318
904	81 72 16	30.0665928	1106195	954	91 01 16	30.8868904	1048218
905	81 90 25	30.0832179	1104972	955	91 20 25	30.9030743	1047120
906	82 08 36	30.0998339	1103753	956	91 39 36	30.9192497	1046025
907	82 26 49	30.1164407	1102536	957	91 58 49	30.9354166	1044932
908	82 44 64	30.1330383	1101322	958	91 77 64	30.9515751	1043841
909	82 62 81	30.1496269	1100110	959	91 96 81	30.9677251	1042753
910	82 81 00	30.1662063	1098901	960	92 16 00	30.9838668	1041667
911	82 99 21	30.1827765	1097695	961	92 35 21	31.0000000	1040583
912	83 17 44	30.1993377	1096491	962	92 54 44	31.0161248	1039501
913	83 35 69	30.2158899	1095290	963	92 73 69	31.0322413	1038422
914	83 53 96	30.2324329	1094092	964	92 92 96	31.0483494	1037344
915	83 72 25	30.2489669	1092896	965	93 12 25	31.0644491	1036269
916	83 90 56	30.2654919	1091703	966	93 31 56	31.0805405	1035197
917	84 08 89	30.2820079	1090513	967	93 50 89	31.0966236	1034126
918	84 27 24	30.2985148	1089325	968	93 70 24	31.1126984	1033058
919	84 45 61	30.3150128	1088139	969	93 89 61	31.1287648	1031992
920	84 64 00	30.3315018	1086957	970	94 09 00	31.1448230	1030928
921	84 82 41	30.3479818	1085776	971	94 28 41	31.1608729	1029866
922	85 00 84	30.3644529	1084599	972	94 47 84	31.1769145	1028807
923	85 19 29	30.3809151	1083424	973	94 67 29	31.1929479	1027749
924	85 37 76	30.3973683	1082251	974	94 86 76	31.2089731	1026694
925	85 56 25	30.4138127	1081081	975	95 06 25	31.2249900	1025641
926	85 74 76	30.4302481	1079914	976	95 25 76	31.2409987	1024590
927	85 93 29	30.4466747	1078749	977	95 45 29	31.2569992	1023541
928	86 11 84	30.4630924	1077586	978	95 64 84	31.2729915	1022495
929	86 30 41	30.4795013	1076426	979	95 84 41	31.2889757	1021450
930	86 49 00	30.4959014	1075269	980	96 04 00	31.3049517	1020408
931	86 67 61	30.5122926	1074114	981	96 23 61	31.3209195	1019368
932	86 86 24	30.5286750	1072961	982	96 43 24	31.3368792	1018330
933	87 04 89	30.5450487	1071811	983	96 62 89	31.3528308	1017294
934	87 23 56	30.5614136	1070664	984	96 82 56	31.3687743	1016260
935	87 42 25	30.5777697	1069519	985	97 02 25	31.3847097	1015228
936	87 60 96	30.5941171	1068376	986	97 21 96	31.4006369	1014199
937	87 79 69	30.6104557	1067236	987	97 41 69	31.4165561	1013171
938	87 98 44	30.6267857	1066098	988	97 61 44	31.4324673	1012146
939	88 17 21	30.6431069	1064963	989	97 81 21	31.4483704	1011122
940	88 36 00	30.6594194	1063830	990	98 01 00	31.4642654	1010101
941	88 54 81	30.6757233	1062699	991	98 20 81	31.4801525	1009082
942	88 73 64	30.6920185	1061571	992	98 40 64	31.4960315	1008065
943	88 92 49	30.7083051	1060445	993	98 60 49	31.5119025	1007049
944	89 11 36	30.7245830	1059322	994	98 80 36	31.5277655	1006036
945	89 30 25	30.7408523	1058201	995	99 00 25	31.5436206	1005025
946	89 49 16	30.7571130	1057082	996	99 20 16	31.5594677	1004016
947	89 68 09	30.7733651	1055966	997	99 40 09	31.5753068	1003009
948	89 87 04	30.7896086	1054852	998	99 60 04	31.5911380	1002004
949	90 06 01	30.8058436	1053741	999	99 80 01	31.6069613	1001001
950	90 25 00	30.8220700	1052632	1000	1 00 00 00	31.6227766	1000000

APPENDIX TABLE V

COMMON LOGARITHMS OF NUMBERS

FROM 1 TO 10000

TO FIVE DECIMAL PLACES

1–100

N	Log	N	Log	N	Log	N	Log	N	Log
0	—	**20**	1.30 103	**40**	1.60 206	**60**	1.77 815	**80**	1.90 309
1	0.00 000	21	1.32 222	41	1.61 278	61	1.78 533	81	1.90 849
2	0.30 103	22	1.34 242	42	1.62 325	62	1.79 239	82	1.91 381
3	0.47 712	23	1.36 173	43	1.63 347	63	1.79 934	83	1.91 908
4	0.60 206	24	1.38 021	44	1.64 345	64	1.80 618	84	1.92 428
5	0.69 897	25	1.39 794	45	1.65 321	65	1.81 291	85	1.92 942
6	0.77 815	26	1.41 497	46	1.66 276	66	1.81 954	86	1.93 450
7	0.84 510	27	1.43 136	47	1.67 210	67	1.82 607	87	1.93 952
8	0.90 309	28	1.44 716	48	1.68 124	68	1.83 251	88	1.94 448
9	0.95 424	29	1.46 240	49	1.69 020	69	1.83 885	89	1.94 939
10	1.00 000	**30**	1.47 712	**50**	1.69 897	**70**	1.84 510	**90**	1.95 424
11	1.04 139	31	1.49 136	51	1.70 757	71	1.85 126	91	1.95 904
12	1.07 918	32	1.50 515	52	1.71 600	72	1.85 733	92	1.96 379
13	1.11 394	33	1.51 851	53	1.72 428	73	1.86 332	93	1.96 848
14	1.14 613	34	1.53 148	54	1.73 239	74	1.86 923	94	1.97 313
15	1.17 609	35	1.54 407	55	1.74 036	75	1.87 506	95	1.97 772
16	1.20 412	36	1.55 630	56	1.74 819	76	1.88 081	96	1.98 227
17	1.23 045	37	1.56 820	57	1.75 587	77	1.88 649	97	1.98 677
18	1.25 527	38	1.57 978	58	1.76 343	78	1.89 209	98	1.99 123
19	1.27 875	39	1.59 106	59	1.77 085	79	1.89 763	99	1.99 564
20	1.30 103	**40**	1.60 206	**60**	1.77 815	**80**	1.90 309	**100**	2.00 000

N	0	1	2	3	4	5	6	7	8	9
100	00 000	043	087	130	173	217	260	303	346	389
101	00 432	475	518	561	604	647	689	732	775	817
102	00 860	903	945	988	*030	*072	*115	*157	*199	*242
103	01 284	326	368	410	452	494	536	578	620	662
104	01 703	745	787	828	870	912	953	995	*036	*078
105	02 119	160	202	243	284	325	366	407	449	490
106	02 531	572	612	653	694	735	776	816	857	898
107	02 938	979	*019	*060	*100	*141	*181	*222	*262	*302
108	03 342	383	423	463	503	543	583	623	663	703
109	03 743	782	822	862	902	941	981	*021	*060	*100
110	04 139	179	218	258	297	336	376	415	454	493
111	04 532	571	610	650	689	727	766	805	844	883
112	04 922	961	999	*038	*077	*115	*154	*192	*231	*269
113	05 308	346	385	423	461	500	538	576	614	652
114	05 690	729	767	805	843	881	918	956	994	*032
115	06 070	108	145	183	221	258	296	333	371	408
116	06 446	483	521	558	595	633	670	707	744	781
117	06 819	856	893	930	967	*004	*041	*078	*115	*151
118	07 188	225	262	298	335	372	408	445	482	518
119	07 555	591	628	664	700	737	773	809	846	882
120	07 918	954	990	*027	*063	*099	*135	*171	*207	*243
121	08 279	314	350	386	422	458	493	529	565	600
122	08 636	672	707	743	778	814	849	884	920	955
123	08 991	*026	*061	*096	*132	*167	*202	*237	*272	*307
124	09 342	377	412	447	482	517	552	587	621	656
125	09 691	726	760	795	830	864	899	934	968	*003
126	10 037	072	106	140	175	209	243	278	312	346
127	10 380	415	449	483	517	551	585	619	653	687
128	10 721	755	789	823	857	890	924	958	992	*025
129	11 059	093	126	160	193	227	261	294	327	361
130	11 394	428	461	494	528	561	594	628	661	694
131	11 727	760	793	826	860	893	926	959	992	*024
132	12 057	090	123	156	189	222	254	287	320	352
133	12 385	418	450	483	516	548	581	613	646	678
134	12 710	743	775	808	840	872	905	937	969	*001
135	13 033	066	098	130	162	194	226	258	290	322
136	13 354	386	418	450	481	513	545	577	609	640
137	13 672	704	735	767	799	830	862	893	925	956
138	13 988	*019	*051	*082	*114	*145	*176	*208	*239	*270
139	14 301	333	364	395	426	457	489	520	551	582
140	14 613	644	675	706	737	768	799	829	860	891
141	14 922	953	983	*014	*045	*076	*106	*137	*168	*198
142	15 229	259	290	320	351	381	412	442	473	503
143	15 534	564	594	625	655	685	715	746	776	806
144	15 836	866	897	927	957	987	*017	*047	*077	*107
145	16 137	167	197	227	256	286	316	346	376	406
146	16 435	465	495	524	554	584	613	643	673	702
147	16 732	761	791	820	850	879	909	938	967	997
148	17 026	056	085	114	143	173	202	231	260	289
149	17 319	348	377	406	435	464	493	522	551	580
150	17 609	638	667	696	725	754	782	811	840	869
N	0	1	2	3	4	5	6	7	8	9

Prop. Parts			
	44	43	42
1	4.4	4.3	4.2
2	8.8	8.6	8.4
3	13.2	12.9	12.6
4	17.6	17.2	16.8
5	22.0	21.5	21.0
6	26.4	25.8	25.2
7	30.8	30.1	29.4
8	35.2	34.4	33.6
9	39.6	38.7	37.8
	41	40	39
1	4.1	4.0	3.9
2	8.2	8.0	7.8
3	12.3	12.0	11.7
4	16.4	16.0	15.6
5	20.5	20.0	19.5
6	24.6	24.0	23.4
7	28.7	28.0	27.3
8	32.8	32.0	31.2
9	36.9	36.0	35.1
	38	37	36
1	3.8	3.7	3.6
2	7.6	7.4	7.2
3	11.4	11.1	10.8
4	15.2	14.8	14.4
5	19.0	18.5	18.0
6	22.8	22.2	21.6
7	26.6	25.9	25.2
8	30.4	29.6	28.8
9	34.2	33.3	32.4
	35	34	33
1	3.5	3.4	3.3
2	7.0	6.8	6.6
3	10.5	10.2	9.9
4	14.0	13.6	13.2
5	17.5	17.0	16.5
6	21.0	20.4	19.8
7	24.5	23.8	23.1
8	28.0	27.2	26.4
9	31.5	30.6	29.7
	32	31	30
1	3.2	3.1	3.0
2	6.4	6.2	6.0
3	9.6	9.3	9.0
4	12.8	12.4	12.0
5	16.0	15.5	15.0
6	19.2	18.6	18.0
7	22.4	21.7	21.0
8	25.6	24.8	24.0
9	28.8	27.9	27.0

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
		150	17 609	638	667	696	725	754	782	811	840	869
		151	17 898	926	955	984	*013	*041	*070	*099	*127	*156
		152	18 184	213	241	270	298	327	355	384	412	441
29	**28**	153	18 469	498	526	554	583	611	639	667	696	724
1 2.9	2.8											
2 5.8	5.6	154	18 752	780	808	837	865	893	921	949	977	*005
3 8.7	8.4	155	19 033	061	089	117	145	173	201	229	257	285
4 11.6	11.2	156	19 312	340	368	396	424	451	479	507	535	562
5 14.5	14.0											
6 17.4	16.8	157	19 590	618	645	673	700	728	756	783	811	838
7 20.3	19.6	158	19 866	893	921	948	976	*003	*030	*058	*085	*112
8 23.2	22.4	159	20 140	167	194	222	249	276	303	330	358	385
9 26.1	25.2											
		160	20 412	439	466	493	520	548	575	602	629	656
		161	20 683	710	737	763	790	817	844	871	898	925
27	**26**	162	20 952	978	*005	*032	*059	*085	*112	*139	*165	*192
1 2.7	2.6	163	21 219	245	272	299	325	352	378	405	431	458
2 5.4	5.2											
3 8.1	7.8	164	21 484	511	537	564	590	617	643	669	696	722
4 10.8	10.4	165	21 748	775	801	827	854	880	906	932	958	985
5 13.5	13.0	166	22 011	037	063	089	115	141	167	194	220	246
6 16.2	15.6											
7 18.9	18.2	167	22 272	298	324	350	376	401	427	453	479	505
8 21.6	20.8	168	22 531	557	583	608	634	660	686	712	737	763
9 24.3	23.4	169	22 789	814	840	866	891	917	943	968	994	*019
		170	23 045	070	096	121	147	172	198	223	249	274
		171	23 300	325	350	376	401	426	452	477	502	528
25		172	23 553	578	603	629	654	679	704	729	754	779
1 2.5		173	23 805	830	855	880	905	930	955	980	*005	*030
2 5.0												
3 7.5		174	24 055	080	105	130	155	180	204	229	254	279
4 10.0		175	24 304	329	353	378	403	428	452	477	502	527
5 12.5		176	24 551	576	601	625	650	674	699	724	748	773
6 15.0												
7 17.5		177	24 797	822	846	871	895	920	944	969	993	*018
8 20.0		178	25 042	066	091	115	139	164	188	212	237	261
9 22.5		179	25 285	310	334	358	382	406	431	455	479	503
		180	25 527	551	575	600	624	648	672	696	720	744
		181	25 768	792	816	840	864	888	912	935	959	983
24	**23**	182	26 007	031	055	079	102	126	150	174	198	221
1 2.4	2.3	183	26 245	269	293	316	340	364	387	411	435	458
2 4.8	4.6											
3 7.2	6.9	184	26 482	505	529	553	576	600	623	647	670	694
4 9.6	9.2	185	26 717	741	764	788	811	834	858	881	905	928
5 12.0	11.5	186	26 951	975	998	*021	*045	*068	*091	*114	*138	*161
6 14.4	13.8											
7 16.8	16.1	187	27 184	207	231	254	277	300	323	346	370	393
8 19.2	18.4	188	27 416	439	462	485	508	531	554	577	600	623
9 21.6	20.7	189	27 646	669	692	715	738	761	784	807	830	852
		190	27 875	898	921	944	967	989	*012	*035	*058	*081
		191	28 103	126	149	171	194	217	240	262	285	307
22	**21**	192	28 330	353	375	398	421	443	466	488	511	533
1 2.2	2.1	193	28 556	578	601	623	646	668	691	713	735	758
2 4.4	4.2											
3 6.6	6.3	194	28 780	803	825	847	870	892	914	937	959	981
4 8.8	8.4	195	29 003	026	048	070	092	115	137	159	181	203
5 11.0	10.5	196	29 226	248	270	292	314	336	358	380	403	425
6 13.2	12.6											
7 15.4	14.7	197	29 447	469	491	513	535	557	579	601	623	645
8 17.6	16.8	198	29 667	688	710	732	754	776	798	820	842	863
9 19.8	18.9	199	29 885	907	929	951	973	994	*016	*038	*060	*081
		200	30 103	125	146	168	190	211	233	255	276	298
Prop. Parts		**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
200	30 103	125	146	168	190	211	233	255	276	298	
201	30 320	341	363	384	406	428	449	471	492	514	
202	30 535	557	578	600	621	643	664	685	707	728	
203	30 750	771	792	814	835	856	878	899	920	942	
204	30 963	984	*006	*027	*048	*069	*091	*112	*133	*154	
205	31 175	197	218	239	260	281	302	323	345	366	
206	31 387	408	429	450	471	492	513	534	555	576	
207	31 597	618	639	660	681	702	723	744	765	785	
208	31 806	827	848	869	890	911	931	952	973	994	
209	32 015	035	056	077	098	118	139	160	181	201	
210	32 222	243	263	284	305	325	346	366	387	408	
211	32 428	449	469	490	510	531	552	572	593	613	
212	32 634	654	675	695	715	736	756	777	797	818	
213	32 838	858	879	899	919	940	960	980	*001	*021	
214	33 041	062	082	102	122	143	163	183	203	224	
215	33 244	264	284	304	325	345	365	385	405	425	
216	33 445	465	486	506	526	546	566	586	606	626	
217	33 646	666	686	706	726	746	766	786	806	826	
218	33 846	866	885	905	925	945	965	985	*005	*025	
219	34 044	064	084	104	124	143	163	183	203	223	
220	34 242	262	282	301	321	341	361	380	400	420	
221	34 439	459	479	498	518	537	557	577	596	616	
222	34 635	655	674	694	713	733	753	772	792	811	
223	34 830	850	869	889	908	928	947	967	986	*005	
224	35 025	044	064	083	102	122	141	160	180	199	
225	35 218	238	257	276	295	315	334	353	372	392	
226	35 411	430	449	468	488	507	526	545	564	583	
227	35 603	622	641	660	679	698	717	736	755	774	
228	35 793	813	832	851	870	889	908	927	946	965	
229	35 984	*003	*021	*040	*059	*078	*097	*116	*135	*154	
230	36 173	192	211	229	248	267	286	305	324	342	
231	36 361	380	399	418	436	455	474	493	511	530	
232	36 549	568	586	605	624	642	661	680	698	717	
233	36 736	754	773	791	810	829	847	866	884	903	
234	36 922	940	959	977	996	*014	*033	*051	*070	*088	
235	37 107	125	144	162	181	199	218	236	254	273	
236	37 291	310	328	346	365	383	401	420	438	457	
237	37 475	493	511	530	548	566	585	603	621	639	
238	37 658	676	694	712	731	749	767	785	803	822	
239	37 840	858	876	894	912	931	949	967	985	*003	
240	38 021	039	057	075	093	112	130	148	166	184	
241	38 202	220	238	256	274	292	310	328	346	364	
242	38 382	399	417	435	453	471	489	507	525	543	
243	38 561	578	596	614	632	650	668	686	703	721	
244	38 739	757	775	792	810	828	846	863	881	899	
245	38 917	934	952	970	987	*005	*023	*041	*058	*076	
246	39 094	111	129	146	164	182	199	217	235	252	
247	39 270	287	305	322	340	358	375	393	410	428	
248	39 445	463	480	498	515	533	550	568	585	602	
249	39 620	637	655	672	690	707	724	742	759	777	
250	39 794	811	829	846	863	881	898	915	933	950	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

	22	21
1	2.2	2.1
2	4.4	4.2
3	6.6	6.3
4	8.8	8.4
5	11.0	10.5
6	13.2	12.6
7	15.4	14.7
8	17.6	16.8
9	19.8	18.9

	20
1	2.0
2	4.0
3	6.0
4	8.0
5	10.0
6	12.0
7	14.0
8	16.0
9	18.0

	19
1	1.9
2	3.8
3	5.7
4	7.6
5	9.5
6	11.4
7	13.3
8	15.2
9	17.1

	18
1	1.8
2	3.6
3	5.4
4	7.2
5	9.0
6	10.8
7	12.6
8	14.4
9	16.2

	17
1	1.7
2	3.4
3	5.1
4	6.8
5	8.5
6	10.2
7	11.9
8	13.6
9	15.3

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	250	39 794	811	829	846	863	881	898	915	933	950
	251	39 967	985	*002	*019	*037	*054	*071	*088	*106	*123
18	252	40 140	157	175	192	209	226	243	261	278	295
1 1.8	253	40 312	329	346	364	381	398	415	432	449	466
2 3.6	254	40 483	500	518	535	552	569	586	603	620	637
3 5.4	255	40 654	671	688	705	722	739	756	773	790	807
4 7.2	256	40 824	841	858	875	892	909	926	943	960	976
5 9.0	257	40 993	*010	*027	*044	*061	*078	*095	*111	*128	*145
6 10.8	258	41 162	179	196	212	229	246	263	280	296	313
7 12.6	259	41 330	347	363	380	397	414	430	447	464	481
8 14.4	**260**	41 497	514	531	547	564	581	597	614	631	647
9 16.2	261	41 664	681	697	714	731	747	764	780	797	814
17	262	41 830	847	863	880	896	913	929	946	963	979
1 1.7	263	41 996	*012	*029	*045	*062	*078	*095	*111	*127	*144
2 3.4	264	42 160	177	193	210	226	243	259	275	292	308
3 5.1	265	42 325	341	357	374	390	406	423	439	455	472
4 6.8	266	42 488	504	521	537	553	570	586	602	619	635
5 8.5	267	42 651	667	684	700	716	732	749	765	781	797
6 10.2	268	42 813	830	846	862	878	894	911	927	943	959
7 11.9	269	42 975	991	*008	*024	*040	*056	*072	*088	*104	*120
8 13.6	**270**	43 136	152	169	185	201	217	233	249	265	281
9 15.3	271	43 297	313	329	345	361	377	393	409	425	441
16	272	43 457	473	489	505	521	537	553	569	584	600
1 1.6	273	43 616	632	648	664	680	696	712	727	743	759
2 3.2	274	43 775	791	807	823	838	854	870	886	902	917
3 4.8	275	43 933	949	965	981	996	*012	*028	*044	*059	*075
4 6.4	276	44 091	107	122	138	154	170	185	201	217	232
5 8.0	277	44 248	264	279	295	311	326	342	358	373	389
6 9.6	278	44 404	420	436	451	467	483	498	514	529	545
7 11.2	279	44 560	576	592	607	623	638	654	669	685	700
8 12.8	**280**	44 716	731	747	762	778	793	809	824	840	855
9 14.4	281	44 871	886	902	917	932	948	963	979	994	*010
15	282	45 025	040	056	071	086	102	117	133	148	163
1 1.5	283	45 179	194	209	225	240	255	271	286	301	317
2 3.0	284	45 332	347	362	378	393	408	423	439	454	469
3 4.5	285	45 484	500	515	530	545	561	576	591	606	621
4 6.0	286	45 637	652	667	682	697	712	728	743	758	773
5 7.5	287	45 788	803	818	834	849	864	879	894	909	924
6 9.0	288	45 939	954	969	984	*000	*015	*030	*045	*060	*075
7 10.5	289	46 090	105	120	135	150	165	180	195	210	225
8 12.0	**290**	46 240	255	270	285	300	315	330	345	359	374
9 13.5	291	46 389	404	419	434	449	464	479	494	509	523
14	292	46 538	553	568	583	598	613	627	642	657	672
1 1.4	293	46 687	702	716	731	746	761	776	790	805	820
2 2.8	294	46 835	850	864	879	894	909	923	938	953	967
3 4.2	295	46 982	997	*012	*026	*041	*056	*070	*085	*100	*114
4 5.6	296	47 129	144	159	173	188	202	217	232	246	261
5 7.0	297	47 276	290	305	319	334	349	363	378	392	407
6 8.4	298	47 422	436	451	465	480	494	509	524	538	553
7 9.8	299	47 567	582	596	611	625	640	654	669	683	698
8 11.2	**300**	47 712	727	741	756	770	784	799	813	828	842
9 12.6											
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
300	47 712	727	741	756	770	784	799	813	828	842	
301	47 857	871	885	900	914	929	943	958	972	986	
302	48 001	015	029	044	058	073	087	101	116	130	
303	48 144	159	173	187	202	216	230	244	259	273	
304	48 287	302	316	330	344	359	373	387	401	416	15
305	48 430	444	458	473	487	501	515	530	544	558	1 1.5
306	48 572	586	601	615	629	643	657	671	686	700	2 3.0
307	48 714	728	742	756	770	785	799	813	827	841	3 4.5
308	48 855	869	883	897	911	926	940	954	968	982	4 6.0
309	48 996	*010	*024	*038	*052	*066	*080	*094	*108	*122	5 7.5
310	49 136	150	164	178	192	206	220	234	248	262	6 9.0
311	49 276	290	304	318	332	346	360	374	388	402	7 10.5
312	49 415	429	443	457	471	485	499	513	527	541	8 12.0
313	49 554	568	582	596	610	624	638	651	665	679	9 13.5
314	49 693	707	721	734	748	762	776	790	803	817	
315	49 831	845	859	872	886	900	914	927	941	955	
316	49 969	982	996	*010	*024	*037	*051	*065	*079	*092	14
317	50 106	120	133	147	161	174	188	202	215	229	1 1.4
318	50 243	256	270	284	297	311	325	338	352	365	2 2.8
319	50 379	393	406	420	433	447	461	474	488	501	3 4.2
320	50 515	529	542	556	569	583	596	610	623	637	4 5.6
321	50 651	664	678	691	705	718	732	745	759	772	5 7.0
322	50 786	799	813	826	840	853	866	880	893	907	6 8.4
323	50 920	934	947	961	974	987	*001	*014	*028	*041	7 9.8
324	51 055	068	081	095	108	121	135	148	162	175	8 11.2
325	51 188	202	215	228	242	255	268	282	295	308	9 12.6
326	51 322	335	348	362	375	388	402	415	428	441	
327	51 455	468	481	495	508	521	534	548	561	574	13
328	51 587	601	614	627	640	654	667	680	693	706	1 1.3
329	51 720	733	746	759	772	786	799	812	825	838	2 2.6
330	51 851	865	878	891	904	917	930	943	957	970	3 3.9
331	51 983	996	*009	*022	*035	*048	*061	*075	*088	*101	4 5.2
332	52 114	127	140	153	166	179	192	205	218	231	5 6.5
333	52 244	257	270	284	297	310	323	336	349	362	6 7.8
334	52 375	388	401	414	427	440	453	466	479	492	7 9.1
335	52 504	517	530	543	556	569	582	595	608	621	8 10.4
336	52 634	647	660	673	686	699	711	724	737	750	9 11.7
337	52 763	776	789	802	815	827	840	853	866	879	
338	52 892	905	917	930	943	956	969	982	994	*007	
339	53 020	033	046	058	071	084	097	110	122	135	12
340	53 148	161	173	186	199	212	224	237	250	263	1 1.2
341	53 275	288	301	314	326	339	352	364	377	390	2 2.4
342	53 403	415	428	441	453	466	479	491	504	517	3 3.6
343	53 529	542	555	567	580	593	605	618	631	643	4 4.8
344	53 656	668	681	694	706	719	732	744	757	769	5 6.0
345	53 782	794	807	820	832	845	857	870	882	895	6 7.2
346	53 908	920	933	945	958	970	983	995	*008	*020	7 8.4
347	54 033	045	058	070	083	095	108	120	133	145	8 9.6
348	54 158	170	183	195	208	220	233	245	258	270	9 10.8
349	54 283	295	307	320	332	345	357	370	382	394	
350	54 407	419	432	444	456	469	481	494	506	518	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	350	54 407	419	432	444	456	469	481	494	506	518
	351	54 531	543	555	568	580	593	605	617	630	642
	352	54 654	667	679	691	704	716	728	741	753	765
	353	54 777	790	802	814	827	839	851	864	876	888
13	354	54 900	913	925	937	949	962	974	986	998	*011
1 1.3	355	55 023	035	047	060	072	084	096	108	121	133
2 2.6	356	55 145	157	169	182	194	206	218	230	242	255
3 3.9											
4 5.2	357	55 267	279	291	303	315	328	340	352	364	376
5 6.5	358	55 388	400	413	425	437	449	461	473	485	497
6 7.8	359	55 509	522	534	546	558	570	582	594	606	618
7 9.1	**360**	55 630	642	654	666	678	691	703	715	727	739
8 10.4	361	55 751	763	775	787	799	811	823	835	847	859
9 11.7	362	55 871	883	895	907	919	931	943	955	967	979
	363	55 991	*003	*015	*027	*038	*050	*062	*074	*086	*098
	364	56 110	122	134	146	158	170	182	194	205	217
	365	56 229	241	253	265	277	289	301	312	324	336
12	366	56 348	360	372	384	396	407	419	431	443	455
1 1.2	367	56 467	478	490	502	514	526	538	549	561	573
2 2.4	368	56 585	597	608	620	632	644	656	667	679	691
3 3.6	369	56 703	714	726	738	750	761	773	785	797	808
4 4.8	**370**	56 820	832	844	855	867	879	891	902	914	926
5 6.0	371	56 937	949	961	972	984	996	*008	*019	*031	*043
6 7.2	372	57 054	066	078	089	101	113	124	136	148	159
7 8.4	373	57 171	183	194	206	217	229	241	252	264	276
8 9.6											
9 10.8	374	57 287	299	310	322	334	345	357	368	380	392
	375	57 403	415	426	438	449	461	473	484	496	507
	376	57 519	530	542	553	565	576	588	600	611	623
	377	57 634	646	657	669	680	692	703	715	726	738
11	378	57 749	761	772	784	795	807	818	830	841	852
1 1.1	379	57 864	875	887	898	910	921	933	944	955	967
2 2.2	**380**	57 978	990	*001	*013	*024	*035	*047	*058	*070	*081
3 3.3	381	58 092	104	115	127	138	149	161	172	184	195
4 4.4	382	58 206	218	229	240	252	263	274	286	297	309
5 5.5	383	58 320	331	343	354	365	377	388	399	410	422
6 6.6											
7 7.7	384	58 433	444	456	467	478	490	501	512	524	535
8 8.8	385	58 546	557	569	580	591	602	614	625	636	647
9 9.9	386	58 659	670	681	692	704	715	726	737	749	760
	387	58 771	782	794	805	816	827	838	850	861	872
	388	58 883	894	906	917	928	939	950	961	973	984
	389	58 995	*006	*017	*028	*040	*051	*062	*073	*084	*095
10	**390**	59 106	118	129	140	151	162	173	184	195	207
1 1.0	391	59 218	229	240	251	262	273	284	295	306	318
2 2.0	392	59 329	340	351	362	373	384	395	406	417	428
3 3.0	393	59 439	450	461	472	483	494	506	517	528	539
4 4.0											
5 5.0	394	59 550	561	572	583	594	605	616	627	638	649
6 6.0	395	59 660	671	682	693	704	715	726	737	748	759
7 7.0	396	59 770	780	791	802	813	824	835	846	857	868
8 8.0											
9 9.0	397	59 879	890	901	912	923	934	945	956	966	977
	398	59 988	999	*010	*021	*032	*043	*054	*065	*076	*086
	399	60 097	108	119	130	141	152	163	173	184	195
	400	60 206	217	228	239	249	260	271	282	293	304
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
400	60 206	217	228	239	249	260	271	282	293	304	
401	60 314	325	336	347	358	369	379	390	401	412	
402	60 423	433	444	455	466	477	487	498	509	520	
403	60 531	541	552	563	574	584	595	606	617	627	
404	60 638	649	660	670	681	692	703	713	724	735	
405	60 746	756	767	778	788	799	810	821	831	842	
406	60 853	863	874	885	895	906	917	927	938	949	
407	60 959	970	981	991	*002	*013	*023	*034	*045	*055	11
408	61 066	077	087	098	109	119	130	140	151	162	1 1.1
409	61 172	183	194	204	215	225	236	247	257	268	2 2.2
410	61 278	289	300	310	321	331	342	352	363	374	3 3.3
411	61 384	395	405	416	426	437	448	458	469	479	4 4.4
412	61 490	500	511	521	532	542	553	563	574	584	5 5.5
413	61 595	606	616	627	637	648	658	669	679	690	6 6.6
414	61 700	711	721	731	742	752	763	773	784	794	7 7.7
415	61 805	815	826	836	847	857	868	878	888	899	8 8.8
416	61 909	920	930	941	951	962	972	982	993	*003	9 9.9
417	62 014	024	034	045	055	066	076	086	097	107	
418	62 118	128	138	149	159	170	180	190	201	211	
419	62 221	232	242	252	263	273	284	294	304	315	
420	62 325	335	346	356	366	377	387	397	408	418	
421	62 428	439	449	459	469	480	490	500	511	521	
422	62 531	542	552	562	572	583	593	603	613	624	10
423	62 634	644	655	665	675	685	696	706	716	726	1 1.0
424	62 737	747	757	767	778	788	798	808	818	829	2 2.0
425	62 839	849	859	870	880	890	900	910	921	931	3 3.0
426	62 941	951	961	972	982	992	*002	*012	*022	*033	4 4.0
427	63 043	053	063	073	083	094	104	114	124	134	5 5.0
428	63 144	155	165	175	185	195	205	215	225	236	6 6.0
429	63 246	256	266	276	286	296	306	317	327	337	7 7.0
430	63 347	357	367	377	387	397	407	417	428	438	8 8.0
431	63 448	458	468	478	488	498	508	518	528	538	9 9.0
432	63 548	558	568	579	589	599	609	619	629	639	
433	63 649	659	669	679	689	699	709	719	729	739	
434	63 749	759	769	779	789	799	809	819	829	839	
435	63 849	859	869	879	889	899	909	919	929	939	
436	63 949	959	969	979	988	998	*008	*018	*028	*038	
437	64 048	058	068	078	088	098	108	118	128	137	9
438	64 147	157	167	177	187	197	207	217	227	237	1 0.9
439	64 246	256	266	276	286	296	306	316	326	335	2 1.8
440	64 345	355	365	375	385	395	404	414	424	434	3 2.7
441	64 444	454	464	473	483	493	503	513	523	532	4 3.6
442	64 542	552	562	572	582	591	601	611	621	631	5 4.5
443	64 640	650	660	670	680	689	699	709	719	729	6 5.4
444	64 738	748	758	768	777	787	797	807	816	826	7 6.3
445	64 836	846	856	865	875	885	895	904	914	924	8 7.2
446	64 933	943	953	963	972	982	992	*002	*011	*021	9 8.1
447	65 031	040	050	060	070	079	089	099	108	118	
448	65 128	137	147	157	167	176	186	196	205	215	
449	65 225	234	244	254	263	273	283	292	302	312	
450	65 321	331	341	350	360	369	379	389	398	408	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	450	65 321	331	341	350	360	369	379	389	398	408
	451	65 418	427	437	447	456	466	475	485	495	504
	452	65 514	523	533	543	552	562	571	581	591	600
	453	65 610	619	629	639	648	658	667	677	686	696
	454	65 706	715	725	734	744	753	763	772	782	792
	455	65 801	811	820	830	839	849	858	868	877	887
	456	65 896	906	916	925	935	944	954	963	973	982
10	457	65 992	*001	*011	*020	*030	*039	*049	*058	*068	*077
1 1.0	458	66 087	096	106	115	124	134	143	153	162	172
2 2.0	459	66 181	191	200	210	219	229	238	247	257	266
3 3.0	**460**	66 276	285	295	304	314	323	332	342	351	361
4 4.0	461	66 370	380	389	398	408	417	427	436	445	455
5 5.0	462	66 464	474	483	492	502	511	521	530	539	549
6 6.0	463	66 558	567	577	586	596	605	614	624	633	642
7 7.0	464	66 652	661	671	680	689	699	708	717	727	736
8 8.0	465	66 745	755	764	773	783	792	801	811	820	829
9 9.0	466	66 839	848	857	867	876	885	894	904	913	922
	467	66 932	941	950	960	969	978	987	997	*006	*015
	468	67 025	034	043	052	062	071	080	089	099	108
	469	67 117	127	136	145	154	164	173	182	191	201
	470	67 210	219	228	237	247	256	265	274	284	293
	471	67 302	311	321	330	339	348	357	367	376	385
9	472	67 394	403	413	422	431	440	449	459	468	477
1 0.9	473	67 486	495	504	514	523	532	541	550	560	569
2 1.8	474	67 578	587	596	605	614	624	633	642	651	660
3 2.7	475	67 669	679	688	697	706	715	724	733	742	752
4 3.6	476	67 761	770	779	788	797	806	815	825	834	843
5 4.5	477	67 852	861	870	879	888	897	906	916	925	934
6 5.4	478	67 943	952	961	970	979	988	997	*006	*015	*024
7 6.3	479	68 034	043	052	061	070	079	088	097	106	115
8 7.2	**480**	68 124	133	142	151	160	169	178	187	196	205
9 8.1	481	68 215	224	233	242	251	260	269	278	287	296
	482	68 305	314	323	332	341	350	359	368	377	386
	483	68 395	404	413	422	431	440	449	458	467	476
	484	68 485	494	502	511	520	529	538	547	556	565
	485	68 574	583	592	601	610	619	628	637	646	655
	486	68 664	673	681	690	699	708	717	726	735	744
8	487	68 753	762	771	780	789	797	806	815	824	833
1 0.8	488	68 842	851	860	869	878	886	895	904	913	922
2 1.6	489	68 931	940	949	958	966	975	984	993	*002	*011
3 2.4	**490**	69 020	028	037	046	055	064	073	082	090	099
4 3.2	491	69 108	117	126	135	144	152	161	170	179	188
5 4.0	492	69 197	205	214	223	232	241	249	258	267	276
6 4.8	493	69 285	294	302	311	320	329	338	346	355	364
7 5.6	494	69 373	381	390	399	408	417	425	434	443	452
8 6.4	495	69 461	469	478	487	496	504	513	522	531	539
9 7.2	496	69 548	557	566	574	583	592	601	609	618	627
	497	69 636	644	653	662	671	679	688	697	705	714
	498	69 723	732	740	749	758	767	775	784	793	801
	499	69 810	819	827	836	845	854	862	871	880	888
	500	69 897	906	914	923	932	940	949	958	966	975
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
500	69 897	906	914	923	932	940	949	958	966	975	
501	69 984	992	*001	*010	*018	*027	*036	*044	*053	*062	
502	70 070	079	088	096	105	114	122	131	140	148	
503	70 157	165	174	183	191	200	209	217	226	234	
504	70 243	252	260	269	278	286	295	303	312	321	
505	70 329	338	346	355	364	372	381	389	398	406	
506	70 415	424	432	441	449	458	467	475	484	492	
507	70 501	509	518	526	535	544	552	561	569	578	9
508	70 586	595	603	612	621	629	638	646	655	663	1 0.9
509	70 672	680	689	697	706	714	723	731	740	749	2 1.8
510	70 757	766	774	783	791	800	808	817	825	834	3 2.7
511	70 842	851	859	868	876	885	893	902	910	919	4 3.6
512	70 927	935	944	952	961	969	978	986	995	*003	5 4.5
513	71 012	020	029	037	046	054	063	071	079	088	6 5.4
514	71 096	105	113	122	130	139	147	155	164	172	7 6.3
515	71 181	189	198	206	214	223	231	240	248	257	8 7.2
516	71 265	273	282	290	299	307	315	324	332	341	9 8.1
517	71 349	357	366	374	383	391	399	408	416	425	
518	71 433	441	450	458	466	475	483	492	500	508	
519	71 517	525	533	542	550	559	567	575	584	592	
520	71 600	609	617	625	634	642	650	659	667	675	
521	71 684	692	700	709	717	725	734	742	750	759	
522	71 767	775	784	792	800	809	817	825	834	842	8
523	71 850	858	867	875	883	892	900	908	917	925	1 0.8
524	71 933	941	950	958	966	975	983	991	999	*008	2 1.6
525	72 016	024	032	041	049	057	066	074	082	090	3 2.4
526	72 099	107	115	123	132	140	148	156	165	173	4 3.2
527	72 181	189	198	206	214	222	230	239	247	255	5 4.0
528	72 263	272	280	288	296	304	313	321	329	337	6 4.8
529	72 346	354	362	370	378	387	395	403	411	419	7 5.6
530	72 428	436	444	452	460	469	477	485	493	501	8 6.4
531	72 509	518	526	534	542	550	558	567	575	583	9 7.2
532	72 591	599	607	616	624	632	640	648	656	665	
533	72 673	681	689	697	705	713	722	730	738	746	
534	72 754	762	770	779	787	795	803	811	819	827	
535	72 835	843	852	860	868	876	884	892	900	908	
536	72 916	925	933	941	949	957	965	973	981	989	
537	72 997	*006	*014	*022	*030	*038	*046	*054	*062	*070	7
538	73 078	086	094	102	111	119	127	135	143	151	1 0.7
539	73 159	167	175	183	191	199	207	215	223	231	2 1.4
540	73 239	247	255	263	272	280	288	296	304	312	3 2.1
541	73 320	328	336	344	352	360	368	376	384	392	4 2.8
542	73 400	408	416	424	432	440	448	456	464	472	5 3.5
543	73 480	488	496	504	512	520	528	536	544	552	6 4.2
544	73 560	568	576	584	592	600	608	616	624	632	7 4.9
545	73 640	648	656	664	672	679	687	695	703	711	8 5.6
546	73 719	727	735	743	751	759	767	775	783	791	9 6.3
547	73 799	807	815	823	830	838	846	854	862	870	
548	73 878	886	894	902	910	918	926	933	941	949	
549	73 957	965	973	981	989	997	*005	*013	*020	*028	
550	74 036	044	052	060	068	076	084	092	099	107	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	550	74 036	044	052	060	068	076	084	092	099	107
	551	74 115	123	131	139	147	155	162	170	178	186
	552	74 194	202	210	218	225	233	241	249	257	265
	553	74 273	280	288	296	304	312	320	327	335	343
	554	74 351	359	367	374	382	390	398	406	414	421
	555	74 429	437	445	453	461	468	476	484	492	500
	556	74 507	515	523	531	539	547	554	562	570	578
	557	74 586	593	601	609	617	624	632	640	648	656
	558	74 663	671	679	687	695	702	710	718	726	733
	559	74 741	749	757	764	772	780	788	796	803	811
	560	74 819	827	834	842	850	858	865	873	881	889
	561	74 896	904	912	920	927	935	943	950	958	966
8	562	74 974	981	989	997	*005	*012	*020	*028	*035	*043
1 0.8	563	75 051	059	066	074	082	089	097	105	113	120
2 1.6	564	75 128	136	143	151	159	166	174	182	189	197
3 2.4	565	75 205	213	220	228	236	243	251	259	266	274
4 3.2	566	75 282	289	297	305	312	320	328	335	343	351
5 4.0	567	75 358	366	374	381	389	397	404	412	420	427
6 4.8	568	75 435	442	450	458	465	473	481	488	496	504
7 5.6	569	75 511	519	526	534	542	549	557	565	572	580
8 6.4	**570**	75 587	595	603	610	618	626	633	641	648	656
9 7.2	571	75 664	671	679	686	694	702	709	717	724	732
	572	75 740	747	755	762	770	778	785	793	800	808
	573	75 815	823	831	838	846	853	861	868	876	884
	574	75 891	899	906	914	921	929	937	944	952	959
	575	75 967	974	982	989	997	*005	*012	*020	*027	*035
	576	76 042	050	057	065	072	080	087	095	103	110
	577	76 118	125	133	140	148	155	163	170	178	185
	578	76 193	200	208	215	223	230	238	245	253	260
	579	76 268	275	283	290	298	305	313	320	328	335
	580	76 343	350	358	365	373	380	388	395	403	410
	581	76 418	425	433	440	448	455	462	470	477	485
7	582	76 492	500	507	515	522	530	537	545	552	559
1 0.7	583	76 567	574	582	589	597	604	612	619	626	634
2 1.4	584	76 641	649	656	664	671	678	686	693	701	708
3 2.1	585	76 716	723	730	738	745	753	760	768	775	782
4 2.8	586	76 790	797	805	812	819	827	834	842	849	856
5 3.5	587	76 864	871	879	886	893	901	908	916	923	930
6 4.2	588	76 938	945	953	960	967	975	982	989	997	*004
7 4.9	589	77 012	019	026	034	041	048	056	063	070	078
8 5.6	**590**	77 085	093	100	107	115	122	129	137	144	151
9 6.3	591	77 159	166	173	181	188	195	203	210	217	225
	592	77 232	240	247	254	262	269	276	283	291	298
	593	77 305	313	320	327	335	342	349	357	364	371
	594	77 379	386	393	401	408	415	422	430	437	444
	595	77 452	459	466	474	481	488	495	503	510	517
	596	77 525	532	539	546	554	561	568	576	583	590
	597	77 597	605	612	619	627	634	641	648	656	663
	598	77 670	677	685	692	699	706	714	721	728	735
	599	77 743	750	757	764	772	779	786	793	801	808
	600	77 815	822	830	837	844	851	859	866	873	880
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**3**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
600	77 815	822	830	837	844	851	859	866	873	880	
601	77 887	895	902	909	916	924	931	938	945	952	
602	77 960	967	974	981	988	996	*003	*010	*017	*025	
603	78 032	039	046	053	061	068	075	082	089	097	
604	78 104	111	118	125	132	140	147	154	161	168	
605	78 176	183	190	197	204	211	219	226	233	240	
606	78 247	254	262	269	276	283	290	297	305	312	
607	78 319	326	333	340	347	355	362	369	376	383	8
608	78 390	398	405	412	419	426	433	440	447	455	1 0.8
609	78 462	469	476	483	490	497	504	512	519	526	2 1.6
610	78 533	540	547	554	561	569	576	583	590	597	3 2.4
611	78 604	611	618	625	633	640	647	654	661	668	4 3.2
612	78 675	682	689	696	704	711	718	725	732	739	5 4.0
613	78 746	753	760	767	774	781	789	796	803	810	6 4.8
614	78 817	824	831	838	845	852	859	866	873	880	7 5.6
615	78 888	895	902	909	916	923	930	937	944	951	8 6.4
616	78 958	965	972	979	986	993	*000	*007	*014	*021	9 7.2
617	79 029	036	043	050	057	064	071	078	085	092	
618	79 099	106	113	120	127	134	141	148	155	162	
619	79 169	176	183	190	197	204	211	218	225	232	
620	79 239	246	253	260	267	274	281	288	295	302	
621	79 309	316	323	330	337	344	351	358	365	372	7
622	79 379	386	393	400	407	414	421	428	435	442	1 0.7
623	79 449	456	463	470	477	484	491	498	505	511	2 1.4
624	79 518	525	532	539	546	553	560	567	574	581	3 2.1
625	79 588	595	602	609	616	623	630	637	644	650	4 2.8
626	79 657	664	671	678	685	692	699	706	713	720	5 3.5
627	79 727	734	741	748	754	761	768	775	782	789	6 4.2
628	79 796	803	810	817	824	831	837	844	851	858	7 4.9
629	79 865	872	879	886	893	900	906	913	920	927	8 5.6
630	79 934	941	948	955	962	969	975	982	989	996	9 6.3
631	80 003	010	017	024	030	037	044	051	058	065	
632	80 072	079	085	092	099	106	113	120	127	134	
633	80 140	147	154	161	168	175	182	188	195	202	
634	80 209	216	223	229	236	243	250	257	264	271	
635	80 277	284	291	298	305	312	318	325	332	339	
636	80 346	353	359	366	373	380	387	393	400	407	
637	80 414	421	428	434	441	448	455	462	468	475	6
638	80 482	489	496	502	509	516	523	530	536	543	1 0.6
639	80 550	557	564	570	577	584	591	598	604	611	2 1.2
640	80 618	625	632	638	645	652	659	665	672	679	3 1.8
641	80 686	693	699	706	713	720	726	733	740	747	4 2.4
642	80 754	760	767	774	781	787	794	801	808	814	5 3.0
643	80 821	828	835	841	848	855	862	868	875	882	6 3.6
644	80 889	895	902	909	916	922	929	936	943	949	7 4.2
645	80 956	963	969	976	983	990	996	*003	*010	*017	8 4.8
646	81 023	030	037	043	050	057	064	070	077	084	9 5.4
647	81 090	097	104	111	117	124	131	137	144	151	
648	81 158	164	171	178	184	191	198	204	211	218	
649	81 224	231	238	245	251	258	265	271	278	285	
650	81 291	298	305	311	318	325	331	338	345	351	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	650	81 291	298	305	311	318	325	331	338	345	351
	651	81 358	365	371	378	385	391	398	405	411	418
	652	81 425	431	438	445	451	458	465	471	478	485
	653	81 491	498	505	511	518	525	531	538	544	551
	654	81 558	564	571	578	584	591	598	604	611	617
	655	81 624	631	637	644	651	657	664	671	677	684
	656	81 690	697	704	710	717	723	730	737	743	750
	657	81 757	763	770	776	783	790	796	803	809	816
	658	81 823	829	836	842	849	856	862	869	875	882
	659	81 889	895	902	908	915	921	928	935	941	948
	660	81 954	961	968	974	981	987	994	*000	*007	*014
	661	82 020	027	033	040	046	053	060	066	073	079
7	662	82 086	092	099	105	112	119	125	132	138	145
1 0.7	663	82 151	158	164	171	178	184	191	197	204	210
2 1.4	664	82 217	223	230	236	243	249	256	263	269	276
3 2.1	665	82 282	289	295	302	308	315	321	328	334	341
4 2.8	666	82 347	354	360	367	373	380	387	393	400	406
5 3.5	667	82 413	419	426	432	439	445	452	458	465	471
6 4.2	668	82 478	484	491	497	504	510	517	523	530	536
7 4.9	669	82 543	549	556	562	569	575	582	588	595	601
8 5.6	**670**	82 607	614	620	627	633	640	646	653	659	666
9 6.3	671	82 672	679	685	692	698	705	711	718	724	730
	672	82 737	743	750	756	763	769	776	782	789	795
	673	82 802	808	814	821	827	834	840	847	853	860
	674	82 866	872	879	885	892	898	905	911	918	924
	675	82 930	937	943	950	956	963	969	975	982	988
	676	82 995	*001	*008	*014	*020	*027	*033	*040	*046	*052
	677	83 059	065	072	078	085	091	097	104	110	117
	678	83 123	129	136	142	149	155	161	168	174	181
	679	83 187	193	200	206	213	219	225	232	238	245
	680	83 251	257	264	270	276	283	289	296	302	308
	681	83 315	321	327	334	340	347	353	359	366	372
6	682	83 378	385	391	398	404	410	417	423	429	436
1 0.6	683	83 442	448	455	461	467	474	480	487	493	499
2 1.2	684	83 506	512	518	525	531	537	544	550	556	563
3 1.8	685	83 569	575	582	588	594	601	607	613	620	626
4 2.4	686	83 632	639	645	651	658	664	670	677	683	689
5 3.0	687	83 696	702	708	715	721	727	734	740	746	753
6 3.6	688	83 759	765	771	778	784	790	797	803	809	816
7 4.2	689	83 822	828	835	841	847	853	860	866	872	879
8 4.8	**690**	83 885	891	897	904	910	916	923	929	935	942
9 5.4	691	83 948	954	960	967	973	979	985	992	998	*004
	692	84 011	017	023	029	036	042	048	055	061	067
	693	84 073	080	086	092	098	105	111	117	123	130
	694	84 136	142	148	155	161	167	173	180	186	192
	695	84 198	205	211	217	223	230	236	242	248	255
	696	84 261	267	273	280	286	292	298	305	311	317
	697	84 323	330	336	342	348	354	361	367	373	379
	698	84 386	392	398	404	410	417	423	429	435	442
	699	84 448	454	460	466	473	479	485	491	497	504
	700	84 510	516	522	528	535	541	547	553	559	566
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
700	84 510	516	522	528	535	541	547	553	559	566	
701	84 572	578	584	590	597	603	609	615	621	628	
702	84 634	640	646	652	658	665	671	677	683	689	
703	84 696	702	708	714	720	726	733	739	745	751	
704	84 757	763	770	776	782	788	794	800	807	813	
705	84 819	825	831	837	844	850	856	862	868	874	
706	84 880	887	893	899	905	911	917	924	930	936	
707	84 942	948	954	960	967	973	979	985	991	997	7
708	85 003	009	016	022	028	034	040	046	052	058	1 0.7
709	85 065	071	077	083	089	095	101	107	114	120	2 1.4
710	85 126	132	138	144	150	156	163	169	175	181	3 2.1
711	85 187	193	199	205	211	217	224	230	236	242	4 2.8
712	85 248	254	260	266	272	278	285	291	297	303	5 3.5
713	85 309	315	321	327	333	339	345	352	358	364	6 4.2
714	85 370	376	382	388	394	400	406	412	418	425	7 4.9
715	85 431	437	443	449	455	461	467	473	479	485	8 5.6
716	85 491	497	503	509	516	522	528	534	540	546	9 6.3
717	85 552	558	564	570	576	582	588	594	600	606	
718	85 612	618	625	631	637	643	649	655	661	667	
719	85 673	679	685	691	697	703	709	715	721	727	
720	85 733	739	745	751	757	763	769	775	781	788	
721	85 794	800	806	812	818	824	830	836	842	848	
722	85 854	860	866	872	878	884	890	896	902	908	6
723	85 914	920	926	932	938	944	950	956	962	968	1 0.6
724	85 974	980	986	992	998	*004	*010	*016	*022	*028	2 1.2
725	86 034	040	046	052	058	064	070	076	082	088	3 1.8
726	86 094	100	106	112	118	124	130	136	141	147	4 2.4
727	86 153	159	165	171	177	183	189	195	201	207	5 3.0
728	86 213	219	225	231	237	243	249	255	261	267	6 3.6
729	86 273	279	285	291	297	303	308	314	320	326	7 4.2
730	86 332	338	344	350	356	362	368	374	380	386	8 4.8
731	86 392	398	404	410	415	421	427	433	439	445	9 5.4
732	86 451	457	463	469	475	481	487	493	499	504	
733	86 510	516	522	528	534	540	546	552	558	564	
734	86 570	576	581	587	593	599	605	611	617	623	
735	86 629	635	641	646	652	658	664	670	676	682	
736	86 688	694	700	705	711	717	723	729	735	741	
737	86 747	753	759	764	770	776	782	788	794	800	5
738	86 806	812	817	823	829	835	841	847	853	859	1 0.5
739	86 864	870	876	882	888	894	900	906	911	917	2 1.0
740	86 923	929	935	941	947	953	958	964	970	976	3 1.5
741	86 982	988	994	999	*005	*011	*017	*023	*029	*035	4 2.0
742	87 040	046	052	058	064	070	075	081	087	093	5 2.5
743	87 099	105	111	116	122	128	134	140	146	151	6 3.0
744	87 157	163	169	175	181	186	192	198	204	210	7 3.5
745	87 216	221	227	233	239	245	251	256	262	268	8 4.0
746	87 274	280	286	291	297	303	309	315	320	326	9 4.5
747	87 332	338	344	349	355	361	367	373	379	384	
748	87 390	396	402	408	413	419	425	431	437	442	
749	87 448	454	460	466	471	477	483	489	495	500	
750	87 506	512	518	523	529	535	541	547	552	558	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	750	87 506	512	518	523	529	535	541	547	552	558
	751	87 564	570	576	581	587	593	599	604	610	616
	752	87 622	628	633	639	645	651	656	662	668	674
	753	87 679	685	691	697	703	708	714	720	726	731
	754	87 737	743	749	754	760	766	772	777	783	789
	755	87 795	800	806	812	818	823	829	835	841	846
	756	87 852	858	864	869	875	881	887	892	898	904
	757	87 910	915	921	927	933	938	944	950	955	961
	758	87 967	973	978	984	990	996	*001	*007	*013	*018
	759	88 024	030	036	041	047	053	058	064	070	076
	760	88 081	087	093	098	104	110	116	121	127	133
	761	88 138	144	150	156	161	167	173	178	184	190
6	762	88 195	201	207	213	218	224	230	235	241	247
1 0.6	763	88 252	258	264	270	275	281	287	292	298	304
2 1.2	764	88 309	315	321	326	332	338	343	349	355	360
3 1.8	765	88 366	372	377	383	389	395	400	406	412	417
4 2.4	766	88 423	429	434	440	446	451	457	463	468	474
5 3.0	767	88 480	485	491	497	502	508	513	519	525	530
6 3.6	768	88 536	542	547	553	559	564	570	576	581	587
7 4.2	769	88 593	598	604	610	615	621	627	632	638	643
8 4.8	**770**	88 649	655	660	666	672	677	683	689	694	700
9 5.4	771	88 705	711	717	722	728	734	739	745	750	756
	772	88 762	767	773	779	784	790	795	801	807	812
	773	88 818	824	829	835	840	846	852	857	863	868
	774	88 874	880	885	891	897	902	908	913	919	925
	775	88 930	936	941	947	953	958	964	969	975	981
	776	88 986	992	997	*003	*009	*014	*020	*025	*031	*037
	777	89 042	048	053	059	064	070	076	081	087	092
	778	89 098	104	109	115	120	126	131	137	143	148
	779	89 154	159	165	170	176	182	187	193	198	204
	780	89 209	215	221	226	232	237	243	248	254	260
	781	89 265	271	276	282	287	293	298	304	310	315
5	782	89 321	326	332	337	343	348	354	360	365	371
1 0.5	783	89 376	382	387	393	398	404	409	415	421	426
2 1.0	784	89 432	437	443	448	454	459	465	470	476	481
3 1.5	785	89 487	492	498	504	509	515	520	526	531	537
4 2.0	786	89 542	548	553	559	564	570	575	581	586	592
5 2.5	787	89 597	603	609	614	620	625	631	636	642	647
6 3.0	788	89 653	658	664	669	675	680	686	691	697	702
7 3.5	789	89 708	713	719	724	730	735	741	746	752	757
8 4.0	**790**	89 763	768	774	779	785	790	796	801	807	812
9 4.5	791	89 818	823	829	834	840	845	851	856	862	867
	792	89 873	878	883	889	894	900	905	911	916	922
	793	89 927	933	938	944	949	955	960	966	971	977
	794	89 982	988	993	998	*004	*009	*015	*020	*026	*031
	795	90 037	042	048	053	059	064	069	075	080	086
	796	90 091	097	102	108	113	119	124	129	135	140
	797	90 146	151	157	162	168	173	179	184	189	195
	798	90 200	206	211	217	222	227	233	238	244	249
	799	90 255	260	266	271	276	282	287	293	298	304
	800	90 309	314	320	325	331	336	342	347	352	358
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
800	90 309	314	320	325	331	336	342	347	352	358	
801	90 363	369	374	380	385	390	396	401	407	412	
802	90 417	423	428	434	439	445	450	455	461	466	
803	90 472	477	482	488	493	499	504	509	515	520	
804	90 526	531	536	542	547	553	558	563	569	574	
805	90 580	585	590	596	601	607	612	617	623	628	
806	90 634	639	644	650	655	660	666	671	677	682	
807	90 687	693	698	703	709	714	720	725	730	736	
808	90 741	747	752	757	763	768	773	779	784	789	
809	90 795	800	806	811	816	822	827	832	838	843	
810	90 849	854	859	865	870	875	881	886	891	897	
811	90 902	907	913	918	924	929	934	940	945	950	
812	90 956	961	966	972	977	982	988	993	998	*004	
813	91 009	014	020	025	030	036	041	046	052	057	6
814	91 062	068	073	078	084	089	094	100	105	110	1 0.6
815	91 116	121	126	132	137	142	148	153	158	164	2 1.2
816	91 169	174	180	185	190	196	201	206	212	217	3 1.8
817	91 222	228	233	238	243	249	254	259	265	270	4 2.4
818	91 275	281	286	291	297	302	307	312	318	323	5 3.0
819	91 328	334	339	344	350	355	360	365	371	376	6 3.6
820	91 381	387	392	397	403	408	413	418	424	429	7 4.2
821	91 434	440	445	450	455	461	466	471	477	482	8 4.8
822	91 487	492	498	503	508	514	519	524	529	535	9 5.4
823	91 540	545	551	556	561	566	572	577	582	587	
824	91 593	598	603	609	614	619	624	630	635	640	
825	91 645	651	656	661	666	672	677	682	687	693	
826	91 698	703	709	714	719	724	730	735	740	745	
827	91 751	756	761	766	772	777	782	787	793	798	
828	91 803	808	814	819	824	829	834	840	845	850	
829	91 855	861	866	871	876	882	887	892	897	903	
830	91 908	913	918	924	929	934	939	944	950	955	
831	91 960	965	971	976	981	986	991	997	*002	*007	
832	92 012	018	023	028	033	038	044	049	054	059	5
833	92 065	070	075	080	085	091	096	101	106	111	1 0.5
834	92 117	122	127	132	137	143	148	153	158	163	2 1.0
835	92 169	174	179	184	189	195	200	205	210	215	3 1.5
836	92 221	226	231	236	241	247	252	257	262	267	4 2.0
837	92 273	278	283	288	293	298	304	309	314	319	5 2.5
838	92 324	330	335	340	345	350	355	361	366	371	6 3.0
839	92 376	381	387	392	397	402	407	412	418	423	7 3.5
840	92 428	433	438	443	449	454	459	464	469	474	8 4.0
841	92 480	485	490	495	500	505	511	516	521	526	9 4.5
842	92 531	536	542	547	552	557	562	567	572	578	
843	92 583	588	593	598	603	609	614	619	624	629	
844	92 634	639	645	650	655	660	665	670	675	681	
845	92 686	691	696	701	706	711	716	722	727	732	
846	92 737	742	747	752	758	763	768	773	778	783	
847	92 788	793	799	804	809	814	819	824	829	834	
848	92 840	845	850	855	860	865	870	875	881	886	
849	92 891	896	901	906	911	916	921	927	932	937	
850	92 942	947	952	957	962	967	973	978	983	988	
N	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Prop. Parts**

Prop Parts	N	0	1	2	3	4	5	6	7	8	9
	850	92 942	947	952	957	962	967	973	978	983	988
	851	92 993	998	*003	*008	*013	*018	*024	*029	*034	*039
	852	93 044	049	054	059	064	069	075	080	085	090
	853	93 095	100	105	110	115	120	125	131	136	141
	854	93 146	151	156	161	166	171	176	181	186	192
	855	93 197	202	207	212	217	222	227	232	237	242
	856	93 247	252	258	263	268	273	278	283	288	293
6	857	93 298	303	308	313	318	323	328	334	339	344
1 0.6	858	93 349	354	359	364	369	374	379	384	389	394
2 1.2	859	93 399	404	409	414	420	425	430	435	440	445
3 1.8	**860**	93 450	455	460	465	470	475	480	485	490	495
4 2.4	861	93 500	505	510	515	520	526	531	536	541	546
5 3.0	862	93 551	556	561	566	571	576	581	586	591	596
6 3.6	863	93 601	606	611	616	621	626	631	636	641	646
7 4.2	864	93 651	656	661	666	671	676	682	687	692	697
8 4.8	865	93 702	707	712	717	722	727	732	737	742	747
9 5.4	866	93 752	757	762	767	772	777	782	787	792	797
	867	93 802	807	812	817	822	827	832	837	842	847
	868	93 852	857	862	867	872	877	882	887	892	897
	869	93 902	907	912	917	922	927	932	937	942	947
	870	93 952	957	962	967	972	977	982	987	992	997
	871	94 002	007	012	017	022	027	032	037	042	047
5	872	94 052	057	062	067	072	077	082	086	091	096
1 0.5	873	94 101	106	111	116	121	126	131	136	141	146
2 1.0	874	94 151	156	161	166	171	176	181	186	191	196
3 1.5	875	94 201	206	211	216	221	226	231	236	240	245
4 2.0	876	94 250	255	260	265	270	275	280	285	290	295
5 2.5	877	94 300	305	310	315	320	325	330	335	340	345
6 3.0	878	94 349	354	359	364	369	374	379	384	389	394
7 3.5	879	94 399	404	409	414	419	424	429	433	438	443
8 4.0	**880**	94 448	453	458	463	468	473	478	483	488	493
9 4.5	881	94 498	503	507	512	517	522	527	532	537	542
	882	94 547	552	557	562	567	571	576	581	586	591
	883	94 596	601	606	611	616	621	626	630	635	640
	884	94 645	650	655	660	665	670	675	680	685	689
	885	94 694	699	704	709	714	719	724	729	734	738
	886	94 743	748	753	758	763	768	773	778	783	787
4	887	94 792	797	802	807	812	817	822	827	832	836
1 0.4	888	94 841	846	851	856	861	866	871	876	880	885
2 0.8	889	94 890	895	900	905	910	915	919	924	929	934
3 1.2	**890**	94 939	944	949	954	959	963	968	973	978	983
4 1.6	891	94 988	993	998	*002	*007	*012	*017	*022	*027	*032
5 2.0	892	95 036	041	046	051	056	061	066	071	075	080
6 2.4	893	95 085	090	095	100	105	109	114	119	124	129
7 2.8	894	95 134	139	143	148	153	158	163	168	173	177
8 3.2	895	95 182	187	192	197	202	207	211	216	221	226
9 3.6	896	95 231	236	240	245	250	255	260	265	270	274
	897	95 279	284	289	294	299	303	308	313	318	323
	898	95 328	332	337	342	347	352	357	361	366	371
	899	95 376	381	386	390	395	400	405	410	415	419
	900	95 424	429	434	439	444	448	453	458	463	468
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
900	95 424	429	434	439	444	448	453	458	463	468	
901	95 472	477	482	487	492	497	501	506	511	516	
902	95 521	525	530	535	540	545	550	554	559	564	
903	95 569	574	578	583	588	593	598	602	607	612	
904	95 617	622	626	631	636	641	646	650	655	660	
905	95 665	670	674	679	684	689	694	698	703	708	
906	95 713	718	722	727	732	737	742	746	751	756	
907	95 761	766	770	775	780	785	789	794	799	804	
908	95 809	813	818	823	828	832	837	842	847	852	
909	95 856	861	866	871	875	880	885	890	895	899	
910	95 904	909	914	018	923	928	933	938	942	947	
911	95 952	957	961	966	971	976	980	985	990	995	
912	95 999	*004	*009	*014	*019	*023	*028	*033	*038	*042	
913	96 047	052	057	061	066	071	076	080	085	090	
914	96 095	099	104	109	114	118	123	128	133	137	
915	96 142	147	152	156	161	166	171	175	180	185	
916	96 190	194	199	204	209	213	218	223	227	232	
917	96 237	242	246	251	256	261	265	270	275	280	
918	96 284	289	294	298	303	308	313	317	322	327	
919	96 332	336	341	346	350	355	360	365	369	374	
920	96 379	384	388	393	398	402	407	412	417	421	
921	96 426	431	435	440	445	450	454	459	464	468	
922	96 473	478	483	487	492	497	501	506	511	515	
923	96 520	525	530	534	539	544	548	553	558	562	
924	96 567	572	577	581	586	591	595	600	605	609	
925	96 614	619	624	628	633	638	642	647	652	656	
926	96 661	666	670	675	680	685	689	694	699	703	
927	96 708	713	717	722	727	731	736	741	745	750	
928	96 755	759	764	769	774	778	783	788	792	797	
929	96 802	806	811	816	820	825	830	834	839	844	
930	96 848	853	858	862	867	872	876	881	886	890	
931	96 895	900	904	909	914	918	923	928	932	937	
932	96 942	946	951	956	960	965	970	974	979	984	
933	96 988	993	997	*002	*007	*011	*016	*021	*025	*030	
934	97 035	039	044	049	053	058	063	067	072	077	
935	97 081	086	090	095	100	104	109	114	118	123	
936	97 128	132	137	142	146	151	155	160	165	169	
937	97 174	179	183	188	192	197	202	206	211	216	
938	97 220	225	230	234	239	243	248	253	257	262	
939	97 267	271	276	280	285	290	294	299	304	308	
940	97 313	317	322	327	331	336	340	345	350	354	
941	97 359	364	368	373	377	382	387	391	396	400	
942	97 405	410	414	419	424	428	433	437	442	447	
943	97 451	456	460	465	470	474	479	483	488	493	
944	97 497	502	506	511	516	520	525	529	534	539	
945	97 543	548	552	557	562	566	571	575	580	585	
946	97 589	594	598	603	607	612	617	621	626	630	
947	97 635	640	644	649	653	658	663	667	672	676	
948	97 681	685	690	695	699	704	708	713	717	722	
949	97 727	731	736	740	745	749	754	759	763	768	
950	97 772	777	782	786	791	795	800	804	809	813	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

	5
1	0.5
2	1.0
3	1.5
4	2.0
5	2.5
6	3.0
7	3.5
8	4.0
9	4.5

	4
1	0.4
2	0.8
3	1.2
4	1.6
5	2.0
6	2.4
7	2.8
8	3.2
9	3.6

Prop. Parts	N	0	1	2	3	4	5	6	7	8	9
	950	97 772	777	782	786	791	795	800	804	809	813
	951	97 818	823	827	832	836	841	845	850	855	859
	952	97 864	868	873	877	882	886	891	896	900	905
	953	97 909	914	918	923	928	932	937	941	946	950
	954	97 955	959	964	968	973	978	982	987	991	996
	955	98 000	005	009	014	019	023	028	032	037	041
	956	98 046	050	055	059	064	068	073	078	082	087
	957	98 091	096	100	105	109	114	118	123	127	132
	958	98 137	141	146	150	155	159	164	168	173	177
	959	98 182	186	191	195	200	204	209	214	218	223
	960	98 227	232	236	241	245	250	254	259	263	268
	961	98 272	277	281	286	290	295	299	304	308	313
5	962	98 318	322	327	331	336	340	345	349	354	358
1 0.5	963	98 363	367	372	376	381	385	390	394	399	403
2 1.0	964	98 408	412	417	421	426	430	435	439	444	448
3 1.5	965	98 453	457	462	466	471	475	480	484	489	493
4 2.0	966	98 498	502	507	511	516	520	525	529	534	538
5 2.5	967	98 543	547	552	556	561	565	570	574	579	583
6 3.0	968	98 588	592	597	601	605	610	614	619	623	628
7 3.5	969	98 632	637	641	646	650	655	659	664	668	673
8 4.0	**970**	98 677	682	686	691	695	700	704	709	713	717
9 4.5	971	98 722	726	731	735	740	744	749	753	758	762
	972	98 767	771	776	780	784	789	793	798	802	807
	973	98 811	816	820	825	829	834	838	843	847	851
	974	98 856	860	865	869	874	878	883	887	892	896
	975	98 900	905	909	914	918	923	927	932	936	941
	976	98 945	949	954	958	963	967	972	976	981	985
	977	98 989	994	998	*003	*007	*012	*016	*021	*025	*029
	978	99 034	038	043	047	052	056	061	065	069	074
	979	99 078	083	087	092	096	100	105	109	114	118
	980	99 123	127	131	136	140	145	149	154	158	162
	981	99 167	171	176	180	185	189	193	198	202	207
4	982	99 211	216	220	224	229	233	238	242	247	251
1 0.4	983	99 255	260	264	269	273	277	282	286	291	295
2 0.8	984	99 300	304	308	313	317	322	326	330	335	339
3 1.2	985	99 344	348	352	357	361	366	370	374	379	383
4 1.6	986	99 388	392	396	401	405	410	414	419	423	427
5 2.0	987	99 432	436	441	445	449	454	458	463	467	471
6 2.4	988	99 476	480	484	489	493	498	502	506	511	515
7 2.8	989	99 520	524	528	533	537	542	546	550	555	559
8 3.2	**990**	99 564	568	572	577	581	585	590	594	599	603
9 3.6	991	99 607	612	616	621	625	629	634	638	642	647
	992	99 651	656	660	664	669	673	677	682	686	691
	993	99 695	699	704	708	712	717	721	726	730	734
	994	99 739	743	747	752	756	760	765	769	774	778
	995	99 782	787	791	795	800	804	808	813	817	822
	996	99 826	830	835	839	843	848	852	856	861	865
	997	99 870	874	878	883	887	891	896	900	904	909
	998	99 913	917	922	926	930	935	939	944	948	952
	999	99 957	961	965	970	974	978	983	987	991	996
	1000	00 000	004	009	013	017	022	026	030	035	039
Prop. Parts	**N**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Index

INDEX

Abscissa, 191
Accounting ratios, 83
Accuracy of estimate, 194, 388, 391
Additive relationships, 473
Aggregative index numbers, 496, 499
Agricultural ratios, 85
Almac, R. B., 491
Amplitude of business cycles, 473, 478; seasonal variation, 449, 453
Analysis of covariance, 586
Analysis of large samples, 93, 98
Analysis of small samples, 522; how obtained, 524; purpose, 524; experimental design, 525; Latin square, 527; *t*-test, 537; arithmetic mean, 537; regression, 542, 544; correlation, 544; percentages, 545; Chi-square, 547; computation of error, 529
Analysis of variance, 557; experimental error, 574; practical use of, 576
Anti-logarithm, 128, 691
Arbitrary origin, 113, 160
Areas of the normal curve of error, Appendix Table II
Arithmetic mean, 113; unorganized data, 115; class intervals, 117; long method, 117; short method, 118; average of group means, 124; weighted, 123, 124; percentages, 72, 81; characteristics of, 125; use of, 125; distribution of, 531; of small samples, 530, 531; standard error of, 532
Arithmetic scale, 17
Array, 97
Average deviation, 153, 154; methods of computation, 154, 156; use, 155 (*See* Mean deviation)
Averages, 113, 136; arithmetic mean, 113; geometric mean, 125, 127; harmonic mean, 131, 135; median, 136, 137; mode, 142, 143; simple averages, 114, 115; weighted averages, 123, 124; root-square-mean, 163
Axes, 191; *X*-axis, 191; *Y*-axis, 191

Babson, Roger W., 417
Band chart, 31, 32
Bar chart, 25, 27; horizontal, 25; vertical, 27, 28; compound, 26, 28; methods of making, 29, 30
Base period, 492
Bennett, K. R., 312
Beta coefficients, 302
Bias, 366; sampling, 368; index numbers, 493, 495, 499
Binomial equation, 352; triangle, 354; distribution, 354; coefficients, 354; application to sampling, 355; error, 379, 387; normal curves, 357
Burnes, Arthur F., 480
Business cycles, 412; measurement, 470, 473; amplitude, 478; deviation from normal, 473, 483; forecasting, 473, 480
Business publications, 514, 520

Captions, 50
Central tendency, 113, 145; measure-

ment of, 113, 125, 136, 144 (*See* Averages)
Charts, 11; line, 12; Lorenz, 24; *Z*-charts, 23; percentage, 16, 18, 22; logarithmic, 16, 18; band, 31, 32; bar, 25, 28; pie, 33, 34; plus and minus, 35, 36; maps, 37, 38, 40; pictographs, 42, 43; cumulative, 66, 69, 141; rules for making, 12, 29, 34, 40
Checking accuracy of data, 612
Chi-square, 549; computation, 547; test of goodness of fit, 551; normality, 553
Chi-square Table, 550
Classification, 50, 54; bases of, 50; time, 50; space, 50; quality, 50; quantity, 50
Class intervals, 99, 100; rules for forming, 99; limits, 103; number, 101; size, 100; equal, 93, 99; unequal, 324; concentration of data, 103; mid-points, 101, 102; by *f* formula, 102; by graphic method, 102
Coding data, 202, 640
Coefficient of correlation 215; determination, 216; relation to regression, 219; Pearsonian coefficient, 224; of ungrouped data, 228; for totals, 228; correlation tables, 230, 231; rank differences, 236; graphic methods, 235; computation of Rho, 296; of *R*, 296; partial correlation, 285, 286; *Z*-functions, Appendix Table III
Coefficient of determination, 215; relation to correlation, 215; multiple, 295, 296; percentage, 296; separate, 298
Coefficient of variation, 172, 173
Coin tossing, 350
Collecting data, 620
Commercial and Financial Chronicle, 417
Compound interest, law of, 130, 131, 269
Continuous series, 94
Controls in sampling, large, 366, 372; small, 525, 528
Corn-hog ratio, 84
Correction, 201, 248, 291
Correlation, 285; use, 286; simple, 223; multiple, 295, 296; positive, 218; negative, 218; linear, 220; tabular, 234; non-linear, 288; limitations, 216, 217
Correlation table, 229, 231
Covariance, 586
Critical ratio, 570
Cumulative charts, 141
Curve fitting, 241, 251; freehand, 241; class averages, 242, 273; least squares, 251; mathematical, 251; linear, 201; simple parabola, 251; cubic parabola, 259; exponential function, 262; hyperbolic, 265, 267; compound interest, 269; Gompertz curve, 269; logistic, 273; time series, 201, 204
Curvilinear regression, 240; freehand, 242; simple, 243; parabolas, 247, 248; mathematical, 248; exponential function, 262; hyperbolic curve, 265; compound interest, 269; Gompertz, 269; logistic, 274; time series, 278; relation to correlation, 244, 260; to determination, 244, 260; to multiple correlation, 312
Cyclical fluctuations, 463; residual method, 465; direct method, 474, 480; measurement, 475, 477; forecasting, 483; amplitude, 478; use in budgeting, 463; multiplicity, 412

Data, 93; primary, 612, 620, 632; secondary, 612, 620; continuous, 94; discrete, 94
Data tables, 318
Decile, 140
Degrees of freedom, 535; *t*-test, 537, 543; computation of error, 536; in small samples, 536; for Chi-square, 550; analysis of variance, 560, 562; analysis of covariance, 595, 598; multiple correlation, 291, 295
Dependent variable, 178; symbols, 191, 192; meaning, 179, 192

Determination, 215, 224; relation to correlation, 215; multiple, 296, 312; separate, 297; standard error of estimate, 288, 295
Deviation, 150; range, 152; quartile, 153; average, 153; mean, 153; from ungrouped data, 154, 156; computation of, 155, 160; standard deviation, 155, 158; from class intervals, 158, 160; root-mean-square, 163
Differences, first, 271, 274; paired, 271, 274
Direct method of computing cycles, 480
Discrete series, 94
Dispersion, 150, 151, 160; measures of, 152, 153, 155; range, 152; quartile deviation, 153; mean deviation, 154; standard deviation, 156, 160; comparison of measures, 161; coefficient of variation, 172, 173
Distribution, 150; frequency, 93, 99, 105; normal curve of error, 165
Doolittle solution of equations, 256; parabolas, 257; multiple correlation, 292
Dun and Bradstreet, Inc., 417

Economics, publications, 513, 515, 520
Editing statistical data, 622, 639
Educational ratios, 86
Equations, 685, 687; binomial, 352; regression, 247; simple non-linear, 247, 250; error, 363; time series, 431, 432, 436; logarithmic, 263; exponential, 267; compound interest, 269; logistic, 272; Gompertz, 269, 271
Error, 346, 379, 381, 388, 395, 529; sampling, 379; normal curve, 358; standard, 363; of mean, 388; standard deviation, 391; coefficient of correlation, 391
Experimental design, 525
Exponential curve, 262, 264
Exponents, Appendix III
Extrapolation, 439
Ezekiel, Mordecai, 294, 311

F-ratio, 578
F-tables, 579, 581
Federal publications, 416, 420
Fisher, R. A., 582, 583, 604
Freehand curves, 180, 182, 183
Frequency curves, 107; histograms, 30; frequency polygon, 30, 105; normal, 358; skewed, 164, 165; equations for, 353, 357, 359
Frequency distributions, 93, 99, 105; location of class intervals, 100, 116; number of classes, 101; graphs of, 104, 105; types of, 105; size of classes, 99, 101; *J*-shaped, 107; characteristics of, 105, 107, 116; normal, 353, 359; skewed, 164, 165
Frequency polygon, 30, 105, 106
Frequency tables, 32, 54, 99, 103, 116; one-way tables, 54; two-way tables, 55; three-way tables, 56; four-way tables, 59; cumulative, 66; percentage, 63

Gauss, Karl F., 357
Geometric mean, 125, 128; definition of, 126; characteristics of, 126, 129, 144; computation of, 127, 128; from class intervals, 128; use of, 129, 130; limitations, 144
Glossary of symbols, Appendix I
Gompertz curve, 269, 271
Goodness of fit, 388; standard error of estimate, 194, 244; Chi-square test, 551
Graphic methods, 141, 422; locating averages, 141; median, 141; quartiles, 141; quintiles, 141; percentages, 141
Graphic presentation, 11, 422; frequency distributions, 104, 105; line graphs, 12, 14, 20; natural or absolute scale, 15, 17, 422; *Z*-chart, 23; pie chart, 33, 34; ratio or logarithmic charts, 16, 18, 423; limitations of, 21

Harmonic mean, 131, 135; characteristics of, 132; computation of, 133, 135; from ungrouped data, 133;

from class intervals, 136; use of reciprocals, 133, 135
Histogram, 30, 104
Hog-corn ratio, 84
Hyperbolic curve, 265, 267

Independent variable, 178, 191; symbols, 191, 287; multiple, 287
Index numbers, 487; nature of, 488; use of, 489, 513; selecting base, 492; type of averages, 492; weights, 499; unweighted, 499; weighted, 499; quantity weights, 499; value weights, 493, 502; tests for, 495; aggregate, 496, 499; average of price relatives, 500; shifting bases, 506; production relatives, 489; quantity indexes, 509; socio-economic indexes, 508; seasonal, 443, 447; wholesale prices, 517; cost of of living, 518
Index of variation, 172, 173
Induction, statistical, 208
Interpolation, mode, 143; trend values, 423

J-curve, 107
Jerome, Harry, 362
Joint relationships, 468; tabular analysis, 329; analysis of variance, 576

Kurtosis, 170, 173; nature of, 173; measurement of, 170, 172

Latin square, 527, 602
Law of large numbers, 367
Least squares, 189, 192, 430; theory, 190, 194; formulas, 192, 247
Leptokurtic distributions, 171
Likelihood (*see* Probability)
Line charts, 12, 17, 18, 22
Linear correlation (*see* Correlation)
Link relatives, 445
Lively, C. E., 491
Logarithmic regression lines, 262, 265, 422
Logarithms, 688; characteristics, 689; common, 690; use of Appendix III; Appendix Table V
Logistic curves, 273, 277
Log lines, 12, 18, 22
Lorenz curve, 24

Mantissa, Appendix III; Appendix Table V
Maps, 29, 37, 38, 39, 40
Mean, 112; arithmetic, 113; definition of, 113; from ungrouped data, 115; from class intervals, 117, 118; long method, 117; short method, 118; weighted, 123, 124; characteristics of, 125; algebraic treatment of, 118; standard error, 388; reliability of, 388
Mean deviation, 153, 156
Mean, geometric, 125; definition of, 126; computation of, 127, 128; use of, 129, 130; from individual items, 127; weighted, 127, 128; from class intervals, 128
Mean, harmonic, 131; definition, 132; computation of, 133; from individual items, 133, 135; weighted, 133; use of, 132, 133; from class intervals, 136
Median, 136; definition, 136; computation of, 137; graphic method, 141; from formulas, 137; relation to arithmetic mean, 145; characteristics of, 136; use of, 145
Methods of research (*see* Averages, Dispersion, Regression, Correlation, Tabulation)
Mid-point of classes, 101, 102
Mills, Frederick C., 298
Mises, Richard von, 348
Mitchell, W. C., 480
Mode, 142; definition, 142; ungrouped data, 142; class intervals, 143; interpolation formula, 143; locational formula, 143; characteristics of, 145; use of, 144
Moments around an arithmetic mean, 167; first moment, 167; second moment, 168; third moment, 168, 169; fourth moment, 169, 170; uses of, 167, 172
Monthly trend values, 432

Moody's, 417, 520
Moving averages, 424; methods of computation, 426, 428; definition, 424; uses of, 425; as trends, 424, 428; in cycles, 425, 428; in seasonal variation, 450, 456; characteristics of, 449, 451
Multiple correlation, 286, 289; meaning of, 286; coefficient of, 295; computation of, 291, 295; relation to net regression, 286; equations, 288; straight line, 286; non-linear, 310; freehand, 311; limitations of, 312; uses of, 312

Naperian logarithms, 357
National Bureau of Economic Research, 474
Natural numbers, Tables of Squares, Square roots and Reciprocals, Appendix Table IV
Net regression coefficients, 288, 292
Non-government indexes, 520
Non-linear correlation, 240, 260
Non-linear regression, 240, 260
Normal curve of error, 107, 165, 357, 358; binomial expansion, 352; equation, 354
Normal equations, 190, 240, 253; derivation of, 192, 247, 253; use of, 353
Normal law of error, 165
Normal probability curve, 165, 358, 693, 694
Normal values, 165, 358, 693, 694
Null-hypothesis, 533; *t*-test tables, 534; for small samples, 533; relation to research, 536

Ordinate, 191
Organization of data, 47, 55, 113; array, 97; tally sheet, 99, 103; class intervals, 100, 101; bases of, 47; tabulation, 49; tables, 54, 55, 56, 59, 68
Organization of statistical problems, 608
Organization tables, 334, 336, 339
Origin, 118, 191, 431; assumed mean, 118; time-series trends, 431, 436

Parabola, 247, 278; simple, 247; equations, 247; uses of, 253; cubic, 253, 255
Parameter, 346, 382
Partial correlation, 285, 297; definition, 297; equations of, 298, 299
Pearson, F. A., 312
Percentage comparisons, tables, 62
Percentages, 72
Percentile, 139, 140, 141
Pictographs, 42, 43
Pie charts, 34
Pin-head maps, 29
Platykurtic, 171
Plus and minus charts, 35, 36
Population ratios, 87
Price relatives, 501
Primary data, 612, 620, 632
Probability, 347, 348, 354; principle of, 348, 350; definition of, 348; additions of, 350; joint, 350; measurements of, 352; normal, 357
Probability curve, 348, 357, 363, 395, 693; equations, 357; use of, 362; physical data, 362; biological data, 362; social data, 364
Probable error, 363
Projected trend, 439 (*see* Extrapolation)
Purposive selection in sampling, 370, 371

Quadratic mean, 163
Qualitative, 47
Quantitative, 47
Quartile, 138, 139; graphic presentation, 141; deviation, 153, 161; range, 153; use of, 138
Questionnaire, Chapter 23
Quintile, 139

Railroad ratios, 88
Random sampling, 372, 373
Randomized blocks, 536
Range of variation, 150, 152; semi-quartile, 153; mean deviation, 153, 156; standard deviation, 155, 160

Rank correlation, 236
Rate of interest formula, 130
Ratio charts, 12, 17, 18, 22
Ratios, 72; bases of, 73, 74, 77; like items, 77; unlike units, 78; averaging, 81; popular ratios, 83, 85, 88
Reciprocals, Appendix Table IV
Regression, 176; definition, 177, 178; plotting, 180; lines, 186, 190; linear, 179, 183, 185; non-linear, 242, 247; computation of, 196, 198; simple, 201, 204; freehand, 179, 181, 183, 184, 186; through class means, 179, 241; least squares, 190, 192, 430; time series, 430; standard error of estimate, 194, 199, 244, 252; use, 204; multiple, 286; tabular analysis, 323
Relative prices, 501, 503
Reliability, 379, 388, 395, 529; means, 388; standard deviation, 391, coefficient of correlation, 391
Residuals, 195, 204
Root-mean-square average, 163

Samples, 366, 395; methods of selection, 368; size of, 367; dependability, 379; random, 369, 372; stratified, 370, 371; purposive, 371; reliability of, 379, 381, 388
Sampling, 366, 367, 395; theory, 367; random, 368, 372; large, 366; small, 522; problems of, 368; purposive, 371; stratified, 369, 370; errors, 379, 386, 387
Scales, 12, 17, 19
Scatter, 150; degree of, 150, 151; measures of, 152, 153, 160
Scatter diagram, 180, 204
Schedules, 633; testing, 637; editing, 639
Science, 178
Seasonal variation, 443, 460; methods of computation, 445; amplitude, 444, 449; use of moving average, 449, 456; adjustments of, 448, 453; use of, 443; uniformity, 444; changing, 460
Secondary data, 612, 620
Secular movement in cycles, 479
Secular trend, 402, 435; measurement of, 431, 439; freehand, 179, 186; least squares, 431, 435, 439; moving averages, 424, 428; monthly trends, 432, 434; parabolas, 431
Selecting of averages, 145, 146
Semi-inter-quartile range, 153
Semi-logarithmic charts, 12, 18, 22
Separate determination, 279
Series of data, 93; time, 399, 410; discrete, 94; continuous, 94; spatial, 47, 48
Shifting monthly values, 433, 434
Significance, 390; t-test, 536
Skewness, 164, 165, 168; definition, 164, 173; measures, 164, 166, 173
Slope of regression line, 247
Small samples, 522; definition, 523; purpose, 524; how obtained, 525; experimental design, 528; Latin square, 529; error, 532
Snedecor, George W., 582, 583
Socio-economic indexes, 508
Sources of data, 115, 513
Spatial data, 48
Square root, 683; Appendix Table IV
Standard and Poors, 418
Standard deviation, 155; definition, 156, 157; computation of, 156, 158, 160; unorganized data, 156; class intervals, 158, 160; long method, 158; short method, 160; characteristics of, 157, 161, 162; use in correlation, 219; error, 388; variance, 557
Standard error, 194, 195, 199; binomial expansion, 352; of mean, 388; standard deviation, 391; correlation, 391; Z-functions, 695
Standard error of estimate, 194, 244; significance, 194; computation of, 195, 200, 207; short methods, 244; zone of estimate, 204; parabola, 252; multiple net regression, 295
Statistic, 8
Statistical inference, 208
Statistical methods, 6

Statistical problems, organization of, Chapter 23
Statistical reports, Chapter 25; scientific, 645; business, 648; popular, 652
Statistical induction, 208
Statistical theory, 6
Statistical unit, 609, 610
Stratified sample, 370
Survey of Current Business, indexes, 515
Symbols, glossary of, Appendix I

Tables, 47, 54, 318; cross classification, complete, 54, 56, 59; incomplete, 60, 61, 62; percentage, 66; *t*-table, 534; Chi-square, 550; *F*-tables, 579, 581; ordinates of normal curve, Appendix Table I; areas under normal curve, Appendix Table II; *Z*-table, Appendix Table III; powers of natural numbers, Appendix Table IV; square roots, Appendix Table IV; reciprocals, Appendix Table IV; logarithms, Appendix Table V
Tabular analysis, 47, 318; data table, 318; organization tables, 320, 324, 327; summary tables, 329, 330, 332; number of class intervals, 333; uniform classes, 99, 101; unequal classes, 324
Tabulation of data, 97, 317; array, 97; tally sheet, 103, 116; class intervals, 100, 101, 117; bases of classifications, 47; types of tables, 49; stubs, 50; captions, 50; positions of emphasis, 51, 52
Tally sheets, 103, 116
Time-series analysis, 399, 410; complexity, 399; removal of trend, 431; removal of seasonal variation, 445; removal of cycles, 467, 469; irregular changes, 399; direct method, 473; residual method, 465
Trends, 176, 185, 191, 325, 430 (*see* Regression)

Ungrouped data, 96
United States Census data, 416
United States Government indexes of data, 622, 624
Unweighted averages, 113
Use of libraries, 621

Validity of samples, 150, 631
Variable, 178, 179; independent, 178; dependent, 179
Variability, measures of, 151, 153, 160
Variance, analysis of, 557, 561
Variation, coefficient of, 172, 173

Weighted averages, 123, 124, 493, 498, 503
Working hypothesis, 304, 305
Worksheets, samples of, 96, 97, 103, 115, 118

X-axis, 191

Y-axis, 191

Z-test, of adjustment of correlation values, 393, 695; Appendix Table III