

UNIVERSAL
LIBRARY

OU_150754

UNIVERSAL
LIBRARY

OUP-67-11-1-68-5,000.

OSMANIA UNIVERSITY LIBRARY

Call No. **311**
D611 Accession No. **2187**
Author **Dittmer, Clarence G.**
Title **Introduction to social statistics**
1926

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

INTRODUCTION TO SOCIAL STATISTICS

BY
CLARENCE G. DITTMER, PH.D.

ASSISTANT PROFESSOR OF SOCIOLOGY,
UNIVERSITY OF WISCONSIN



CHICAGO & NEW YORK
A. W. SHAW COMPANY
LONDON, A. W. SHAW AND COMPANY, LIMITED
1926

COPYRIGHT 1926, BY A. W. SHAW COMPANY
PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

THE writer of this *Introduction to Social Statistics* has kept continually in mind the public for which he was writing it. He is interested neither in startling the statistician nor in delighting the mathematician. He has been teaching the elements of statistical method to sociology students at the University of Wisconsin and realizes that few of them will ever become professional statisticians but that all of them will have to deal with statistical studies and present their findings in an intelligible manner. He has in mind also the army of volunteer and professional social workers who today are expected to deal with facts rather than with fancies. This is not a public of mathematicians; indeed, most of the users of this book will have long since forgotten their algebra and most of them would be completely stumped by a problem involving the extraction of the square root. This being the case, the writer has endeavored to make as simple as possible the use of only the most essential mathematical formulas and has made no attempt to explain their derivation.

Students usually approach a course in statistical method with a decided "complex." It is often required of them, they recognize it as a necessary evil, they feel that it will be hard and that a subject "as dry as statistics" cannot possibly be interesting. The task of the teacher is to dissolve this "complex" and make the subject live. In the first place, an elementary or introductory course in statistics should be both *elementary* and *introductory*. There is no object in swamping the student with heavy theory and method which the average student is incapable of understanding and which, in all probability, he will never use. In the second place, the teaching of statistical method should be carefully *motivated*. Hours spent in the statistical laboratory are

worth far more than the best of lectures on the subject. The writer has met with the best response from his students by using his one weekly lecture period for the purpose of discussing the laboratory work and relating it to actual and practical social problems. In the laboratory one statistical project is carried through the entire semester. One semester, for example, schedules were placed in the hands of the students, calling for information regarding the age, age at marriage, nationality, occupation, and cause of death (if dead) of both parents and grandparents; also the number of children in the family and their ages. Such schedules, filled in by the students and their friends, were assembled and carefully edited. The entire semester was spent in working up the data. Toward the end of the semester, on the basis of the semester's experience, the students criticized and revised the original schedule and drew up complete plans for a statistical investigation in some other field.

Another semester the class worked in much the same manner with birth certificate records supplied by the State Board of Health. Each student worked with a different county but with records covering exactly the same period of time. This method better motivates the work of the laboratory than isolated problems. Statistics are not dry unless made so.

There may be some question as to the propriety of naming this book *Introduction to Social Statistics*. An *introduction* it is, but it is a *social* statistics only in so far as it aims to meet the needs of those who must understand and use statistical method in elementary sociological research.

The writer is indebted to Professor Richard T. Ely, who made this book possible, to Professors Edward A. Ross and John L. Gillin for inspiration and guidance, and to Professors Harry Jerome, Jacob Perlman, and Roy J. Colbert for critical reading of the manuscript.

CLARENCE G. DITTMER

Madison, Wisconsin, May 4, 1926

CONTENTS

PREFACE	iii
---------------	-----

I

STATISTICS AND THE SOCIAL SCIENCES	3
--	---

Abstract speculation *versus* the scientific study of facts. Need in the social sciences. Statistics as an old science. The need for a new application. The causes of the present importance of statistics. Statistics as "social bookkeeping." The purpose of statistics. The scientific method in the exact and the social sciences. The limitations of statistics. Statistics and intellectual integrity.

II

THE FIELD OF SOCIO-STATISTICAL FACT	11
---	----

The people who should be interested in social statistics. Social statistics studies society as much in health as in disease. Previous analyses of the field of socio-statistical fact. A new analysis. The whole field of social relationships for all social purposes. Analyses of related fields: business statistics; vital statistics; medical statistics; educational statistics; mathematical statistics.

III

STATISTICAL ACCURACY	22
----------------------------	----

Statistics as the exact scientific treatment of approximations. Mechanical and tabular aids to accuracy. The use of round numbers. Continuous and discontinuous series. Sources of error: fictitious accuracy; cumulative and compensating errors; the mislocation of decimal points; fictitious accuracy in totals; personal bias. The standard of accuracy. Editing for accuracy. Accuracy and the statistician's reputation.

IV

GATHERING THE DATA	30
--------------------------	----

Sampling *versus* accounting. The law of statistical regularity. The law of the inertia of large numbers. Secondary sources and their use. Primary data and how to gather them. Types of primary investigation. A practical example of the schedule maker's art. Some samples from the hands of amateurs.

V

TABULATION	53
------------------	----

"The shorthand of statistics." The work-sheet. The tally sheet. The perfect title. Good mechanical form. Where to put totals.

Tabular perspective. Simple, partial, and complete classification. Chronological, geographical, qualitative, and quantitative classification. The expressing of class limits. Cumulative frequency. Mechanical tabulation.

VI

GRAPHIC PRESENTATION 77

The purpose of graphic presentation. The pictogram or statistical picture. The limits of pictogrammatic presentation. The cartogram or statistical map. The circle and sector or "pie" chart. The bar chart. The line graph. Criteria and standards of good graphic presentation.

VII

NUMERICAL SUMMARIZATION 112

The meaning of numerical summarization. Types of numerical summarization. Averages: the arithmetic mean; the mode; the median. The characteristics and limitations of averages. The weighted average.

VIII

NUMERICAL SUMMARIZATION (*Continued*) ... 125

Measures of dispersion: the range; average deviation; quartile deviation; standard deviation. Coefficients of deviation. The measure of skewness. Coefficients of correlation. Rates. Index numbers. Probable error.

IX

GRAPHIC AND TABULAR SUMMARIZATION .. 146

The moving average. The straight-line trend (line of least squares). Short-time oscillation. The correlation or double-entry table; the scatter diagram. The Lorenz curve. The ratio or logarithmic chart.

INDEX 163

ILLUSTRATIONS

CHARTS

1. Ranging and frequency distribution	65
2. Illustrated vertical bar chart	79
3. Pictogramatic table	81
4. Chart and table combined—relativity	82
5. Two statistical maps or cartograms	83
6. Graduated dot chart	85
7. Circle and sector or pie chart	86
8. Shaded circle and sector chart	87
9. Table and bar chart combined	88
10. Horizontal subdivided bar chart	89
11. Horizontal bar chart in perspective	91
12. Subdivided percentage horizontal bar chart	92
13. Vertical bar chart with zero line	93
14. Subdivided percentage vertical bar chart	94
15. Vertical bar chart	95
16. Vertical bar chart	96
17. Graph and vertical bar chart combined	97
18 <i>a</i> . Bar Chart, frequency polygon, and smoothed curve combined	100
18 <i>b</i> . Frequency polygon or straight-line graph	101
18 <i>c</i> . Smoothed curve	102
18 <i>d</i> . Bar chart or rectangular histogram, straight-line graph or frequency polygon and smoothed curve combined	103
19. Chronological line graph—shaded	105
20. Chronological line graph—shaded	106

21.	Chronological line graph—simple	107
22.	Pictogram and cartogram combined	109
23.	Cumulative frequency graph	122
24.	A comparison of the dispersions of frequency distributions	128
25.	Chart showing the tendency of the three averages to separate under conditions of skewness	132
26.	The moving average	147
27.	The line of least squares	150
28.	Short-time oscillation	152
29.	The scatter diagram	155
30.	The Lorenz curve	157
31.	The ratio chart	159
32.	The arithmetic chart	160

SCHEDULES

1.	Industrial department educational survey	46
2.	Out-of-state tourist survey	47
3.	Confidential information—wages and employment ..	48
4.	Fresh air camp survey	48
5.	Jewish investigation committee's schedule letter	51
6.	Child welfare association's letter	52

FORMS

1.	Survey of tuberculosis cases schedule	39
2.	Punch cards, adapted and unadapted	72
3.	Punch card adapted to special problem	72

TABLES

1.	The Work-Sheet	54
2.	The Tally Sheet	56
3.	Cumulative Frequency Distribution	57
4.	Simple Single Frequency Table	58
5.	Simple Single Frequency Table	59
6.	Double Complete Cross-Classification Table	60
7.	Triple Complete Cross-Classification Table	61
8.	Triple Complete Cross-Classification Table	62
9.	Quadruple Complete Cross-Classification Table	63
10.	Conjugal Condition; Double Complete Cross-Classification Table	73
11.	Special-Purpose Table	74
12.	Special-Purpose Table	75
13.	Special-Purpose Table	76
14.	Who Gets the Money? (Illustrating Chart 3)	81
15.	Distribution of North China Families by Size and Income	104
16.	Sample Distribution for the Estimation of the Arithmetic Mean	114
17.	Process for the Estimation of the Arithmetic Mean from a Frequency Table	115
18.	Sample Distribution of Unassembled Items	116
19.	Sample Distribution for the Estimation of the Mode	117
20.	Sample Distribution for the Estimation of the Median	120
21.	Sample Distribution for the Estimation of Average Amount of Deviation	126

22.	Sample Distribution for the Computation of the Standard Deviation	131
23.	Sample Paired Series for the Computation of the Pearsonian Coefficient of Correlation	135
24.	Table Showing the Computation of the Pearsonian Coefficient of Correlation	135
25.	Determination of the Moving Average	148
26.	Determination of the Line of Least Squares from a Sample Distribution	151
27.	Determination of the Short-Time Oscillation or Trend from a Sample Distribution	153
28.	The Correlation or Double-Entry Table	154
29.	Sample Distribution for the Estimation of the Lorenz Curve	156
30.	Sample Distribution for the Illustration of the Ratio Chart	158

**INTRODUCTION TO
SOCIAL STATISTICS**

I

STATISTICS AND THE SOCIAL SCIENCES

Abstract speculation *versus* the scientific study of facts. Need in the social sciences. Statistics as an old science. The need for a new application. The causes of the present importance of statistics. Statistics as "social book-keeping." The purpose of statistics. The scientific method in the exact and the social sciences. The limitations of statistics. Statistics and intellectual integrity.

ANY "cracker-barrel" philosopher of the village grocery store variety can tell you just what is wrong with the world, and exactly what to do about it. I am afraid, also, that he has colleagues in far more spectacular environments, who are using his identical methods of deduction. He has lived, he has observed life about him, he has read. On the basis of these experiences and his environmental stimuli, he weighs and evaluates as best he can and then just comes to conclusions. His scientific method may be faulty, but his convictions are none the less strong.

Since the beginning of time, men have speculated about qualitative fact in an attempt to straighten out their natural and spiritual environment. It is only natural that "the study of the vital, fundamental, inescapable phenomena of human life has gripped man from the very earliest time." Through this method, which might be called the philosophical method, men have made great discoveries, but most of their conclusions would have to be classed as opinions. Even what we know today as the "exact sciences" had their beginning in this manner and became scientific and exact only as they applied methods of exact measurement to the field of quantitative fact.

Today we refer to that field of endeavor which has to do with the study of human relationships as the Social Sciences,

and it is a question in the minds of many, whether in their present state of development, they justify being so designated. Some even contend that human relationships are of such variable and highly complicated order that it will never be possible to reduce their study to the status of an exact science. Be this as it may, modern society is demanding less of the "cracker-barrel" philosophy, and more of the exact conclusions which come from the unbiased, unemotional study of measurable, quantitative data. We are now no longer content to "plant by the moon," and incantation forms no part of the chemist's laboratory equipment. The physicist tells us less of what he thinks, and more of what he has demonstrated to be the case. The present prestige enjoyed by economics is due to the fact that it has largely given over the method of abstract speculation and insists on making a scientific analysis of facts. If the other social sciences are to justify their name, or even to hold their own, it will be because of the continued and further application of statistical method in their research projects.

Statistics is not a new science. The word itself is nearly two centuries old, having first been used by Aschenwall. Previous to that time the term Political Arithmetic was current, and describes far more accurately its earlier function. The word comes from the Latin word "status," meaning the state, and the Italian word "statistica," meaning one versed in matters of the state, which is an indication of the fact that statistics were used mainly in earlier times as aids to rulers in the art of state-craft. They were even used by primitive tribes as a sort of historical reminder in the form of notched sticks, knots tied in cords, piles of stones, and various other methods. Modern statistical method probably had its beginning in the periodical census, which, however, did not come until near the beginning of the eighteenth century, when Frederick William the First began gathering semiannual reports on the economic condition of his people.

In earlier times censuses were occasional and extend into remote antiquity. Even in the Bible we have our book of "Numbers" and we read in I Chronicles, Chapter 21, in interesting detail how David numbered his people for the purpose of determining their fighting strength. "And Satan stood up against Israel and provoked David to number Israel." Accordingly, having been prompted of the devil, David commissioned Joab to number his people, but Joab demurred, expressing his desire that "The Lord make his people a hundred times more than they be." "Nevertheless, the king's word prevailed" and Joab accepted his commission. The results of his investigation were of primary interest. He found the total population to number 1,000,000, of whom one out of ten were soldiers. "But Levi and Benjamin counted he not among them: for the king's word was abominable to Joab." Interestingly enough "God was displeased with this thing: therefore he smote Israel" and for three days the plague raged and of the 1,000,000 of the population, not counting the tribes of Levi and Benjamin, the specific mortality rate from this plague was 70. Thus did God punish the children of Israel for David's sin in making this earliest recorded census.

We may read into this old story an antipathy to censuses probably of long standing. They were quite properly taken as "works of the devil" by harassed populations who possessed no other explanation and who saw in them merely new aids to further burdens. The census meant two things: that the king had in mind a new military draft and that new taxes were in the making. May we conclude that Joab, in opposing the census, was a better statesman than the king?

During the middle of the seventeenth century, John of Gaunt made important discoveries from the study of mortality statistics. This led Edmund Halley to lay the foundations for modern scientific life insurance. Before this time life insurance had been known, but it was little more than gambling on the lives of sea captains.

A history of statistics as it has been used through the ages would be interesting, but our purpose here is to show how it has come into a new use. We no longer use it merely for purposes of historical reminder, or as an aid to statecraft, important as these phases are, but for the general welfare of the social population; for the purpose of understanding and bettering its condition. The modern census is very different from the old. Not only is it periodically collected, and does it gather information with regard to economic conditions, and fighting strength, but also with regard to social conditions, and general welfare as well. Lester F. Ward presents to us the gospel of Telic Progress, but it will be only through exact statistical knowledge that society will be able to promote its own progress—to lay out its own path—not through mere philosophizing.

There are three causes of the present importance of statistics. In the first place, *this is an age of large numbers*. Social relationships are so extended and knowledge has so increased that we are faced with the necessity for scientifically classified knowledge. It is now impossible for every man to know everything. It is impossible for him to know thoroughly more than a small part, even in his own highly specialized field. Today nations are composed of many millions of people, whereas in the old days, they were composed of thousands. We are compelled to think and to talk in terms of such large numbers that many times it is impossible for us even to grasp their significance. An individual is usually limited to what he can himself see, or experience, but modern means of communication have greatly extended his seeing and experiencing faculties. But this very extension of knowledge increases the difficulty in forming judgments and coming to conclusions. He is balked by the fact, and is compelled to turn to statistical analysis or classified knowledge even to understand what they are all about. In the second place, *scientific method demands it*. Society is no longer satisfied with haphazard guesses, and

demands that those who are trying to regulate our social affairs be as well versed and know as well what they are about as those who are working in the field of the natural sciences. In the third place, *just the business of living is becoming more and more a scientific process* and social programs, matters of scientific procedure. It is easy enough to either hold or to refute opinions, and it is altogether too easy for both to be based upon prejudice and superstition. People in general are still prone merely to react, because "thinking is such hard work." Statistical facts, whether founded or unfounded, seem to be unanswerable. On the one hand, we accuse statistics of lying, but on the other, we clinch our arguments with them where we can.

Statistics have been called "social bookkeeping." Society needs balance sheets, cost accounting, and sales reports as much as does the merchant. The general storekeeper at the crossroads can get along very well without complicated systems of bookkeeping; it is possible for him to keep many of his accounts in his head, but the great merchant or the head of a large corporation is completely lost without them. So it is with a small community. In the earlier city states, which were composed of only a few thousands of individuals, it was possible for the rulers to know at first hand, most of the conditions of their people. But in the modern great state, composed of hundreds of millions of people, government cannot proceed by guesswork. The facts must be at hand. Chapin has well said that "Controversies over social problems usually give rise to more heat than light"; a deep truth which indicates our need for an exact statistical measurement of social phenomena.

Jerome has said that statistics "is the art of extracting significant truths concealed in masses of numerical facts." Long as statistics have been known and used, it is an interesting fact that the modern science of statistics is scarcely a half-century old. We have at the present time great masses of statistical data but with so little standardization

of method as to make them almost useless for comparative purposes. "During the past decade, however, methods of dealing with statistical data have been so greatly improved that it is becoming possible to interpret social, economic, and vital relations, and even to forecast future trends." They are used to predict economic crises, to study unemployment, to ascertain wage rates, to determine the reasons for and the conditions of poverty and dependency, as a basis for granting pardon and parole, to aid social workers in dispensing relief, and to check up on most of our philosophies and theories of general social welfare. But statistics must not be thought of as merely an aid in the study of socio-pathological conditions. Statistics are needed as much or more for the study and understanding of the normal functioning of the social group in a state of health.

Pages of uncorrelated figures are bewildering and capable of wide misuse and misinterpretation. It is the function of statistical method to standardize the technique of presentation and interpretation, and to safeguard against their misuse. The science of statistics, however, is not an end in itself; it is merely an essential means to a supremely important end. Its aim is to designate relative importance, to evaluate, to distinguish between important facts, to place first things first, and to give perception.

Social statistics deals with the measurement of social forces. It is the exact scientific method applied to the study of human relationships and is fast becoming the universal language of the social sciences. As a science its method is far more difficult than the scientific method as applied to the field of the natural sciences. There the scientist deals with forces which are beyond the control of man, such as physical forces and chemical reactions. Other things being equal, the law of gravity always operates in exactly the same manner. Chemicals of a given purity always react in combination as they have done before. But social science, in dealing with social phenomena, is dealing with a group of

forces which are man-controlled. It is not always the case that they act in the same manner; the reaction of a force in one community will not necessarily be the same as in another, nor in the same community at different times.

For the sociologist the laboratory method is either impossible or very limited. Human relationships cannot be isolated in test tube and tested in flame. He must study passions, attitudes, and activities as and where he finds them. He cannot direct them, he cannot even bring them into being at will. Because of these difficulties sociologists have in some cases decided to content themselves with what they have mistakenly called a descriptive science.

Again "statistics are extremely dry"; every one knows that—we suffer a sort of a psychological complex in the presence of large groups of figures. But if statistics are dry, or are thought to be, it is mainly because of faulty method. Now it is the rôle of the modern statistician to study the psychology of statistical presentation; the problem of how to carry and impress the message. If they are of supreme importance, if they are necessary to the guidance of our lives, if it is impossible for the social sciences to be scientific without statistical method, they cannot be dry.

A word of warning is now in order. What statistics can actually accomplish in its present state of development and what we desire to accomplish through its use are two very different things. At best, the science of statistics is capable of measuring measurable phenomena, of discovering correlations, and other relationships. But to this must be added a large measure of common sense. We might discover, for instance, a direct correlation between the increased consumption of automobiles and the increased amount of divorce, but the mere fact proves nothing. It may be only an accident. Again, we have an amusing illustration of the use of statistics for philosophical purposes, in the publication of Johann Peter Sussmilch, who in 1741 attempted to demonstrate statistically the doctrine of "Natural Order"

in a document entitled "Reflections on the Divine Order in the Mutations of the Human Race as Indicated by Its Birth, Death, and Propagation." Noticing an approximate equality in the sexes at the time of marriage, he saw proof of a divine command in favor of monogamy. Noticing also that the death-rate in the city is higher than that in the country, he decided that in the cities vice and luxury flourish, bringing down the wrath of God, and decided in favor of the country. Faulty as his methods were, we have here the beginning of the use of statistics for definitely social purposes.

Finally a word about statistics and intellectual integrity. It seems impossible that there could be better training in honesty and in exactitude than in the practice of statistics. In order that our statistics be of any value whatsoever, it is necessary that the statistician be scrupulously honest in the use of his figures. He must use exactly what he finds, and report the same. No statistician can be a statistician in the highest sense of the word, and juggle his figures. He must have all of the qualities which we would expect to find in any scientist. He must be willing to face facts, he must be open-minded, he must possess an unbiased zeal for the truth, and he must be capable of application. He must be scrupulously exact in his use of the data in hand, and he must be blessed with the same type of constructive imagination that is used by the exact scientist in his laboratory.

II

THE FIELD OF SOCIO-STATISTICAL FACT

The people who should be interested in social statistics. Social statistics studies society as much in health as in disease. Previous analyses of the field of socio-statistical fact. A new analysis. The whole field of social relationships for all social purposes. Analyses of related fields: business statistics; vital statistics; medical statistics; educational statistics; and mathematical statistics.

STATISTICAL studies have been written for the use of educators, psychologists, biologists, health workers, economists, political scientists, demographers, and many others.

In general the separate fields are fairly clear, but to whom would an *Introduction to Social Statistics* be of interest? To what fields of human endeavor should it be aimed? Perhaps, first of all, we should mention the sociologist, whose aim is the scientific study of society; to discover trends, uniformities, and, if possible, laws. The formal sociologists themselves, however, are greatly in the minority. Scattered over the world are men and women by the tens of thousands who are engaged in active service to society. First in this class come the trained social workers, who are actually engaged in social welfare work. Second, all those whose endeavor is primarily religious: the clergy, church workers, Y.M.C.A. secretaries, and many others. In the third case we might mention teachers, in so far as their work is a definite social endeavor. Teachers are interested not only in the technique of teaching, the technique of the best presentation of their particular subject, but are also interested in the welfare of their students. They are interested in knowing where their students come from, their home backgrounds, their cultural environment, their physical care and nourishment, what are the stimuli which are affecting

their school work, and what can best be done to improve their condition. Teachers eminently need social statistics, and teachers are beginning to use them. Finally we come to that great group of individuals which includes every one from the business man to the club woman, people who, in addition to their normal activities, are interested in doing what they can to improve social conditions. To this large class of people, then, who are interested in the problems of social welfare, who are interested in knowing exactly what the social conditions are, and how they can be improved, this book on social statistics is addressed.

Modern social life is complex; it is no longer socially permissible for a person out of the goodness of his heart or the softness of his emotions merely to dole out alms to the needy. We realize that such activities merely result in pauperization; that we must, if we are scientifically and properly to ameliorate the conditions which surround us, know the facts of the case at first hand. Again, our social organism is not only so complex, but so large, that it is impossible for us, without the aid of exact statistical information, to know the real scope of our task, and what should be done. Secrist has said, "The possession and use of facts in business today show the difference between success and failure," and in the same manner in our social or religious work it is only the possession of facts which makes success at all possible. The business man who keeps his accounts in his head is fast going under in the economic struggle for existence. Likewise the sociologist who bases his theories on broad generalities and "philosophical reasoning" is becoming a thing of the past. The social worker, in whatever field of endeavor, who is ruled by emotions rather than by facts, is finding the modern social problem far more than he can handle.

Our first task, then, is that of defining the field of socio-statistical fact, the field within which the student of society may expect to find data of his major interest. First, let us

notice that since sociology represents a general, rather than a specialized, field, as is the case with economics and political science, the field will be much broader and indeed will contain certain information already covered by these special social sciences. Not only is it general in nature, but it is also far broader than it has been in the past conceived. The modern interest in the statistical study of society and its problems probably received its first definite impetus from Mr. Booth's *Life and Labor of the People*, which appeared in 1889. Mr. Booth was first of all interested in, and studied conditions of, poverty. Since his time statistical method has been increasingly used in social studies, but still mainly for purposes of amelioration of adverse social conditions. Some one has said that the best way to understand the physical body in sickness is to study it in health. Similarly, the best way to understand society in an abnormal state is the study of society in a perfectly normal state. Thus statistics must not be limited merely to a study of social pathology but must include also a study of society in health. Again, we must not limit ourselves merely to studies in demography and vital statistics. The study of the social organism as a whole in all of its manifestations must be our field of endeavor. The social survey must aim to reveal the good as well as the bad.

It might be well now to notice the organization of the field as presented by Mayo-Smith in his *Statistics and Sociology*, which appeared in 1904. He analyzes the field in the following manner:

1. *Population* in itself is a very general conception, and does not carry us far in describing social organization. It is necessary to proceed at once to classification. The most important lines of classification are as follows:

(a) *Demographic classes*. The individuals of the population are distinguished according to differences of sex, age, conjugal condition, and physical health. Connected

with this classification we have the social phenomena of births, deaths, marriages, and sickness. All these are important facts influencing the social organization of every community.

(b) *Social classes*. Men differ in social position, religious confession, occupation, material condition, and ethical conduct. The differences in societies in these respects often mark the stages of social well-being and make for progress, peace, and stability, on the one hand; or retrogression, discord, and instability, on the other.

(c) *Ethnographic classes*. Men differ in race, blood, and nationality; hence societies differ in ethnic composition, race character, national ideals, and political allegiance. These facts have attracted much attention because they underlie political science.

2. *Physical environment*. Population is possible, of course, only as conditioned by physical environment. The relations of population organized in society to the land are important facts for sociology, and these relations have to be followed out in all the divisions of population noted above.

3. *Social environment*. As men live together in social relations they develop language, customs, institutions, social and ethical standards and ideals. These constitute what Herbert Spencer calls the superorganic environment. By them all the phenomena of social life are modified, and the new relations thus established are, perhaps, the most important elements of social organization. The relation of the population to the physical environment is also very much modified by these influences.

4. Finally, we have the fact that *social relations are changing*. It is necessary to trace these changes and how they affect social organization. It is here that we come across the doctrine of social forces and need some scientific method of measuring their strength. This is sometimes called dynamic sociology, but the distinction is not an im-

portant one for us, as in our study we shall find that all society is dynamic.

Bailey, in a similar effort which appeared a few years later, divides the field much more simply as follows:

1. Sex, age, and conjugal condition
2. Birth
3. Marriage
4. Death
5. Growth of population

I am afraid that, excellent as these works are, we must conceive of the field of socio-statistical fact in even larger terms. The very largeness of the field may result in some discouragement, but at least it indicates to the socially minded individual that source from which he can draw exact statistical information for the discovery of facts, uniformities, and laws. The following outline aims to be suggestive rather than all-inclusive:

THE FIELD OF SOCIO-STATISTICAL FACT

1. *Population quantity facts:*

Births, deaths, immigration, emigration, density of population, population shift.

2. *Population quality facts:*

Defect, subnormality, disease, normality.
Leadership, genius, talent.
Illiteracy, poverty, pauperism.

3. *Population composition facts:*

Age, sex, color, race, nativity, parentage, conjugal conditions, occupation.

4. *Socio-ethical facts:*

Marriage, divorce, desertion, illegitimacy.
Crime, suicide.
Philanthropy, charity organization.
Educational association, religious association, political association, fraternal association, social cooperation.

5. *Socio-economic facts:*

Hours and earnings, conditions of labor, industrial acci-

dents, compensations, child and female labor, unemployment, industrial unrest, national prosperity, distribution of wealth, standard of living, housing, occupations.

6. *Socio-political facts:*

Social legislation—facts indicating the social efficiency of political organization, elections, shift of allegiance, use of franchise, citizenship.

7. *Socio-psychological facts:*

Lynchings, revivals, panics, fashions, crazes, booms.
Constructive cooperative effort.

The whole field of social relations, then, whether they have to do with the population itself or with ethical, economic, political, or psychological factors which rest thereon, is the field of endeavor with which the social statistician has to contend.

The earliest censuses on record have to do with the gathering of population quantity facts, for the purpose of aiding the monarch in determining his potential wealth and power both in times of peace and in times of crisis. These facts are of as great value to us today for the same purposes, but we are beginning to realize that they are of even greater value to us as indexes to many specific conditions of general social welfare.

It is interesting to note, for example, that in neither France nor China does the population seem to be increasing at any very marked rate. In the case of France, this may be and probably is due to the exercise of prudential restraint, whereas in China it is a strong indication of the tremendous pressure of the population on the food supply and a high death-rate. Many seemingly strange social conditions in China can be explained directly in terms of population pressure, while in our own earlier national life the absence of population pressure and the presence of free land had their very definite effect in forming our social structure. It is only now and as a result of feeling this pressure

that we are beginning to revise our opinions on the desirability of free and unrestricted immigration. We have gone further and are studying the problem of our own genetic increase in the light of statistical facts. We have discovered that the "race-suicide" specter is less of a danger than we had thought.

The registration of births and deaths was once left to the haphazard records which were kept by the clergy, while today most of the states in the union are in the compulsory birth and death registration area. We have long been numbering our live stock, but now we are beginning to realize the necessity for knowing what is happening to our children. With regard to the quality of the social population we know very little, but the next generation will know vastly more. We have made a start in gathering statistics of sex, disease, and subnormality, but we are still neglecting the field of leadership, genius, and talent. With regard to the composition of the social population we have done much in America, probably due to the fact of the composite nature of our population, which has in itself given rise to many of our social problems.

Let us now turn to some typical analyses of related fields. These will be suggestive to us as indicating what we should expect to accomplish within the field of sociology.

The first and most suggestive is contributed by Secrist in his book entitled *Statistics in Business*. His analysis of that field is as follows:

1. *Labor facts:*

- Wages, wage-rates, and earnings
- Hours
- Union membership, and types
- Strikes and lockouts
- Accidents
- Compensation, industrial and health insurance
- Unemployment
- Turnover
- Immigration

2. *Capital facts:*

Capitalization

Resource

Earnings

3. *Production facts:*

Agricultural crops

Raw materials extracted

Manufactures

Stocks on hand

4. *Population facts:*

Geographical distribution

Occupations

Earnings

5. *Market facts:*

Money

Banking

Credit

Trade

Commodity prices

Security prices

Competition

Failures

Forecasting

6. *Land and building facts:*

Land and building values

Taxes

Restrictions on use

7. *Transportation facts:*

Foreign rates and fares

Foreign facilities

Domestic rates and fares

Domestic facilities

Domestic control

8. *Law and policy facts*, as expressed through law and the activities of:

State railroad and utility commissions

State tax commissions

State industrial commissions

Federal Trade Commission

United States Department of Agriculture

United States Geological Survey
Interstate Commerce Commission
United States Tariff Commission

The field of *vital statistics* is here illustrated by Whipple's book under that title. In this book, in addition to a discussion of statistical method, he discusses the following types of statistical data as applicable to the field:

1. Under the head of "Enumeration and Registration," he discusses the United States Census; Census Data; Census Schedules; Registration of Births, of Deaths, of Marriage, of Morbidity, of Notifiable Disease; Venereal Diseases; and Morbidity from Non-reportable Diseases. Also Registration Laws; the Registration Area; Sickness Surveys; and the Need of National Statistics.

2. Under the head of "Population," he discusses the Enumeration of Population; the Estimation of Population; the Rate of Increase, or Decrease; Immigration; Urban and Rural Population; Density of Population; Classification of Population, as to color, race, nativity, and parentage; Sex Distribution; Dwellings and Families; Age Distribution; Age Groups; Ratios and Standards.

3. The prediction of future population.

4. General death-rate, birth-rate, and marriage-rate.

5. The city death-rate.

6. Causes of deaths.

7. Analysis of death-rates.

8. Statistics of particular diseases.

9. Study of deaths by age group, and finally life tables.

Pearl, in his *Medical Biometry and Statistics*, seems to limit his particular field of investigation to the causes of death. This he calls "The Raw Data of Bio-Statistics." He suggests the following simplified classification as to causes of death:

1. Circulatory system, blood and blood-forming organs
2. Respiratory system

3. Primary and secondary sex organs
4. Kidneys and related excretory organs
5. Skeletal and muscular systems
6. Alimentary tract and associated organs concerned in metabolism
7. Nervous system and sense organs
8. Skin
9. Endocrinal system
10. All other causes of death

Still another field is that of *Education*. In this field the educator using statistics will be interested in the measurement and comparison of intelligence coefficients, and with all that group facts which have to do with the growth and development of the intelligence of the student.

Let us now turn to West's *Introduction to Mathematical Statistics*. Here we find an entirely different problem.

The *mathematical statistician* is interested in no particular type of data, but in all data, of whatsoever sort, in as far as they may be statistically treated. He is interested in the uniformities, uniformity laws, and methods which may be applied equally to the fields of educational statistics, social statistics, economic statistics, vital statistics, or of whatever sort they may be. It is interesting to notice that West analyzes his field under the following heads:

1. Curve plotting
2. Frequency curves
3. Averages
4. Forms of distribution
5. The normal probability curve
6. The correlation table
7. The correlation ratio
8. The coefficient of correlation
9. Correlation from ranks
10. The moments of distribution
11. The theory of correlation
12. The method of contingency

From this brief analysis of the field of statistical fact and of statistical method it will be noticed that each one of the statisticians will use the method of mathematical statistics

in so far as any of these methods are applicable to his particular field. Therefore, in social statistics we shall expect to find a sufficient amount of mathematical statistics to deal with our particular data. For instance, the rates and ratios which are so largely used by the vital statistician are of little importance to the economic statistician, while the method of index numbers which plays a large part in the field of economic statistics is of less importance to the social statistician. Other illustrations might be given, but this will be sufficient for our purpose.

III

STATISTICAL ACCURACY

Statistics as the exact scientific treatment of approximations. Mechanical and tabular aids to accuracy. The use of round numbers. Continuous and discontinuous series. Sources of error: fictitious accuracy; cumulative and compensating errors; the mislocation of decimal points; fictitious accuracy in totals; and personal bias. The standard of accuracy. Editing for accuracy. Accuracy and the statistician's reputation.

For the person who is going to use statistical method there are numerous pitfalls. He must be on constant guard against fictitious accuracies of many sorts. The following discussion will point out some of these dangers and will indicate methods whereby greater accuracy can be attained. The practical application of this discussion will appear in the succeeding chapters.

THE RELATIVE NATURE OF STATISTICS

The numerical terms in which the results of statistical analyses are presented indicate a larger amount of accuracy than, in fact, is usually present. Their very definiteness carries with it weight and conviction, and the one person who should not be overimpressed with their significance is the statistician himself. Chaddock has remarked that "a healthy skepticism is probably the most essential quality of a statistician." The author of one of the most widely quoted statistical analyses of the present generation is reported as remarking that he wished he were as sure of his results as the people who so frequently quoted them. Statistics is a science of averages. It deals with average relationships, and no single numerical value can adequately represent either the entire universe of a variable or even a small por-

tion of it. The mathematician can, with a fair degree of accuracy, demonstrate that the product of two and two is four. But, when the statistician says that two and two equals four, he means that on the average, in so far as the two's are really representative of the universe, on the average and within certain limits, the product will be four. The United States Census is probably one of the most carefully gathered bodies of statistical data in existence. It aims to deal with universes rather than with samples. By a universe we mean the ideally entire group as distinguished from a representative sample of that group. In studying divorce in the United States, for example, the universe would be represented by all divorces granted within the period of time studied. Usually, however, the statistician has to content himself with the study of a random sample which he has reason to believe will prove representative of the universe of divorce. Still we know that even so simple a matter as numbering the population cannot be accomplished with exact accuracy, and when we come to the estimation of rates and values, the possibility of error is greatly increased. While, then, the statistician deals with relative truths and approximate accuracies, he has no excuse for slipshod methods and careless work. By error we do not mean here, the errors which result from bungling work, we mean rather that unavoidable amount of difference by which the character of the sample departs from the normal character of the whole. The statistician must use care to inject no more inaccuracy than is necessary in dealing with data which are at best representative. Statistics, then, is the exact scientific treatment of approximations.

MECHANICAL AIDS

Let us consider the mechanical devices which the statistician has at his disposal to aid him in the computations which he must make. Perhaps the first and most important

of these is the slide-rule. Even the 10-inch rule, which costs but a few dollars, is of inestimable value in working out rates and ratios. Only three digits can be read on the lower portion of the rule, and two on the upper, but the fourth or the third place can, with a fair degree of accuracy, be approximated. On the more complicated Fuller and Thatcher slide-rules four places can with ease be read and the fifth estimated. Usually this is all that is necessary for the estimation of rates. For more complicated computations, calculating machines are available. With these, absolute accuracy to a large number of places is possible. The statistician also needs a standard adding machine. His hand-multiplications and divisions are apt, on the average, to be more accurate than his sums. The adding machine also provides him with a printed slip which he may use for the purpose of checking his data, and provided he has made no mistake in operating the machine, the machine is incapable of giving him erroneous returns. Tabulating machines, elsewhere discussed, will seldom be at the disposal of the statistician, but where they are, another possible source of error is eliminated.

TABLES

One is not surprised when a person does not know how to extract square or cube root; indeed, much time would be wasted if he attempted to do so by hand. Squares, and sometimes cubes, can be read directly from slide-rules, and tables of squares and cubes are easily available. If he familiarizes himself with the use of logarithms in mathematical processes, he will find the use of the tables of logs of great value to him, both in increasing accuracy and reducing the time element.

ROUND NUMBERS

For some purposes the use of round numbers is preferable to the use of numbers of numerous digits. In reducing com-

plicated numbers to round millions, thousands, or hundreds, or in reducing numbers in which decimal places are present to full units, the last digit beyond the point so designed, if it be more than one-half, is considered as one full unit, or if it be less than one-half, is dropped off entirely. In case the number be exactly one-half, that is, 5, it is a good practice to raise alternate fives to unity and drop out the others. Thus, the number 1692.62 in terms of units would be read 1693; in terms of hundreds, would be read 17; and in terms of thousands, would be read 2. Primarily, however, *round numbers are of use not in the making of mathematical calculation, but in presenting ultimate results:*

385.56	385.6	386.	390.	400.
177.77	177.8	178.	180.	200.
122.22	122.2	122.	120.	100.
<u>685.55</u>	<u>685.6</u>	<u>686.</u>	<u>690.</u>	<u>700.</u>

CONTINUOUS AND DISCONTINUOUS SERIES

Series of variable are of two sorts, *continuous* and *discontinuous*, and should not be confused in their statistical treatment. A continuous series is one which is capable of fine measurements of magnitude. An example of this would be the height of individuals. It would be possible to measure them to thousandths of an inch and no two heights would of necessity fall at exactly the same point. *There is a continuous range of possibility.* An example of discontinuous series, on the other hand, would be that of the size of families. The sizes are sure to fall on whole units. For example, a family may be 4, 5, 6, or 10 in size but never $7\frac{3}{4}$ or $8\frac{1}{2}$. Again, income cannot be measured in smaller units than cents and they are usually measured in round dollars. A continuous series may be rendered discontinuous by compromising on the degree of accuracy with which the variable will be measured. If it is the case of the individual already mentioned, we may decide to measure him to the nearest inch, or if it be the distance between cities, we will com-

promise on the nearest mile. In the tabulating and graphing of continuous and of discontinuous series there are some differences in treatment which must be observed.

SOURCES OF ERROR

Fictitious accuracy is one of the main vices of the pseudo-statistician. If, for instance, three men are 60, 65, and 70 inches tall respectively, it is accurate, within the limits of the accuracy of the individual measurements, to say that their average height is 65 inches; but if one of them be 61 inches and the other two 65 and 70 inches, the same mathematical process by which we total their height in inches and divide by 3 will give us a result 65.333333 indefinitely continued. This is manifestly inaccurate, since the indefinitely continued decimal places indicate a degree of accuracy which was not present in the original items. The presence of digits beyond the decimal point indicates accuracy in the original items to the same extent and would be so interpreted by the statistician. It may be stated, therefore, as a general rule, that no total or product should show a higher degree of accuracy than is present in the original items or at best not more than one decimal point beyond. We would be willing to allow, for example, on the basis of heights measured in units of inches an average measurement of 60.5 inches, but not one of 60.55. In the same manner, if accuracy in the individual item be measured to the third decimal place and the sum or the product resulting is such a figure as 7.000 the three zeros must be retained. They may not be dropped off, because they indicate an accuracy of the 7 to that fineness of measurement.

Cumulative and compensating errors. Two more errors which may be pitfalls to the statistician are the *cumulative* and *compensating errors*. An example of the cumulative error would be the case of a yardstick which was one-tenth of an inch too long or too short. This is a cumulative error

in that the error would accumulate with every measurement in which the yardstick was used. Surveyors' chains are capable of stretching and their instruments of being off level and this results in the accumulation or the amassing of error. A compensating error, on the other hand, is usually an error of judgment in which it is as apt to fall on one side as on the other. A mistake in one direction, therefore, is apt to be compensated by a similar and equal mistake in the other direction. It is fortunate for the statistician that compensating error plays so large a part in the making of surveys, but the statistician is warned not to rely too much on the compensating nature of his errors.

Location of decimal points. Another error—the necessity of mentioning it at all is unfortunate—lies in the *location of the decimal point* in the simple mathematical process of multiplication and division. For instance, in the squaring of .25 the numerical result is 625, but all too many people are puzzled to know just where to place the decimal point. In slide-rule operations where the mathematical process is absent, this problem of correctly placing the decimal point is increased.

Accuracy in totals should also be considered. In adding similar variables of different degrees of accuracy the total can never be more accurate than its least accurate element, as in like manner, no chain is stronger than its weakest link. Consequently, in totaling, each item will be entered as it stands, but the sum or the total will be recorded only to the degree of accuracy of the least accurate item:

175.006	175.
25.5	25.
921.43	921.
100.1	100.
<u>92.</u>	<u>92.</u>
1314.036	1313.

From the above two computations it will be seen that it is unnecessary to correct each item in terms of the least accurate item since this injects unnecessary inaccuracy in

the total. Enter the respective items as they stand and correct only the total. Thus the correct total would be 1314 and neither 1313 or 1314.036.

Personal bias. Another source of error in the making of statistical surveys is that of *personal bias*. The investigator must be sure, first of all, that bias in his own case is absent, and next, that methods of collection have been so systematized and standardized that personal bias in the case of enumerators or of individuals filling in schedules will be reduced to the minimum. In this connection all units used within the schedule should be carefully defined. In using the word dwelling-house, for instance, we must be sure that each enumerator is using and interpreting the term in exactly the same manner. (See pages 40ff.)

STANDARD OF ACCURACY

Another group of inaccuracies have to do with the collection of data. Before any statistical investigation is undertaken or as soon as the preliminary survey has been made, *standards of accuracy* should be definitely decided upon. Decide, first, what degree of accuracy is possible, and second, what degree of accuracy is desirable. Shall ages, for instance, be recorded to the nearest day, nearest month, or merely in terms of years? In the latter case shall the year be that of the last birthday, in terms of which ages are usually given, or in terms of the nearest birthday? Care should be taken that the standard of accuracy chosen is a possible accuracy and that the results show accuracy only to that point.

EDITING

When schedules have been returned to the investigator, they must first of all be most carefully *edited*. In some cases the entire schedule will be so defective that it must be entirely thrown out. In others, only portions of the schedules can be used. Recently in a class in statistics the

students were given schedules calling for the age, age at marriage, nationality, occupation, and cause of death (if dead) of both parents and grandparents. The schedules were carefully discussed in the class and each student was given the task of filling out one schedule for himself and of acting as enumerator and having four schedules filled out by student friends. Of the 125 schedules returned, 35, or 28%, were sufficiently defective in one or more respects to warrant discarding them. If this can happen with a group of budding statisticians, it is merely an indication of the care which must be taken in editing schedules which come in by the correspondence method. The investigator is not a mind-reader and, if he is an honest statistician, he will not "fudge" his results. He will proceed to the statistical analysis only of authentic statistical data, will treat them without bias, whether they show what he is looking for or not, and give an honest report to the world. He will not attempt to compare the incomparable, and he will not draw sweeping conclusions from inadequate samples.

EXACTITUDE

The reputation of the statistician is most at stake in evidences of *inaccuracy in tabulation and graphic presentation*. If it is evident from the scrutiny of a table that totals do not actually total or that points on a graph or chart do not correspond with the data which they represent, suspicion is at once cast on the entire project and the reader can only conclude that the statistician was capable of careless work. The statistician cannot view his own work with too great suspicion or exercise too great care in the exactitude of his methods.

IV

GATHERING THE DATA

Sampling *versus* accounting. The law of statistical regularity. The law of the inertia of large numbers. Secondary sources and their use. Primary data and how to gather them. Types of primary investigation. A practical example of the schedule maker's art. Some samples from the hands of amateurs.

HAVING considered the vital importance of statistical analysis and the legitimate field of socio-statistical research, we now arrive at the first problem which confronts the statistician, that of obtaining the data on which his research will be based. The first point to be observed is that statisticians seldom deal with a field of data in its entirety or what the statistician calls a complete "universe." They draw their conclusions, in the great majority of cases, from *samples*, and indeed were it not for the statistical law of samples and their application, the modern science of statistics would be impossible.

THE SAMPLING PROCESS

Statistical analysis is not primarily an accounting proposition. The aim of the accountant is a complete count which is accurate to the last detail. He must weigh and count every item, while "statistics endeavor to carry the formation of significant judgments beyond the limit where perfect and complete count is possible." It is seldom possible even in limited fields for the statistician to attempt what is done by the United States Census, in counting every individual and facts concerning him, living within the borders of the country at a given time, and even here it must be remembered that not only is an absolutely complete count physically im-

possible, but the count does not remain accurate for one day after it has been made.

The primary assumption on which the statistician proceeds and which possesses the regularity of a law is that "*a moderately large number of items chosen at random from among a very large group are almost sure on the average to have the characteristics of the larger group.*" In other words, adequate samples do represent the universe, and on the basis of adequate samples general trends may be discovered. If this is true, and if the samples prove to be adequate, then successive samples within the same group should show approximately the same tendencies. This is one method of testing the sample to see whether or not it is adequate. Ascertain whether or not 50% to 75% of the sample shows approximately the same characteristics as the entire sample.

Another interesting fact which makes statistics possible is that of the *inertia of large numbers*. This is illustrated by the fact that while in given communities we may find considerable variations from year to year, in birth, death, and marriage rates for the country as a whole, we find these rates to be increasing or decreasing at a very regular rate, or remaining stationary. Again, while we may have a crop failure in one given portion of the country, the production of wheat, let us say, for the entire country will not show a similar variation. This is the law which makes all types of insurance a scientific rather than a gambling proposition. The discovery of this law was made possible through the sampling process. Because adequate samples are truly representative of the whole, it has been possible for us to engage in extensive cost of living studies, to make market analyses and forecasts, to map out social programs, and to make the host of social studies on which proposed national social progress is based.

Whipple has the following to say which bears upon the subject:

In using statistics we necessarily employ the methods of logical

thinking comprised in what is termed "induction" methods, by which general tendencies and laws are found from accumulations of facts.

Statistical induction may be said to be one of the most potent weapons of modern science. Referring to it Royce said that the technique of statistical induction consists wholly in learning how to take fair samples of the facts in question, and how to observe these facts accurately and adequately.

Statistics are being constantly invoked for testing hypotheses in all branches of science. This involves four distinct processes: (1) the choice of a good hypothesis; (2) the computation of certain consequences, all of which must be true if the hypothesis is true; (3) the choice of a fair sample of these consequences for a test; (4) the actual test of each of these chosen hypotheses.

Deductive reasoning as well as inductive reasoning is involved in the use of vital statistics. It is perhaps the natural order of mental processes for the mind pursuing an inductive study to leap ahead to some conclusion and then fill the intervening steps by working backward by deduction.

It is by the application of the principles of logic that the statistician is able to keep his conclusion within reasonable bounds.¹

Since so much depends upon the statistical sample used, it stands to reason that the greatest care should be exercised in the choice of that sample. In the first place, the sample should be truly representative of the problem involved. It should be typical rather than exceptional. It would be possible to prove absolutely anything by statistics if great enough care were taken to choose a sample which would represent the point which it is desired to have demonstrated. In the second place, in order to be truly representative, the sample must be of adequate size. In the third case, the selection of items must be purely random. This is to avoid any personal bias in their selection.

It is well to note that merely doubling the size of the sample does not double its precision. In order to do this, it is necessary to increase the size of the sample by four.

As has been indicated by the foregoing, statistical studies are, in the main, of three sorts, (1) accounting studies of

¹ Whipple, *Vital Statistics*, pp. 14-15.

complete universes; the United States Census is representative of this type. (2) We have what Whipple calls the Monographic method, which is a complete and detailed study of a particular group. And (3) we have the sampling or representative method which studies a sample as representative of the whole. This last is the more common statistical practice.

THE USE OF SECONDARY SOURCES

Let us next turn our attention to the technique of gathering data for statistical samples. Statistical data are of two sorts, primary and secondary. By primary we mean data which are originally gathered by the investigator directly for original study. By secondary data we mean all of those original studies which have already been made; collections of data which may be brought together and used for the purpose in hand. It is important to call attention to the fact that due to undeveloped and unstandardized statistical method, much of this secondary statistical data cannot with entire safety be used.

In making statistical studies which are based on secondary sources, the careful statistician will always approach these sources with the greatest suspicion. Even though they have been made by reputable scholars, he must assure himself that they knew the technique of statistical analysis and presentation. He must take into account possible bias, the size and location of samples, methods of gathering them, the time element or period represented, how accurately it represents its universe, and finally and most important of all, the definition of units. Even the United States Census is capable of error. Two identical studies of the standard of living of laborers, for instance, cannot be compared until we know how each investigator has defined the terms "laborer" and "standard of living"; wide variation is here possible. With regard to this type of investigation from secondary

sources, we will content ourselves by saying that the greatest care must be taken in combining data from various sources to assure us that they are really comparable.

THE PRIMARY OR ORIGINAL INVESTIGATION

It has been well said that the most important part of any investigation is the preliminary work done before any attempt is made to gather the data. In fact, it is a common practice where large investigations are under contemplation to organize a preliminary survey in order to test out the methods proposed and the data which it is hoped to secure. The careful student of society does not indulge in the haphazard collection of data from which he hopes to stumble on to social laws and explanations. Before he begins his investigation he is very careful to map out his field, to prepare and test his schedules as carefully as possible. He will have in mind very definitely the exact problem that he is attempting to analyze. He will study it in all of its aspects, in order to come to a definite conclusion as to what he wants to find out. In this connection he will, first of all, study the available secondary sources at hand, in order to see what others have done with this problem and from these he may gain many valuable hints as to methods of procedure and sources of information as well as some of the difficulties which may be met. He may find that this problem has already been solved or that portions of it have been adequately studied. After having defined his problem and having in mind his secondary sources, he will proceed to study the field for the purpose of determining where he can best secure the data which he desires. These data may be in the form of official or private records, or they may exist merely in the minds of men. In the latter case it will be necessary for him carefully to plan his approach to them for the purpose of getting from them the exact and uncolored data which he desires.

His next problem will be that of selecting the best method for gathering the data. In some respects the most effective method is that of the intensive one-man survey, where the investigator goes from individual to individual, personally gathering the information which he desires. This method has the advantage of a consistent personal point of view but it has the disadvantage of being a slow and cumbersome method which is decidedly limited as to the size of the field which can be covered. This was the method which was used by Engel in his now famous "Standard of Living" investigation.

A variation of this method is that of securing and training investigators who will take given groups or districts and carry on parallel investigations. This has the result of speeding up the work and makes it possible to cover a larger territory, but naturally it has the disadvantage of increasing cost and the consistency of the personal point of view, since no two investigators, no matter how carefully trained, will react to the same situation in exactly the same manner.

A third and at present a too carelessly used method is that of the mail-questionnaire, in which there is no personal contact between the investigator and the person investigated, schedules being prepared and sent out by mail to the individual possessing the information and they returning them with their answer. Through this method the maximum of speed is often obtained. It has also the advantage of being cheap and making it possible to cover a wide territory. The disadvantage, however, is the same as in the preceding case, for the personal bias is now as large as the number of persons who are returning replies. Answers, too, tend to be careless when not checked up by an enumerator.

An additional variation may be noticed. Investigations may be of the above cross-section type, in which information is desired from a given territory or group of people representing a given instant or period of time. On the other

hand, enumerators or investigators may be secured for the purpose of making periodical reports. This method, which is more properly called registration, would be the one used in the securing of crop or weather reports, school attendance, and vitality facts. The result is a continuous or a historical picture rather than a cross-section at a given time.

All methods are used by the United States Government in the statistical study of its population and their problems. In the main, however, the method of gathering data for the decennial census is that of giving more or less training to paid investigators who secure, with the aid of carefully prepared schedules, the information desired. In addition to this, corps of permanently employed enumerators periodically report on weather, market, crop, vitality, and other facts. The last method, however, valuable and desirable as it is, is limited in its application because of the great expense entailed. Only a government, large corporation, or foundation can afford to use it. But whatever the method which may be used, the main problem is that of the careful preparation of schedules on which the data will be entered, the definition of units where there is any possibility of misunderstanding, and complete and adequate instructions for enumerators which will be 100% fool proof.

The schedule should, in the first place, whether it is placed in the hands of individual enumerators or sent by mail, be as simple and concise as possible. Complicated schedules covering pages of printed material and requiring detailed information invite confusion, resentment, and error. Information should not be called for merely because it is interesting, or because it might possibly prove germane to the investigation. The information desired should be reduced to its simplest terms. No question should be asked unless it is certain that it is going to be needed. Care should be taken again in asking questions on which there is danger of personal bias. It is useless to ask Democrats what they think of a Republican administration. Do not ask questions

which are unnecessarily inquisitorial or delicate in nature. Ask a woman, for instance, what year she was born, but never ask her how old she is. Make it a general practice to ask questions which can be answered in terms of a number or in as few words as possible. Questions which can be answered by "yes" or "no" are far preferable to those which require a sentence. The form of the schedule should also be given attention. It will be possible to so arrange the blanks for answers that the schedule will easily adapt itself to rapid reading and editing. Where possible, it is a good practice to so arrange the schedule that the data desired can be entered directly into blank tabular forms prepared for that purpose. Again, when numerical data are in the main desired, it is a good practice to so arrange the schedule that the figures will appear in one column at the left of the schedule rather than scattered in a haphazard manner over the sheet wherever a question happens to end. Definite rules cannot and should not be laid down. Each schedule, and this applies to statistical presentation in general, represents a specific problem which must be solved in the light of its particular needs.

Schedules sometimes are so complicated as to require more time on the part of the informant to find out what they are about than in securing and entering the data. One of the best schedules the writer has seen was contained on one side of a post-card. Its clarity and brevity was a pleasing surprise which got results. A 5-by-8-inch card or one side of a sheet of typewriter size paper will more easily get results than a 4-page folder.

The questions asked should be so clearly stated that there can be no possibility of misinterpretation. Where the schedules are placed in the hands of enumerators, manuals of *very complete* instructions must accompany them. *Never take anything for granted* either with enumerators or informants unless it be that they will always make the wrong interpretation when there is any choice.

Where schedules are mailed out in the mail-questionnaire method, a letter which is a psychological work of art must accompany them. It must predispose the person being investigated to answer the questionnaire rather than to throw it in the waste-basket. An appeal to his interest must be made, and he must be convinced that the investigation is not only worth while to the investigator, but that it may prove worth while to him to spend his time on it. In addition, the letter must contain such information as cannot be embodied in the schedule, for the purpose of helping the informant to give the information which is desired. Due to the fact that the mail-questionnaire has been so widely used, it is necessary always to enclose a stamped addressed envelope, if the investigator hopes for many replies. The writer has in mind one investigation of this sort in which a dollar bill was pinned to the schedule indicating that the investigator realized that the time of the person being surveyed was worth paying for, and that he considered his investigation sufficiently worth while to pay for it. Of course, this is very exceptional and is a practice which will not and cannot be widely used. But in this day of bombardment by questionnaires it is necessary sometimes to do something drastic in order to secure replies. Ordinarily a 25% to 35% response to even the best of questionnaires, sent to persons definitely interested, is considered a good return.

According to King¹ the questions chosen should be:

1. Comparatively few in number;
2. Require an answer of a number or yes or no;
3. Simple enough to be readily understood;
4. Such as will be answered without bias;
5. Not unnecessarily inquisitorial;
6. As far as possible corroboratory;
7. Such as directly and unmistakably cover the point of information desired.

Form 1 is a practical example of the schedule-maker's art. The schedule itself is well-balanced, compact, and concise.

¹ *Elements of Statistical Method*, p. 57.

SURVEY OF TUBERCULOSIS CASES

1. Place _____ 2. Address of Patient _____ 3. Date of Record _____

4. Name of Patient _____ 5. Sex _____ 6. Race _____ 7. Date of Birth _____

8. Married, Single, Widowed, Divorced, Separated (Check). 9. Country of Birth (a) of Case _____ (b) of Parents _____

10. Legal Residence _____ 11. Length of time (a) in County _____ (b) in particular community _____

12. (a) Number of Children (under 16) in Household _____ 12. (b) No. of Child. who have had med. exam. _____

13. (a) No. of Other Cases of Tuberculosis in Family _____ 12 (d) No. of Adults who have had med. exam. _____

14. (a) Occupations Prior to Illness _____ 13. (b) Specify Relationship to Patient _____

Diagnosed _____

Suspected _____

Dead _____

15. Approximate Date of First Symptoms _____ (b) No. of Years in those Occupations _____

16. (a) Date of Diagnosis _____ (b) Where Diagnosed _____ (c) Type _____ (d) Sputum _____

17. (a) Present Occupation _____ (b) Employer _____

18. Sanatorium Experience (a) where _____ (b) Dates _____

19. Clinic Experience (a) where _____ (b) Dates _____

20. Source of Record _____ 21. Date of Filling Out this Card _____

22. Present Disposition of Case _____

23. Notes _____

Copyright, 1924, National Tuberculosis Association

Form 1: An excellent example of a complete, compact, and concise schedule.

It can be contained on a 5-by-8-inch card. It is the result of careful thought and goes right to the heart of the problem by asking for the exact information needed and no more. It would seem that one might safely place this schedule in the hands of any competent enumerator, yet notice the detailed instructions which Miss Whitney (statistician, National Tuberculosis Association) has evidently found it necessary to prepare. She has anticipated the injunction "to take nothing for granted." Only experience with schedules and questionnaires can demonstrate the above need for exactitude.

INSTRUCTIONS TO ENUMERATORS USING
"SURVEY OF TUBERCULOSIS" SCHEDULES

1. *Place.* This refers to the minor civil division—city, township, or borough, in which the patient lived at the close of the survey year.

2. *Address of patient.* This item is for the convenience of the surveyor. Give the last address known and if newer information is received substitute this for the first record. In urban communities the house number and street name make the locating of the person easier; in the country, distances from specific urban communities, the number of houses beyond a certain crossroad or identifying church or school building, are most helpful.

3. *Date of record.* This refers to the date on which the case is reported to the public health authority, of whatever title, whose duty it is to record cases of communicable disease.

4. *Name of patient.* Give surname first. Follow with title and given names in order, as Jones, Mrs. Mary Ellen. Titles are not always given but they are particularly helpful in certain sections of the country. In the South the name Willie for a girl is common and likewise Peter has been known as a cognomen for a woman. Full names are less apt to cause confusion than initials.

5. *Sex.* State M. or F.

6. *Race.* This refers to the basic racial stocks, White, Black, Indian. The abbreviations W. B. Ind., are sufficient. "The term 'white' as used in the census reports refers to persons understood to be pure-blooded whites. A person of mixed blood is classified according to the non-white racial train, or if the non-white blood itself is mixed, according to his racial status as adjudged by the community in which he resides. Thus a person of mixed white and Indian descent, or of mixed

white and Negro descent, is classified as an Indian or as a Negro, as the case may be, regardless of the amount of white blood; and a person of mixed Indian and Negro blood, or of mixed white, Indian and Negro blood, is classified either as an Indian or as a Negro, according to his racial status in the community in which he lives."—*U. S. Bureau of the Census, Population, 1920, p. 10.*

7. *Date of birth.* State month, day, and year of birth. Be sure to write out the month, as September 10, 1924. In abbreviating by numbers there is apt to be error. Some persons would indicate the above date as 9/10/20, others as 10/9/24. And errors are made more easily when figures rather than letters are involved. Do not endeavor to obtain the exact date of the month except in cases where the patient died or moved away within the survey year. Likewise it is not necessary to press for the exact age if the age as given will fall within one of the age groups, such as 25-29, 30-34, 35-39, used in tabulating the material. It will be helpful in future to the local worker to know just when the patient was born, but for the purposes of the report it is not necessary to be so specific if the facts can be tabulated as reported.

8. *Check the word applicable.* Do not check more than one.

9. *Country of birth* (a) of Case. (b) of Parents. Give country only, as Ireland, Italy, Norway, United States, Russia. In cases where applicable, indicate Jewish stock by writing Jew in parenthesis after the country as Russia (Jew).

10. *Legal residence.* By "legal residence" is meant the minor civil division in which the patient last resided for one full year prior to the date on which his whereabouts were last known, either at the close of the survey year, or the time the patient died or moved from the country. Give township or city if the patient's legal residence was within the area being surveyed, otherwise the city or county and state.

11. *Length of time* (a) *in county.* Give total length of time patient resided in county prior to the close of the survey year or to his death or departure from the county with length of each period of residence, as: 16 yrs. 5 yrs. 3 mos. This might indicate that the patient was either a resident or non-resident. To clarify this point it helps to note the time absent from the county thus: 16 yrs. (away 10 yrs.) 5 yrs. (away 3 yrs.) 3 mos. This would indicate a non-resident. If the figure in the second parenthesis were 3 mos. instead of 3 yrs. the patient would be a resident. If the last period of stay has been less than one year, indicate in months, otherwise the number of years is sufficient unless there is some unusual circumstance involved.

(b) *In particular community.* Give length of time (as above) during which the patient resided in the last minor civil division in which he lived in the county.

12. (a) *Number of children (under 16) in household, and (c) number of adults in household.* Include the patient in whichever group he falls and after the number indicate that he is included as "6 (incl. pat.)." This latter is not necessary but will be found to serve as a check. As a further check it is helpful to list the members of the household on the back of the card with their names if possible, relationship to patient, and age. This is particularly helpful later if other members of the household are also cases included in the survey. By household is meant all those persons coming under one domestic head. Persons doing light housekeeping in one or two rooms of a house constitute one household as do also the members of a boarding house where the occupants really form a common group.

12. (b) and (d). Note the number in each class who have had a thorough physical examination including a careful chest examination.

13. (a) *Number of other cases of tuberculosis in family.* By family are meant all blood relatives and only blood relatives. (a) Is simply the sum of those cases noted under 13 (b) as Diagnosed and Dead, exclusive of those Suspected.

13. (b) *Specify relationship to patient.* Under Diagnosed are included all those cases active or arrested who have been reported to have been definitely diagnosed as tuberculous by some physician. Under Dead are included those who died before the close of the survey year who were said to have been diagnosed as tuberculous. The facts under 13 (a) and (b) cannot be absolutely accurate because of the fallibility of the informants, and it is not necessary nor would it be possible to obtain diagnoses from physicians in all cases at this time. The relationship should be specified under each, and by note here or elsewhere on the card the name of the relative ought to be entered. This makes it easier to add information on other cards when desired.

14. (a) *Occupations prior to illness.* "Occupation" means "gainful occupation." "Prior to illness" means "prior to approximate date of first symptoms" as given in 15.

List all of the occupations specifically, as winder, cotton mill; salesman, grocery store; stemmer, cigar factory; R. R. brakeman; R. R. clerk; and so forth. All of the occupations from the time the patient began to work should be listed. Do not enter such non-gainful occupations as "school child" or "housewife." If there has been no gainful occupation write "none."

14. (b) *Number of years in those occupations.* Give the length of time of each of those occupations listed in 14 (a). Sometimes this will have to be expressed in months. The tabulations will be made for the predominant occupation, the occupation in which the patient engaged

for the longest period of time, and where there is any closeness in the amounts of time spent in two or more occupations try to ascertain this information called for with particular care.

15. *Approximate date of first symptoms.* Give best date in your opinion. The onset will in many cases be fixed in the patient's mind in connection with a bad cold, influenza, measles, pleurisy, or such, sometimes with an accident. If this is noted on the card, it will help in checking with the approximate date of onset as given by others. The month and year should be given, if possible.

16. (a) *Date of diagnosis.* This refers to the first date on which the patient was ever diagnosed as having tuberculosis and will be given, say, as of September, 1923. In subsequent conversation it will be learned that the patient was diagnosed as having tuberculosis of the bone, or a tuberculous gland, in December, 1905. The 1905 date is the one which should be given on the card.

16. (b) *Where diagnosed.* This refers to the place, in or out of the survey area, where the patient was diagnosed on the date given under 16 (a). It is sometimes interesting to have indicated here whether that diagnosis was made by a tuberculosis clinic or private physician or other agency when it is particularly desirable to evaluate the work of the clinic.

16 (c) *Type.* Here should be included all the forms of tuberculosis which the disease has at any time prior to the close of the survey year assumed in this patient—such as pulmonary, bone, joint, skin, gland.

16 (d) *Sputum.* If sputum was ever examined prior to the close of the survey year, enter Positive, Negative, or Examined—unknown, as the case may be, in this space. If not examined, or if it was unknown whether an examination had been made or not, enter accordingly. Suitable abbreviations may be used. When an examination has been made or more than one, it is helpful for the local person who may subsequently use the information to have the date written opposite the result. No particular effort, however, should be made to obtain the date. If any sputum examination has resulted in a positive finding, it is not necessary to enter the result of any other examination which may have been negative or unknown.

17. (a) and (b). This refers to the occupation of the patient at the close of the survey year. If patient had died or moved away prior to this time, merely enter a dash under each. If the patient was not actively engaged in any gainful occupation, enter "none" under (a) and dash under (b).

18. *Sanatorium experience (a) where and (b) dates.* The answer

to this item should include all time spent in any hospital or sanatorium for treatment of tuberculosis in any form prior to the close of the survey year or the patient's death or departure from the survey area. Exact dates of admission and discharge should be secured if possible, for the patients are apt to misstate the length of stay when the latter is expressed in terms of weeks or months. If desired for some local purpose, the condition of the patient at discharge might be noted when known. If the patient is still under care at the close of the survey year, this fact should be indicated.

19. *Clinic experience (a) where and (b) dates.* Here should be noted any and all clinics which the patient has attended in connection with his illness of tuberculosis in any form, together with the exact dates in attendance, if possible; if the exact dates are not readily obtainable, at least the dates of admission and discharge should be entered.

20. *Sources of record.* Under this heading, in the order in which the information came, should be listed each agency or person contributing any fact to the schedule, as D. C.; H. O.; F. W. S.; Carlisle Clinic: mother; patient. (D. C., Death Certificate; H. O., Health Officer; F. W. S., Family Welfare Society.)

21. *Date of filling out this card.* This has been held to apply to the date or dates on which the home visit or visits were made. If the local organization wishes, it might be stretched to cover the date on which information from any source was added, the date being given in connection with each source.

22. *Present disposition of case.* Under this heading note whether the patient is an active or arrested case, or has died or moved away. In the case of death enter the date, place, and cause of death as given on the death certificate. In the case of a removal enter the date of moving and if possible the place to which the patient moved.

23. *Notes.* Enter here any amplifications of points which will make for clearness and any data of special interest in connection with the case.

In connection with the selection of enumerators *for this particular type of survey* Miss Whitney has the following to say:

The selection of the person to make a case-finding survey is the most delicate, and at the same time the most important part of the study. Handbooks and methodology will not avail if the surveyor be ill-chosen.

Even at the risk of being charged with the description of a

"white blackbird" it might be well to mention some of the qualities and characteristics which a successful case-finding surveyor must have. He (or she) should have had one or more years' experience in the social side of public health work; he should have a working knowledge of statistics; he should be a person who can gain the support of physicians by the thoroughness of his method and by his intelligent knowledge of the medical aspects of tuberculosis; he should have a tenacity of purpose in getting whatever information is available, added to tact and graciousness, without which the patients themselves will not cooperate; and he should top it all with tireless energy.¹

The following are typical schedules and letters to accompany questionnaires. With the exception of the second they are student laboratory exercises and, while far from perfect, represent the type of work which can be done by any person who is familiar with the underlying principles. Number one is a questionnaire which was prepared by a student and later actually used. It was handed to the industrial girls in the institution indicated and it brought results. Number two is the questionnaire referred to earlier in the chapter. It was printed on a post-card and was most effective in spite of the wags who, in holiday mood, took advantage of the opportunity to indulge in their wit. Three was accompanied by a letter. Four, to be filled in by enumerators, was fully explained in a manual of instructions. Five and six are letters in which the attempt, at least, was made to inspire cooperation.

The main defect of these schedules is the chief vice of all inexperienced investigators, that of asking too many questions. The aim should always be "few questions and essential information" rather than "many questions and much information."

¹ *Tuberculosis Case-Finding Surveys*, National Tuberculosis Association, 1925.

SOCIAL STATISTICS

SCHEDULE NUMBER I

INDUSTRIAL DEPARTMENT EDUCATIONAL SURVEY

Madison Young Women's Christian Association

Girls! We want to give you what you want. Help us by letting us know what that is. Return this questionnaire to Miss Simpson as soon as you can fill it out, and accept the thanks of every girl in the association.

1. Name _____
 Address _____ Tel. _____
 Age last birthday _____ Present occupation _____
 Nationality: Father _____ Mother _____ Self _____
2. Education outside the United States.
 Where _____ Years _____
3. Education in the United States.
 Grade school _____ years _____ months
 High school _____ years _____ months
 College _____ years _____ months
 Commercial _____ years _____ months
 Night school _____ years _____ months
 Vocational _____ years _____ months
 Continuation _____ years _____ months
 Correspondence _____ years _____ months
 Summer school _____ years _____ months
 Y.W.C.A. Educational Courses; Check
 English _____ Psychology _____
 Hygiene _____ Bible _____
 Economics _____ Current Events _____
 Why did you leave school? _____

4. Are you satisfied with the future of your job? _____

5. If you could, would you like to return to school? _____
 What courses of study interest you most?
 1. _____ 3. _____
 2. _____ 4. _____
 Why would you like more learning?
 To get a better job? _____
 To make more money? _____
 Other reasons: _____

We thank you very much, indeed!

GATHERING THE DATA

47

SCHEDULE NUMBER 2

OUT-OF-STATE TOURIST SURVEY

Requested by the Wisconsin Highway Commission

TO OUR VISITORS: The Wisconsin Highway Commission desires to serve traffic better each year and asks your cooperation.

Constructive suggestions on Marking, Signing, Maintenance, and Camp Grounds will be appreciated. Will you kindly fill out and mail this card? Thank you.

1. Why did you come to Wisconsin? _____

2. Did you camp or stop at hotels or resorts? _____

3. How many persons in car? _____

4. Amount of money expended in Wisconsin by occupants of car? _____

5. Entered State _____ 192 _____ Departed _____ 192 _____

6. How many miles did you travel in Wisconsin? _____

7. Remarks: _____

Signed _____ Address _____

Date _____ Occupation _____

SOCIAL STATISTICS

SCHEDULE NUMBER 3

CONFIDENTIAL INFORMATION

FOR THE _____ INDUSTRIAL COMMISSION

The information below is required to ascertain the average weekly wage and hours of employees in industries in _____ for the purpose of bettering industrial conditions.

1. Name _____ Address _____
2. Factory where employed _____
3. Hours each day _____
4. Number days per week _____
5. Are you employed during the entire year? _____
6. Number of days lost per year _____
7. Reason for time missed (Check)
 - (a) Sickness _____
 - (b) Sickness in family _____
 - (c) Vacation _____
 - (d) Help not needed in factory _____
8. Wage per hour _____
9. Full time weekly wage _____
10. Are you paid for overtime? _____

SCHEDULE NUMBER 4

FRESH AIR CAMP SURVEY

Being a study of vacation needs of Jewish children

1. Name (family) _____
2. Present address _____
3. Previous address (if any) _____

(No)
(Street)

(No)
(Street)

(City)
(State)
4. Length of time in Madison (Years) _____

5. Present occupation of father, mother, or guardian_____

6. Where employed?_____

7. Contributors to weekly income beside wage-earners_____

8. Weekly income_____

Name	Where Employed	Wage	Amount Contributed
1.			
2.			
3.			
4.			
5.			
6.			

9. Amount saved weekly_____

10. Children:

(a)

Name	Age	Number of Times at Camp	Attend School-	Summer School-
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

- (b) Would the children benefit by the vacation? _____
(c) Physical fitness (to be filled in by medical examiner) _____

11. Has the family at any time received aid from the United Jewish Charities? _____

12. Mother:

- (a) What are the mother's duties in the home? _____

(b) Does she work? _____

(c) Would she benefit by the vacation? _____

13. Condition of home:

- (a) General condition (orderly, clean, disorderly, dirty) _____

(b) Sleeping conditions:

(1) Number in each room _____

(2) Number in each bed _____

(c) Number living in the home _____

14. Surroundings:

(a) How far from a school? (blocks) _____

(b) How far from a park? (blocks) _____

(c) Distance from a recreation center (blocks) _____

(d) Is there any back yard or place in which to play in front of home? _____

(e) Do children play in street? _____

(f) If not, where do children play? _____

15. Remarks:

Information received by _____

Date_____

SCHEDULE NUMBER 5

JEWISH INVESTIGATION COMMITTEE

Dear fellow student:

Do you know what is the matter with the Jewish student body? If you do, you will be doing a service to your fellow members by filling out this questionnaire.

This committee is endeavoring to find the cause for the dissatisfaction evidenced by the Jewish students. After finding out the cause, we shall seek the remedy. We propose a community house, but wish to find out what you would suggest. The friction and lack of cooperation among the Jewish students must stop, and we are enlisting your aid in this problem.

Please fill out the questionnaire and mail so that it will be in our hands by the first of December, 19—, using the stamped and enclosed envelope. We assure you that any information you may give us will be strictly confidential, and the material will be published in mass form only and no names will be used.

Our slogan is "THE JEWISH STUDENT BODY ONE UNIT."

Sincerely yours,

_____(Sec'y)

SOCIAL STATISTICS

SCHEDULE NUMBER 6

CHILD WELFARE ASSOCIATION

November 5, 192—

TO THE SCHOOL TEACHER:

The Child Welfare Association of the city of _____ is conducting a survey to determine, as accurately as possible, the extent to which Moving Picture Entertainments influence the child, during the formative years of his life.

With this purpose in view we ask your cooperation in obtaining the desired information. We ask you to distribute the enclosed questionnaires to the pupils of your grade.

You will observe that no name is required on the questionnaire. This was purposely omitted with the belief that the information so given would be more accurate. In order to further increase the accuracy, we ask you to assure your pupils that this survey is conducted by an organization not connected with the school, and to emphasize the fact that you will not read the questionnaires after the pupils have filled them out.

The questionnaires are to be filled out on Monday, April 11, to show attendance at Moving Picture Theatres during the week of April 4-10.

A stamped addressed envelope is enclosed for your convenience in mailing the questionnaires to our office. Your cooperation in the survey will be greatly appreciated.

Sincerely yours,

(Signed) _____
President.

V

TABULATION

"The shorthand of statistics." The work-sheet. The tally sheet. The perfect title. Good mechanical form. Where to put totals. Tabular perspective. Simple, partial, and complete classification. Chronological, geographical, qualitative, and quantitative classification. The expressing of class limits. Cumulative frequency. Mechanical tabulation.

THE statistical table has been called the "shorthand of statistics." The purpose of using statistical methods is that of bringing facts into relationship with each other so that we can understand their importance and relative significance. This cannot be done merely from the scrutinizing of lists of facts or paragraphs describing them. But it can be done when facts are brought into relationship with each other by means of tables, graphs, charts, maps, and so forth. We might, for instance, describe in great detail the number of convicts in our penitentiaries, the causes for which they are there, the length of their prison terms, their race, nationality, sex, and other facts concerning them. But, be these facts ever so carefully described, they cannot be placed before us in the same perspective that they would appear if in a well-arranged table. The table gives us at once and at a glance all of these facts and we are able to draw conclusions beyond those which perhaps the writer of the paragraph would consider significant. We are able to use them as he used them, and, in addition, to make our own interpretations. The cross-classification table, then, contains all of the valid information. The graph or chart visualizes by another method this same data. We might even go so far as to list them in columns and give totals, averages, and rates.

THE WORK-SHEET

The first step in the tabulation of statistical data is that of the arrangement of the data gathered into the form of a work-sheet, hand-tally cards, or punch cards (where mechanical tabulation is possible). Of the work-sheet (see Table 1) we may say that it is merely a statistical table in

TABLE 1
WORK-SHEET FOR WISCONSIN BIRTH-RECORD STUDY

Adams County																Oakland County																
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1
-	-	-	1	3	✓	31	30	1	2	2	-	3	0	0	-	-	-	-	1	7	✓	35	33	1	3	3	-	6	0	1	✓	-
-	-	-	1	8	✓	35	32	1	1	3	-	5	0	0	-	-	-	-	1	1	✓	20	19	1	1	3	-	1	0	0	✓	-
-	-	-	1	4	✓	30	24	2	1	3	-	4	0	0	✓	-	-	-	1	11	✓	52	42	3	1	3	-	1	0	0	✓	-
-	-	-	1	5	✓	46	31	2	2	3	-	5	0	0	✓	-	-	-	1	8	✓	42	39	1	1	3	-	8	0	0	✓	-
-	-	-	1	10	✓	50	43	1	2	3	-	9	1	0	✓	-	-	-	1	1	✓	42	40	1	1	3	-	1	0	0	✓	-
-	-	-	1	6	✓	40	39	1	1	3	-	6	0	0	✓	-	-	-	1	6	✓	35	35	1	1	3	-	5	1	0	✓	-
-	-	-	1	1	✓	29	24	1	1	3	-	1	0	0	✓	-	-	-	1	13	✓	42	38	1	1	3	-	7	6	0	✓	-
-	-	-	1	4	✓	39	38	1	1	3	-	4	0	0	✓	-	-	-	1	18	✓	42	36	1	1	3	-	9	9	0	✓	-
-	-	-	1	5	✓	44	42	1	1	3	-	5	0	0	✓	-	-	-	1	9	✓	38	36	1	1	3	-	9	0	0	✓	-
-	-	-	1	5	✓	37	38	1	1	3	-	5	0	0	✓	-	-	-	1	12	✓	30	38	1	1	3	-	1	1	0	✓	-
-	-	-	1	5	✓	44	40	1	1	3	-	5	0	0	✓	-	-	-	1	12	✓	44	39	1	1	3	-	12	0	0	✓	-
-	-	-	1	5	✓	38	29	1	2	3	-	4	1	0	✓	-	-	-	1	1	✓	32	32	1	1	3	-	1	0	0	✓	-
-	-	-	1	2	✓	29	20	2	1	1	-	2	0	0	✓	-	-	-	1	1	✓	37	25	1	1	3	-	1	0	0	✓	-
-	-	-	1	1	✓	32	30	2	1	3	-	1	0	0	✓	-	-	-	1	1	✓	41	24	1	1	2	-	1	✓	✓	✓	-
-	-	-	1	1	✓	24	20	2	1	3	-	1	0	0	✓	-	-	-	1	1	✓	25	20	1	1	3	-	1	✓	✓	✓	-
-	-	-	1	2	✓	24	19	1	2	3	-	1	0	0	✓	-	-	-	1	4	✓	31	30	2	5	2	-	4	0	0	✓	-
-	-	-	1	7	✓	37	31	1	2	3	-	7	0	0	✓	-	-	-	1	3	✓	28	23	2	2	2	-	3	0	0	✓	-
-	-	-	1	2	✓	37	30	1	1	3	-	2	0	0	✓	-	-	-	1	2	✓	27	24	1	1	2	-	2	0	0	✓	-
-	-	-	1	2	✓	58	28	1	2	3	-	2	0	0	-	-	-	-	1	8	✓	44	39	1	1	1	-	6	2	0	✓	-
-	-	-	1	4	✓	31	26	8	2	4	-	4	0	0	✓	-	-	-	1	2	✓	30	25	1	2	3	-	1	1	0	✓	-
-	-	-	1	4	✓	29	33	1	1	3	-	4	0	0	✓	-	-	-	1	4	✓	29	24	1	1	2	-	4	0	0	✓	-
-	-	-	1	5	✓	32	23	1	1	3	-	5	0	0	✓	-	-	-	1	1	✓	46	20	1	1	5	-	1	0	0	✓	-
-	-	-	1	1	✓	28	23	1	1	4	-	1	0	0	✓	-	-	-	1	9	✓	44	40	3	1	2	-	9	0	0	✓	-
-	-	-	1	6	✓	46	42	1	1	3	-	6	0	0	✓	-	-	-	1	1	✓	40	36	1	2	2	-	1	0	0	✓	-
-	-	-	1	9	✓	57	38	1	1	3	-	9	0	0	✓	-	-	-	1	2	✓	21	31	1	1	3	-	2	0	0	✓	-
-	-	-	1	1	✓	30	28	2	2	3	-	1	0	0	✓	-	-	-	1	6	✓	36	30	3	3	2	-	6	0	0	✓	-
-	-	-	1	8	✓	55	44	6	1	1	-	6	2	0	-	-	-	-	1	6	✓	40	29	1	1	2	-	5	1	0	✓	-
-	-	-	1	3	✓	34	20	1	6	3	-	2	1	0	✓	-	-	-	1	9	✓	48	34	8	1	1	-	7	2	0	✓	-
-	-	-	1	6	✓	29	28	1	1	1	-	4	2	0	✓	-	-	-	1	3	✓	33	30	1	1	5	-	3	0	0	✓	-
-	-	-	1	9	✓	42	36	1	1	1	-	7	1	1	✓	-	-	-	1	2	✓	24	24	1	1	4	-	2	0	0	✓	-
-	-	-	1	3	✓	34	33	1	1	2	-	3	0	0	✓	-	-	-	1	1	✓	24	28	5	1	5	-	1	0	0	✓	-

A portion of the work-sheet used in assembling 16 items each from birth certificates. Each number heading a column represents a definite item and information is entered in the form of a symbol or a number. It was possible by this method to enter 500 cases on one sheet. (See tallying key on page 55.)

TALLYING KEY FOR WISCONSIN BIRTH-RECORD STUDY

1. Stillborn	(yes or no)
2. Deformed or defective	(yes or no)
3. Sex	(male or female)
4. Twin, triplet, or other	(1 2, 3, or 4)
5. Number in order of birth	(number)
6. Legitimate	(yes or no)
7. Father's age	(age)
8. Mother's age	(age)
9. Father's birthplace	(Special key 1)
10. Mother's birthplace	(Special key 1)
11. Father's occupation	(Special key 2)
12. Is mother gainfully employed?	(yes or no)
13. Number born alive and now living	(number)
14. Number born alive and now dead	(number)
15. Number stillborn	(number)
16. Is report written in ink?	(yes or no)

UNIVERSITY LIBRARY

SPECIAL KEY NUMBER 1

- 1 = Wisconsin
- 2 = Other States
- 3 = Germany
- 4 = Ireland
- 5 = Great Britain
- 6 = Norway
- 7 = Sweden
- 8 = Poland
- 9 = Italy
- 10 = Russia
- 11 = All others

SPECIAL KEY NUMBER 2

- 1 = Unskilled Man'l.
- 2 = Semiskilled Man'l.
- 3 = Skilled Man'l.
- 4 = Routine Mental
- 5 = Skilled Mental

GENERAL KEY

- ✓ = Yes
- = No
- = Female
- ✓ = Male
- \ = Information Omitted
- x = Information Unknown

which each individual case is entered separately and in which there is no relation between parts aside from that of entering the data from one schedule on one line. The object of whatever method used is to gather the data into such form that they can easily be incorporated into tables where direct and cross relationships can be observed.

THE TALLY SHEET

The next step is the tally sheet. This is a preliminary table in rough form in which the facts are entered by means of scores or checks. The tally sheet having been completed, the scores are counted and entered into the final table. An example of the tally sheet used in the study of

the ages of mothers recorded on birth certificates will be found below, (Table 2).

In an original investigation the data contained in the schedules represent the facts in their primary raw form.

TABLE 2
TALLY SHEET FOR WISCONSIN BIRTH-RECORD STUDY

Mother		Sheboygan	Taylor	Thompson	Verona
15	I	I	I	I	I
16	II	2	I	I	
17	III	3	I	I	III
18	IIII	4	II	5	II II
19	II II II	14	II II	7	II II
20	II II II II II I	26	II I	6	II II
21	II II II II	17	II II II	12	II II I
22	II II II II II III	28	II II	7	II II
23	II II II II II II III	33	II I	6	II II II
24	II II II II II II	22	II I	6	II II
25	II II II II II II II II	32	II III	9	II II
26	II II II II II II II II III	34	II I	6	II II I
27	II II II II II II I	26	II II	10	II II
28	II II II II II II II III	33	II III	9	II II
29	II II II II II	17	II II II	15	II II
30	II II II II II II II	25	II I	6	II II
31	II II II II	13	II	2	II II
32	II II II II I	16	II II	4	II II
33	II II II II II I	21	II II	7	II II
34	II II II II II II	20	I	I	II
35	II II II II II III	18	II II	8	II II
36	II II II II II	12	II III	9	II III
37	II II	5	II II	7	II II
38	II II II	7	II II	7	II II
39	II II II II	12	II II	5	II II
40	II II	4	II	2	II
41	II II I	6	II II	5	I
42	II II	3	I	I	II
43	II II	2			I
44	II II	3			
45					I
46					I
47	I	I			
48					
49					
50					

Data are taken from the work-sheet by means of the tally sheet. From the tally sheet the table is easily constructed. The above section of a tally sheet is the one used in classifying mothers by ages from the preceding work-sheet.

The work-sheet or other method of concentration of material represents the first step toward digestion, the tally sheet represents the second, and the table represents the third.

THE TITLE

The first and most important step in table-making is that of proper identification of the contents. Trouble is often encountered in the use of statistical tables which have been made by others for other purposes. This is due to the fact that these tables do not contain titles of sufficient detail to give all of the essential information necessary before the data can be correlated and used in connection with other data independently gathered. Two dangers must be avoided in the making of these titles. In the first place, there is no excuse for verbosity and the use of flowery language. Titles may be picturesque if purposes of propaganda demand it, but they must be concise and pointed. The second danger is the opposite extreme of too great conciseness. Titles must not be concise at the expense of information. The title is "the sign-post of the table designed to catch the eye of the reader, to suggest noteworthy features and to encourage further perusal." It should not contain information which is already evident in the body of the table itself. "The ideal title is reasonably concise, catchy, unambiguous, not misleading, and adequate." First, *all tables should be numbered*. This facilitates indexing and cross-reference to them. In the second place, the title should indicate *what* the table

TABLE 3
THE CUMULATIVE FREQUENCY DISTRIBUTION*

Class	Simple Frequency	Cumulative Frequencies	
		Down	Up
0— 4.9	2	2	30
5.0— 9.9	5	7	28
10.0—14.9	11	18	23
15.0—19.9	7	25	12
20.0—24.9	5	30	5

* See page 69.

is about. If the data are expressed in terms of units, thousands, tens of thousands, million or billions, the fact should be recorded in the subtitle. Units should, if not self-evident, be properly defined. If, for instance, the word *farm* is used, and only farms over 60 acres in size are included in the classification, this point should be made clear. If, again, we are using the term *age* and are understanding this to mean present age or age at death, this fact also should be made clear. In the third place, the title should make clear *when*, or over what period of time, the data represented were selected. And, finally, in the fourth place, *where*, or the geographical area, represented.

MECHANICAL FORM

Next in importance to the proper designation of the table through title, footnotes, caption, and stubs is that of its mechanical form. When counting and sorting have been accomplished, the next step is the arrangement of the data in tabular form, in columns and rows of figures. Each

TABLE 4
FREQUENCY TABLE*

Size Class	Frequency
0—4.9	1
5.0—9.9	3
10.0—14.9	5
15.0—19.9	9
20.0—24.9	5
25.0—29.9	3
30.0—34.9	1

* See page 70.

column is properly identified by means of a *caption*, and each *row* by means of a *stub heading*. While there is a general technique to be followed, each table represents an individual problem to be solved. Tables are composed of words, lines, and

figures. An important part of the problem of tabular presentation is that of attaining proper perspective. Through the use of different styles and sizes of type, it is possible to make the various descriptive elements of the tables stand out in their proper relationship. The main portion of the title should be in type of such sort that it will at once attract attention to itself as the most important part.

Box and stub-headings should naturally be in smaller type but sufficiently large to be easily read. Where tables are printed, the printer has at his disposal a wide variety of type. Where the printing is done by hand, care should be taken so to vary the lettering as to attain this objective. In

Table 5

**THE TOTAL POPULATION OF
MADISON, WISCONSIN**

1920 and 1925 *

Year	Population
Total	
1920	
1925	

* (Source)

the past the statistician has been required to be a draftsman as well, but today, through the use of the typewriter, quite satisfactory results may be attained. To be sure, the variety of type available is not so great, but with the judicious use of capitals, combinations of capitals and small letters, and small letters alone, good tables can be made. Through the use of a typewriter having removable or changeable type shuttles, quite remarkable results are possible. With regard to the ruling of the table, it is possible to embody heavy lines, light lines, double lines, and red lines. The object in using lines of different sorts is to make the different elements of the tables stand out in various perspectives.

Naturally we want to set off the title of the table from the rest of the table by either a heavy or a double line. Again,

NOTE: Tables 5, 6, 7, 8, and 9 are one unit attempting to show the evolution of the statistical table from the simple one or single cross-classification to the complete four or quadruple cross-classification table. (See pages 70-71.)

Table 6

DISTRIBUTION OF THE TOTAL POPULATION

By Year and Sex

Madison, Wisconsin

1920 and 1925 *

Year	Sex		
	Total	Male	Female
Total			
1920			
1925			

* (Source)

total columns or total rows may be so designated, and between the various parts of multiple classification tables a heavier line may be used than within the classification itself. No definite rules can be laid down as to just what sorts of type shall be used or what combination of lines. Each table represents an individual problem and that problem is to set off the material contained in the table to the best advantage so that those portions of the table which are the most important will naturally stand out. The statistician will easily develop a technique of his own and will also observe the technique employed in such statistical studies as those of the United States Bureau of Census and other statistical organizations.

In form there are two general types of table, the open and

Table 7

DISTRIBUTION OF THE TOTAL POPULATION

By Year, Sex, and Income

Madison, Wisconsin

1920 and 1925

Income and Year		Sex		
		T	M	F
Total	Total			
	1920			
	1925			
Under \$2000.	Total			
	1920			
	1925			
\$2000. and over	Total			
	1920			
	1925			

* (Source)

T = Total M = Male F = Female

Table 8

DISTRIBUTION OF THE TOTAL POPULATION

By Year, Sex, and Nativity

Madison, Wisconsin

1920 and 1925 *

Year	Sex and Nativity								
	Total			Native Born			Foreign Born		
	T	M	F	T	M	F	T	M	F
Total									
1920									
1925									

* (Source) T = Total M = Male F = Female

the closed variety. The closed table has a single or double line completely about it with the title either within or without the boundary. This sort of table is more attractive than the open table which has no lines on the sides, but it is doubtful whether or not this additional expenditure of time and energy in making the table is worth while except for publicity purposes. In the United States Census and in most commercial reports tables are of the open type.

LOCATION OF TOTALS

Another matter of importance to be noted is that of the location of totals. In adding up a column of figures the total is usually placed at the bottom of the column, but in a statistical table rows of totals should be at the top of the table and columns of totals at the left. This is for the pur-

Table 9

DISTRIBUTION OF THE TOTAL POPULATION

By Year, Sex, Income, and Nativity

Madison, Wisconsin

1920 and 1925 *

Income and Year		Sex and Nativity								
		Total			Native Born			Foreign Born		
		T	M	F	T	M	F	T	M	F
Total	Total									
	1920									
	1925									
Under \$2000.	Total									
	1920									
	1925									
\$2000. and over	Total									
	1920									
	1925									

* (Source)

T = Total

M = Male

F = Female

pose of giving them the maximum emphasis in the table. This is a rule which should invariably be followed and which will always be found in good statistical presentation.

LOCATION OF TABLE

Since the table is in a way a work of art, the size of the table should be suited to the space which it is to occupy. Whether it be a single table on a page, a group of tables on a page, or a single table set in the contents of the page, the

width of the columns and rows should be so adjusted that it will look as though it belongs there. Secrist well sums up the important points in the art of table-making in the following words: "In the clearest way possible, project into a table form the message which is wished to be conveyed. Make the totals prominent by placing them first and show the relative importance of the classification by the ruling and spacing of major and minor headings. Make tables attractive, concise, and emphatic. Set them approximately to standard sized pages . . . and above all give suitable brief but explicit titles."

TYPES OF CLASSIFICATION

Finally, tables may be classified into three groups, *simple*, *partial cross-classification*, and *complete cross-classification*. *Simple tables* are those in which numerical data are given for but one classification. For example, the death-rate for various age groups, the numbers employed in various occupations, or the per capita wealth of different countries. A *partial cross-classification table* is one in which two or more sets of numerical data are given for the same classification but without attempt to relate them to each other. A partial cross-classification table is in reality a combination of two or more simple tables. For instance, we might desire to show for the various age groups in successive columns, first, morbidity, and second, mortality rates. Or we might, for the years 1920 and 1925, divide the population of a given city into male and female, native-born and foreign-born, and those having incomes of less than \$2,000 and those having incomes of \$2,000 or more. It would be entirely possible to incorporate this information into one table without showing how many males were native-born and had incomes of over \$2,000, or how many females were foreign-born and had incomes of less than \$2,000. If, on the other hand, we were to attempt to make this classification and carry it out

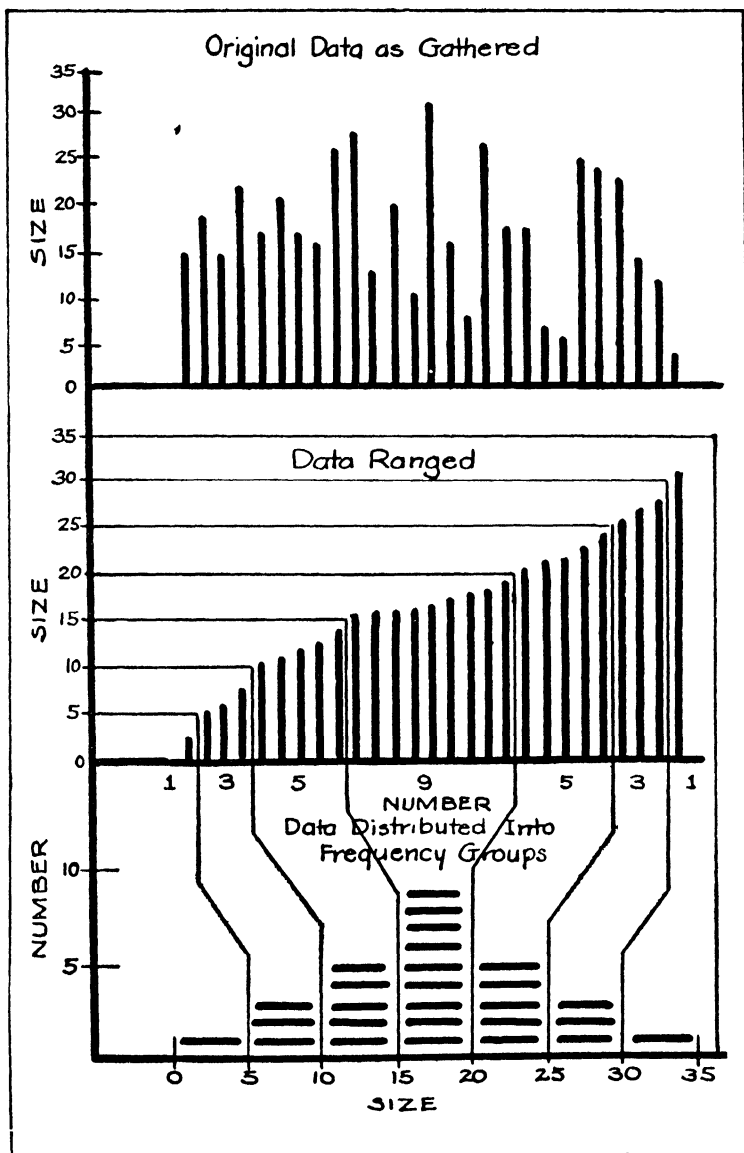


Chart I: Ranging and frequency distribution. (See page 69.)

to its farthest detail so that every item could be read in terms of every other item, then we would have accomplished a *complete cross-classification*. It is possible, of course, to make complete cross-classification tables composed of any number of elements, but experience has shown that four is about the maximum size practicable. (See Table 9, page 63.)

In the above suggested quadruple complete cross-classification we have this type of table reduced to its minimum size, since each of the classifications is composed of but two elements. Adding to this a third column, or a row representing a total for each, we will find ourselves with a table composed of nine rows and nine columns, exclusive of stubs and captions and titles. If, now, we are to add a fifth classification, say marital condition, dividing it into married and unmarried with a row for a total, we will then have increased our table just three times in size. Another of a similar sort would increase it another three. It is observed that, as we add classifications, the size and complexity of the table very rapidly mounts. Seldom, indeed, will a four cross-classification table be used, not to say one which is larger. Since also most tables are for specific purposes, general complete cross-classifications of this sort are usually unnecessary. They are desirable for general purposes of research, but for the specific purpose in hand they are usually superfluous. Complicated tables should on general principles be avoided.

BASES OF CLASSIFICATION

There are four general bases of classification in operation. First, *chronological* classification, in which the time element is basic. For instance, the growth of the population by decades or immigration into the United States by years. Second, *geographic*, of which birth- or death-rates by states or countries would be an example. Third, *qualitative*, illus-

trated by such classifications as sex, marital condition, and employment. And finally *quantitative*, or division according to some numerical attribute. An example of this would be classifying men according to salary and dividing them into salary groups of \$500 or \$1,000 each.

CLASS LIMITS

In any of the above four types of classification the main problem is that of the proper definition of class limits. Classes must be mutually exclusive if valid statistical results are to be obtained. Chronological classifications by days, months, years, decades, or centuries are perhaps the most simple. But even here, in dealing with the year as the unit, for example, we must know what point was taken as the beginning of the year. It need not be from January 1 to December 31, we may be speaking of the school year which might be from September 1 till June 30. In geographical classifications such as Wisconsin or Minnesota there is little possibility of confusion, but just what do we mean by such a classification as "The United States"? Are we including or excluding our island possessions and what about territories? Would the word Milwaukee refer to the city or to the county of which the city forms the major portion? In qualitative classifications male and female are absolute and mutually exclusive classifications, but what about married and unmarried? Where will you place the widow or divorcee in this classification?

In these three types of classification it is evident that our problem is that of the proper definition of the units which designate these classifications. There is no specific rule to follow with regard to the manner in which the above widow or divorcee should be classified. The important point is that the statistical study should show definitely how she was placed or, perhaps, for the purpose of the particular study, that those widowed or divorced were not included at all.

LIMITS OF QUANTITATIVE CLASSIFICATIONS

The problem of setting the class limits of quantitative classifications is more a matter of the standardization of technique than of the definition of units. When we say that salaries are between one and two thousand, between two and three thousand, and between three and four thousand dollars each it is difficult to know just where the line of demarcation between classes takes place. And yet the majority of tables fed to the public from semiscientific sources through a myriad of publications commit this error. Here it is that much statistical inaccuracy has rendered many interesting investigations valueless as secondary sources for further study.

Class limits in frequency distributions should always be so clearly stated that misunderstanding cannot result. Instead of dividing classes into groups of from 0 to 5, 5 to 10, 10 to 15, and 15 to 20, it is more desirable to form the practice of stating them as from 0 to 4.9, 5.0 to 9.9, 10.0 to 14.9, with the understanding that each class includes everything up to, but not including, the absolute beginning of the next class. As a general practice there is no cogent reason for breaking this general rule which results in clear definition and reasonable certainty as to what is to be included within a class. A table thus composed results in what we may call a frequency distribution. This frequency is usually composed of classes of equal size or interval, the limits of which, as above suggested, are to be clearly indicated.

Unequal class intervals are in general to be avoided, though there are occasions when they may be allowed. For instance, the United States Census uses the following age classification: (1) under 5 years of age; (2) under 1 year; (3) 5 to 9 years; (4) 10 to 14 years; (5) 15 to 19 years; (6) 20 to 44 years, and (7) 45 years and over. Again, in the study of incomes we would be justified in using a smaller class interval for the smaller incomes than for the larger

incomes. Up to a limit of ten thousand dollars, incomes might be classified in thousand-dollar groups. Between ten and one hundred thousand dollars they might be classified in ten-thousand-dollar groups and beyond that point in one-hundred-thousand-dollar groups. While this is an entirely possible procedure, there is no justification for it where equal class intervals will work as well. In general, equal class intervals should, if possible, be employed.

Another type of table, the cumulative frequency table, for certain purposes is quite useful. In the ordinary frequency table the frequency for each class is given individually, but in the cumulative frequency table the frequencies are accumulated until the last class represents the total for the whole. Cumulation may proceed either up or down, and the frequency designated opposite any one class will represent the total number of frequencies up to, or down to, and including that particular class (Table 3, page 57).

Chart 1 visualizes the process whereby the individual items in a random sample are, first, arranged in order according to size—this is called ranging—and second, classified into size groups, which is called frequency distribution. In the middle portion of Chart 1 (data ranged) we have divided the vertical axis, which represents the magnitude of the items, into equal divisions of five units each. It will be noticed that in the first group, which extends from 0 to, but not including, 5 (0-4.9), we have one case, while in the middle section, 15.0-19.9, we have nine cases. In this middle section the vertical axis directly indicates the length of each individual bar. The next step reverses the arrangement of the axes, showing not the size of each individual item, but the size class into which it falls, along the base line or the horizontal axis. The vertical axis now merely indicates how many of the items fall within each class. We may think of this third portion of the chart as representing the bars or items no longer standing on end and ranged but the bars of a given size class lying on their sides and piled.

The third portion of the chart can now be readily converted into the form of a frequency table as is illustrated in Table 4 on page 58.

COMPLETE CROSS-CLASSIFICATION GENERAL-PURPOSE TABLES

Tables 5 to 9, on pages 59 to 63, represent the process whereby the complete four cross-classification is developed from the simple one cross-classification table. This is an illustration of what can be done rather than what will usually be done. Since most tables are for the special purpose of some particular investigation, they will tend to be partial or incomplete cross-classifications. Attention is again called to the fact that these tables have purposely been reduced to their simplest form. Each classification is composed of a total and of two elements; for instance, total, male and female. Of course, it is rather unusual to divide incomes into just two classes with \$2,000 as the dividing line, but it is interesting to note what would happen to the size of our table if we had incomes more logically divided into a dozen or fifteen groups. The number of items in the various classifications must always be taken into consideration in determining the number of cross-classifications to be attempted.

The first two tables occupy relatively more space than the others, due to the fact that more space is available. In the same manner, the third table might be made wider, but there would be no object in increasing the length of the fourth.

These tables are purposely typed to demonstrate what can be done by any stenographer or clerk who is in the possession of a bit of patience and ingenuity. Their main lack is the absence of any variation in the type of line used.

Attention should also be called to the fact that this complete four cross-classification table represents the four types of classification: chronological, qualitative, quantitative, and geographical.

SPECIAL-PURPOSE CROSS-CLASSIFICATION TABLES

Tables 10, 11, 12, and 13 are taken from the National Tuberculosis Association's pamphlet, *Tuberculosis Case-finding Surveys*, which was prepared by their statistician, Miss Jessamine S. Whitney. They are special-purpose tables and represent both complete and partial cross-classification. They further illustrate the type of work which may be expected from the printer.

In Table 10, note the unequal interval of the age groups due to the requirements of the particular problem. This is a double complete cross-classification table.

Table 11 is also a complete cross-classification of three elements. Note the equal class interval of the age grouping and the designation of the class limits.

Tables 12 and 13 are partial cross-classifications. Note the manner in which the first column of Table 12 accommodates the three classifications: (a) white and negro, (b) resident and non-resident, and (c) dead, alive, arrested, removed. This arrangement takes less horizontal space in the table than the method used in the preceding typewritten tables but is less clear. It is, nevertheless, by far the more common method.

If the titles of these tables seem not entirely clear and self-explanatory, it is because they appear in the body of a survey which furnishes this additional information.

MECHANICAL TABULATION

In mechanical tabulation, facts are entered directly from the schedule or other source upon the punch card which is illustrated in Form 2. The amount of information which can be recorded upon a punch card is definitely limited by its 45 columns and the type of data which are being recorded. Qualitative data such as cause of death, occupation, nationality, and so forth, can be coded so as to require only one or two columns. Quantitative data can also be coded,

[illegible]

Unadapted Punch Card

[illegible]

Adapted Punch Card Used at the University of Wisconsin

Form 2: An example of the adaptation of the punch cards to the needs of a university registrar.

[illegible]

Form 3: The adaptation of the punch card to a specific need.

TABLE 10
CONJUGAL CONDITION

CONJUGAL CONDITION	AGE GROUPS			
	Total	0-19 Years	20-24 Years	25 Years and Over
All persons.....				
Single.....				
Married.....				
Widowed.....				
Divorced.....				
Separated.....				

provided we are dealing only in round numbers or frequency classes, but usually they are not and require as many columns on the punch card as there are places in the largest figure to be recorded. Names, addresses, and dates may be written at the top of the card or may be coded and punched. The University of Wisconsin codes the name of each of its 8,000 students for the purpose of mechanically returning them to alphabetical order.

The economical adjustment of a group of items to the punch card is a task which requires careful thought. Sometimes it will be possible to enter all the data on a schedule, but more often it requires several cards or a selection of data for one card.

Punching is done by means of a mechanical device which works with numbered keys and automatically advances the card one column each time. The mechanical process of punching the cards and checking them for errors requires time, but once they are made we have a permanent record which can be put through the tabulating machines any number of times.

Tabulating machines are highly complicated electrical devices which sort the cards and operate adding machine units by means of electrical contacts which are made through the punched holes. They are able to sort and tally the cards for as many classifications as are desired and give a half-dozen or more totals at one operation. A machine of this

TABLE 11
AGE AND SEX OF TUBERCULOUS PERSONS, TOGETHER WITH THEIR
STATUS ON (*Date Survey Year Ended*)

AGE GROUPS	ALL PERSONS			STATUS														
				Dead			Active			Arrested			Removed					
All Ages.....																		
Under 5 years.....																		
5- 9 years.....																		
10-14 years.....																		
15-19 years.....																		
20-24 years.....																		
25-29 years.....																		
30-34 years.....																		
35-39 years.....																		
40-44 years.....																		
45-49 years.....																		
50-54 years.....																		
55-59 years.....																		
60-64 years.....																		
65-69 years.....																		
70-74 years.....																		
75 years and over.....																		

An example of a triple complete cross-classification table.

TABLE 12

EXTENT TO WHICH PERSONS COMING IN CONTACT WITH TUBERCULOUS PATIENTS WERE EXAMINED, ACCORDING TO STATUS OF LATTER WHEN LAST KNOWN

STATUS OF TUBERCULOUS PATIENTS WHEN LAST KNOWN	PERSONS* COMING IN CONTACT WITH TUBERCULOUS PATIENTS			
	Children (under sixteen years)		Adults	
	In House-holds	Examined	In House-holds	Examined
All patients				
White				
Residents				
Dead				
Active				
Arrested				
Removed				
Non-residents				
Dead				
Active				
Arrested				
Removed				
Negro				
Residents				
Dead				
Active				
Arrested				
Removed				
Non-residents				
Dead				
Active				
Arrested				
Removed				

*When a person came in contact with more than one patient, he has been enumerated only once and then as a contact of the patient whose condition was the most serious.

An example of partial cross-classification. Compare the arrangement of the three classifications in the column at the left with that of the two classifications in Table 9, page 63.

sort will, once the cards are punched, do as much work in an hour as a clerk could do by hand in a week. This is the method which has made the modern census possible. It is used by large business concerns to great advantage and even great universities are adapting it to their needs. Form 3 shows how it might be adapted to a specific problem.

TABLE 13
TUBERCULOUS PERSONS WITH SPECIFIED SANATORIUM OR HOSPITAL EXPERIENCE
ACCORDING TO RESIDENCE AND STATUS

RESIDENCE AND STATUS	PERSONS WITH SPECIFIED SANATORIUM OR HOSPITAL EXPERIENCE						
	TOTAL	NO KNOWN EXPERIENCE	KNOWN EXPERIENCE				Duration Unknown
			Total	Less than Six Weeks	Six Weeks and Less than Six Months	Six Months and Over	
All persons.....							
Residents.....							
Dead.....							
Active.....							
Arrested.....							
Removed.....							
Non-residents.....							
Dead.....							
Active.....							
Arrested.....							
Removed.....							

Special-purpose, partial cross-classification table. Compare with Tables 9 and 12, pages 63 and 75.

VI

GRAPHIC PRESENTATION¹

The purpose of graphic presentation. The pictogram or statistical picture. The limits of pictogrammatic presentation. The cartogram or statistical map. The circle and sector or "pie" chart. The bar chart. The line graph. Criteria and standards of good graphic presentation.

GRAPHIC presentation is the problem of visualizing tabulated or tabulable data. It exists both for the purpose of analysis and propaganda. No type of statistical presentation has been more widely applied and none has lent itself more fully to misrepresentation in the hands of the propagandist, or vague half-truths in the hands of the tyro. King remarks that "figures, at best, are not easy things for the mind to hold long enough for the purpose of comparison. When read to an audience they become practically meaningless. In a book, their essential indications can only be ascertained by careful scrutiny. One of the chief aims of statistical science is to render the meaning of masses of figures instantly clear and comprehensible at a glance." To tell a listening audience the numerical loss of life suffered by the various nations in the Great War, or present the same numerical data in tabular form to a reading public requires considerable mental effort on the part of listener and reader;

¹ With the exception of Charts 4 and 18, the charts used in this chapter have been selected very much at random as actual examples of current statistical practice. They represent good, bad, and indifferent statistical presentation. The reader will profit most by sizing them up in the light of the criteria laid down. The writer is indebted to The American Appraisal Company for Charts 2, 7, 8, 19 and 20; to the University of Wisconsin Agricultural Experiment Station for Charts 3 and 6; to *University of Nebraska Bulletin No. 180* for Charts 5 and 10; to the *Quarterly Journal of Economics* for Chart 9; to the Chamber of Commerce of the United States for Charts 11, 13, 14, 15, 16, and 17; to the United States Department of Labor, Children's Bureau, for Chart 12; to the Wisconsin State Board of Health for Chart 21; and to the United States Department of Agriculture for Chart 22.

first, in comprehending the figures, and second, in making the comparison. In the case of the audience there is no possibility that they will be capable of retaining the figures in mind long enough for this purpose or that they would be capable of this purely mental comparison if they did remember the figures. The human mind, it must be remembered, at best is lazy and refuses to operate unless adequately stimulated. A suggestion here to public speakers is in order. It is useless to rattle off figures of numerous integers. The mind refuses to grasp them and they become confused. "One billion, five hundred sixty-eight million, one hundred and ten thousand, two hundred sixty-four" may be the exact truth, but to your audience "a bit over one and one-half billion" will convey the idea much more surely and make comparison possible. Graphic presentation, then, rests on the psychology of "fixation of attention." It meets also the psychological problem of illusion, and the statistician must be careful to avoid methods which seem to show what is not there or which confuse the eye. Methods of graphic presentation should not be used independently of the data concerned, since they are merely illustrative or for the purposes of visualization. They should either be used in conjunction with the tables which they illustrate or contain within themselves the data.

Care must always be taken in selecting the most appropriate method of graphic presentation. The criterion may be scientific, or it may be purely popular, the question being, what will best appeal to the particular group with its particular group background?

THE PICTOGRAM

The most commonly used method, in the past at least, is that of the *pictogram* or *statistical picture*. We probably first became familiar with this method through the agency of our old geographies which, in the back, compared the

status of countries in the matter of production of wheat, cotton, and other commodities through the agency of pictures of bales of cotton of various sizes, bags of wheat, piles of coal, big men and little men, and squares, cubes, and circles of various dimensions. You will find them today in any copy of *Literary Digest* or other magazines which popularize statistical facts. There is no question that they make a psychological appeal far stronger than that of the most sci-

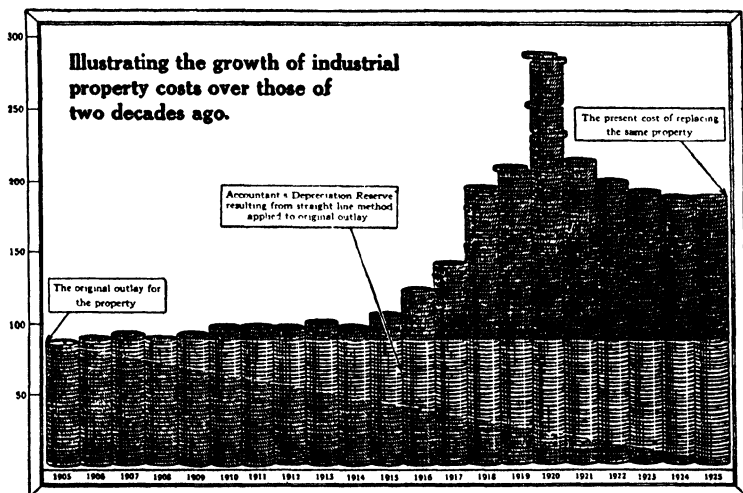


Chart 2: A pictogrammatic bar chart.

entifically constructed charts and graphs. They capture the attention, direct it to the data, when the data happens to be present, and fix upon the mind an image. The trouble is that this image is apt to be capable of various interpretations which are usually erroneous. The reader will have in mind that most common of all pictograms which illustrates the sizes of various standing armies of the world. Here we have in order first the giant soldier and then soldiers of various sizes down to the pigmy which represents the standing army of the nation with the smallest number. What is the comparison which is expected of the reader? Is he to compare

height, or area, or volume? The smallest soldier may be one-tenth the height of the largest soldier but in area and volume he is very much smaller. Pictograms of this sort seldom indicate which comparison has been used, and even where the figures accompany the pictogram it is difficult to discover which was meant. About the only idea which the reader can get is that either a large or a small amount of difference is present. It is of value only in capturing the attention, and then only in case he studies the data and corrects the optical illusion which is sure to be present. It may be stated with the force of a law that it is bad statistical practice but not bad propaganda to employ methods of graphic presentation which use more than one dimension for purposes of comparison. Pictograms should never be used except in conjunction with, or for the purpose of, the modification of legitimate methods of graphic presentation.

The writer has in mind a statistical picture which appeared some time ago in a popular magazine and which so impressed him that he made note of it for future reference. It was labeled "Relativity." Four figures were represented. First came the pugilist with his determined face, bull neck, broad shoulders, brawny arms, and the information that he earns \$750,000 per year. Second, the movie star, a Jackie Coogan, who came well above the pugilist's shoulder and was labeled \$500,000. Jackie's head was the larger of the two and he was almost as tall, but, even with arms akimbo he did not compare with the fighter in breadth. Third in line was the baseball player who came to Jackie's elbow and was labeled \$100,000. And last in the list, the college professor who stood as high as the baseball player's knee and was labeled \$3,500.

Now we will admit that the picture was striking and certainly told a story. At least college professors would be apt to remember it. But it was a picture in which the numerical values bore very little relation to pictures by which they were represented. It is interesting to note that the pro-

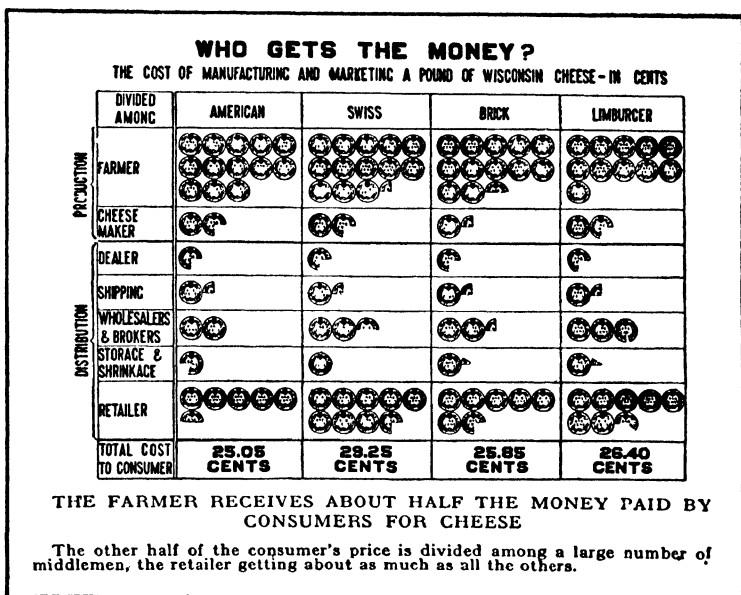


Chart 3: A pictogrammatic table the advantages of which are doubtful.

TABLE 14
WHO GETS THE MONEY?
The Cost of Manufacturing and Marketing a Pound of Wisconsin Cheese—in Cents

Divided Among	American	Swiss	Brick	Limburger
Production—				
Farmer.....	13.00	13.25	12.50	11.00
Cheese Maker.....	1.75	1.75	1.25	1.75
Distribution—				
Dealer.....	.75	.75	.75	.75
Shipping.....	1.25	1.25	1.25	1.25
Wholesaler & Broker	2.00	2.50	2.25	2.95
Storage & Shrinkage	.80	1.00	1.10	1.05
Retailer.....	5.50	8.75	6.75	7.65
Total Cost to Consumer.	25.05	29.25	25.85	26.40

*The above table is constructed on the basis of the totals and the pictogrammatic information contained in Chart 3. What are the advantages of each?

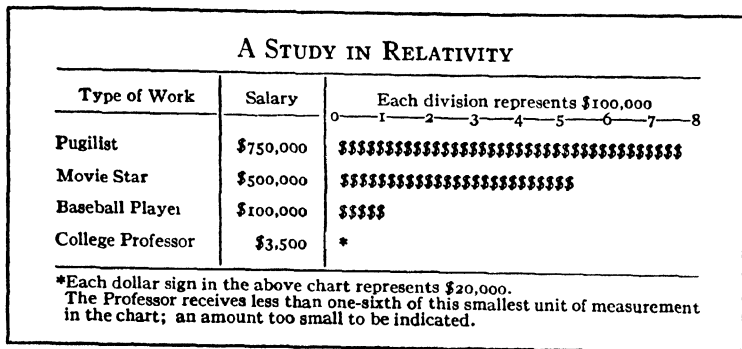


Chart 4: A suggestive bar chart which carries a message.

fessor was as large as the baseball player's shin, or the movie star's forearm, or the pugilist's foot—but merely interesting. Chart 4 shows the manner in which the statistician might have attempted to popularize the same information.

THE CARTOGRAM

Cartograms or *statistical maps* partake somewhat of the nature of the pictogram in that they present a familiar geographical area in map form and show the distribution of the data concerned. Cartograms most commonly have to do with the distribution of density or frequency. Methods are employed which range all the way from definitely differentiated cross-hatching to mere shading. In dot maps one dot may represent a single unit or any designated number of units. Where the dots are too frequent to be easily counted, the result is a mere density of shading which gives but a vague impression of greater or lesser amount. Where one dot represents a large number of units, for example, one hundred, one thousand, or one million, it is a common practice to use half and quarter dots in conjunction with full dots. This makes it possible to gain the impression of density and also to have a small enough number of dots so that they can be counted. This, however, is a cumbersome

method and seldom, indeed, will the interpreter of the cartogram put himself to the trouble of counting the dots, quarter dots, and half dots, and estimating the total. As a general rule, definitely designated cross-hatchings are preferable in the making of cartograms.

In the choice of different types of cross-hatching two

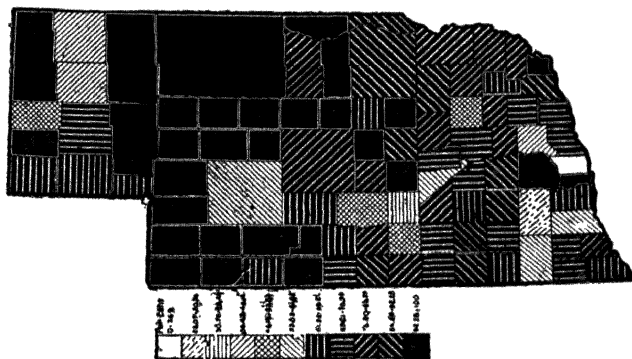


FIG. 3.—NEBRASKA'S UNLIBRARIAN POPULATION

Less than two-fifths have access to public library facilities. The sparser the population the greater its dependence on home-owned reading matter.

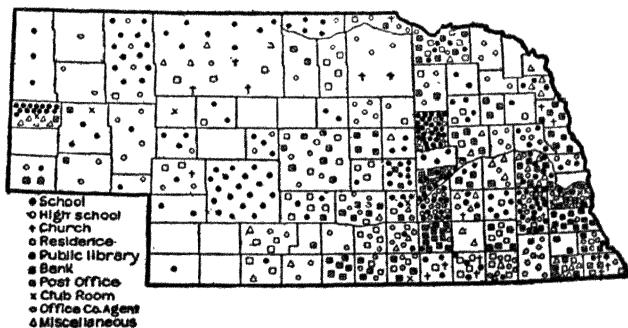


FIG. 4.—WHERE THE TRAVELING LIBRARIES WENT

Each symbol shows where the Nebraska Public Library Commission sent a traveling library during the two years ended November 30, 1920. A large share went into library stations accessible to farmers.

rules may be followed. In the first place each type of cross-hatching should be definitely recognizable and capable of differentiation from all others. In the second place, the greatest density or the highest degree of frequency should be represented by the most definite or blackest cross-hatching while the lightest density or smallest degree of frequency should be represented by the lightest cross-hatching or perhaps no cross-hatching at all.

A variation of the cartogram is the *pin map* in which glass-headed pins of different colors take the place of dots. This is purely an office map for administrative or propaganda purposes where it is desirable to keep pace with fluctuating conditions. Every cartogram should contain an easily understandable key and where possible a table containing the data on which it is based. Undoubtedly the most unscientific and least desirable method of cartogramatic presentation is that in which circles or dots of different sizes are used. Obviously, it is impossible to compare the area of circles with any degree of exactness.

THE CIRCLE AND SECTOR

Another widely used and little justified method of graphic presentation is that of the *circle and sector chart*. This is used in the main where it is desirable to show a portion of the whole in such studies as "where the dollar goes." About the only advantage of this method is that the circle may be made to represent and visualize the dollar. By dividing the circle into sectors and cross-hatching these sectors into different designs it is possible to indicate relationships, but not so exactly as good statistical presentation demands. It is possible by the means of concentric circles of different sizes to make more than one comparison, but usually these result in such complicated diagrams that they are very difficult to interpret. Indeed we may say of the circle and sector diagram that it is capable of portraying little which cannot be

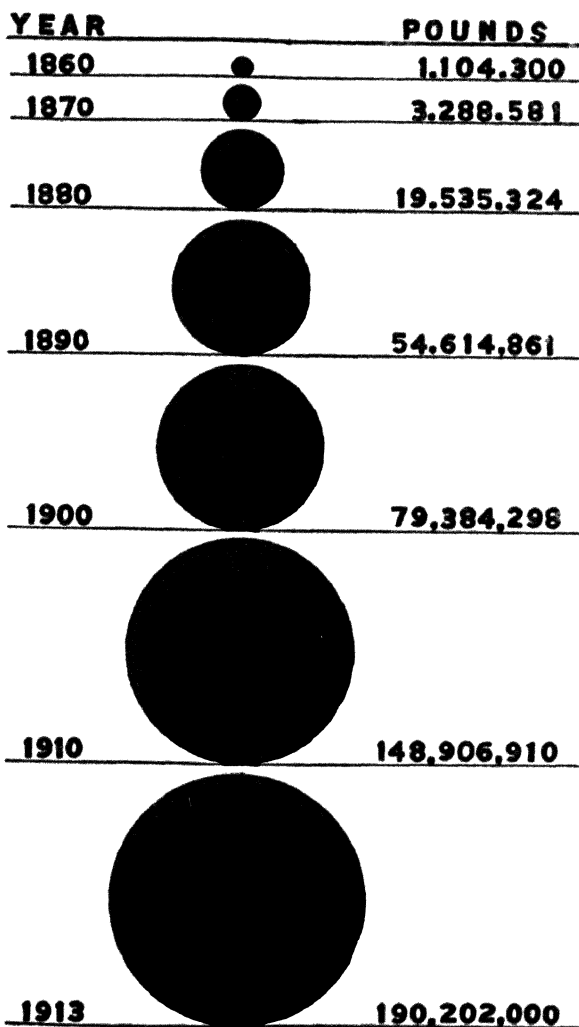


FIG. 1.—HOW THE WISCONSIN CHEESE INDUSTRY HAS GROWN
 Figures for 1860 to 1910 from the Federal Census; for 1913 from estimates
 made by the Wisconsin State Dairy and Food Commissioner.

Chart 6: These dots show nothing that could not have been portrayed to far better advantage by bars.

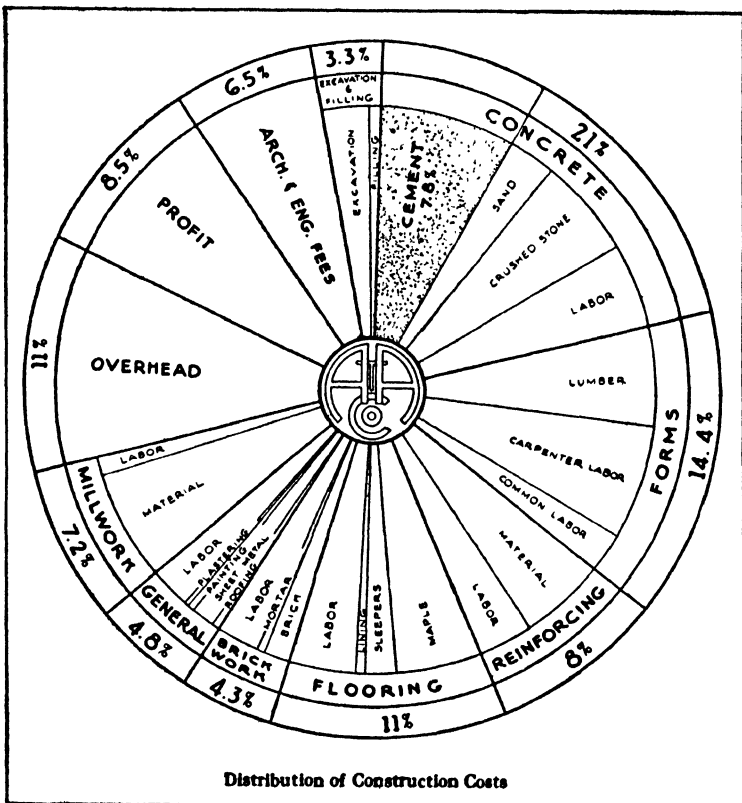


Chart 7: This chart represents a subdivided circle and sect or diagram as effectively done as is possible with this type of chart. Made in the form of a horizontal bar chart it would have been capable of readier comparison but still complicated.

presented to far greater advantage by the next method which we are to discuss. The main advantage of the circle and sector chart lies in its slightly more popular appeal.

THE BAR CHART

The *bar chart* visualizes by means of bars of varying length but of the same width, either horizontally or verti-

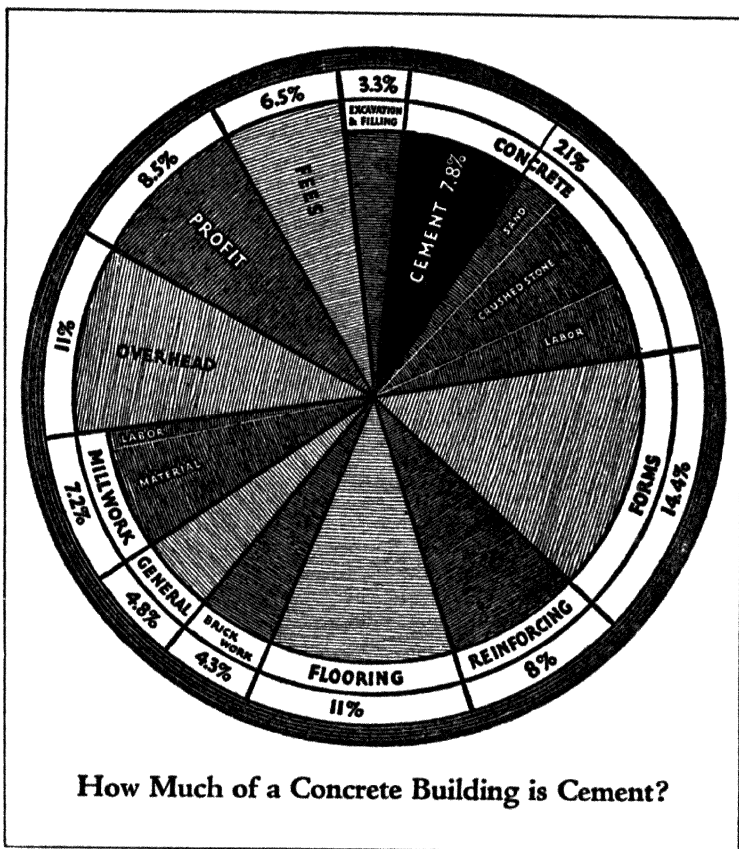
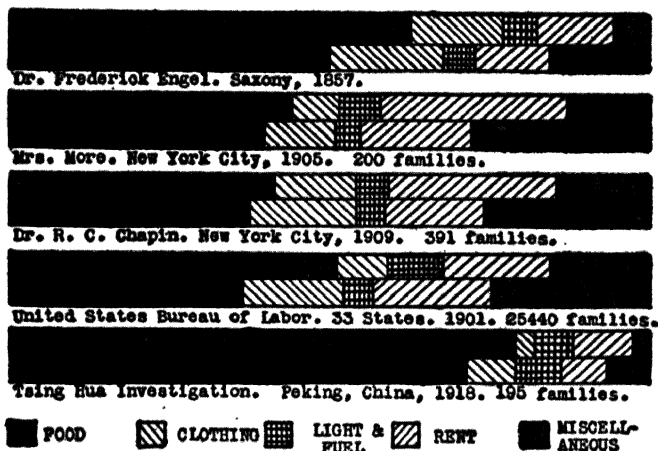


Chart 8: An example of the advantageous use of cross-hatch shading.

cally arranged, the magnitude of the data in question. In this chart we have the ideal method of presentation of most kinds of statistical data. Comparison is made by means of but one dimension, that of the length of the bar. Comparison is aided by a scale extending along the length of the first bar, either horizontally or vertically, as the case may be and by scale lines which run through as many of the bars as is necessary. Again, we may point out that there is little in graphical presentation which cannot be accomplished by its

A COMPARISON OF THE EXTREMES OF VARIATION IN THE PERCENTAGE OF EXPENDITURE FOR VARIOUS PURPOSES BETWEEN THE HIGHEST AND THE LOWEST INCOME GROUPS IN FIVE INVESTIGATIONS¹



THE EXPENDITURES IN FIVE INVESTIGATIONS

Investigation	Income Groups A — Lowest B — Highest	Per cent of Total Expenditure				
		Food	Clothing	Light and Fuel	Rent	Miscellaneous
Engel A	\$225- 300	62.0	16.0	5.0	12.0	5.0
..... B	750-1000	50.0	18.0	5.0	12.0	15.0
Mrs. More A	200- 400	44.2	7.3	6.5	30.5	11.5
..... B	1200-1500	39.5	11.3	4.5	18.0	26.7
Dr. Chapin A	400- 499	40.8	13.0	5.6	26.8	14.8
..... B	1500-1599	36.8	16.8	4.1	16.3	26.0
U. S. B. of L. A	Under 200	50.9	8.7	7.9	16.9	15.6
..... B	1200 Over	36.5	15.7	5.0	17.4	25.4
Tsing Hua A	30- 49(a)	79.0	3.4	6.0	9.9	1.3
..... B	130- 149	70.8	8.5	7.0	7.1	6.6

(a) Chinese dollars. Exchange value fluctuates greatly but is normally about \$.50.

Chart 9: A startling comparison made possible by the use of the horizontal bar chart.

proper use. Even the *line graph* which we are to discuss next is but a refinement of this same bar chart. The only advantage, in addition to that of visual appeal, that the circle and sector diagram has over the bar chart is that the circumference of the circle represents the total, whereas in a bar chart, especially if it is composed of a large number of bars, it is impossible to use a bar to indicate the total without making it out of all proportion to the remaining bars which follow.

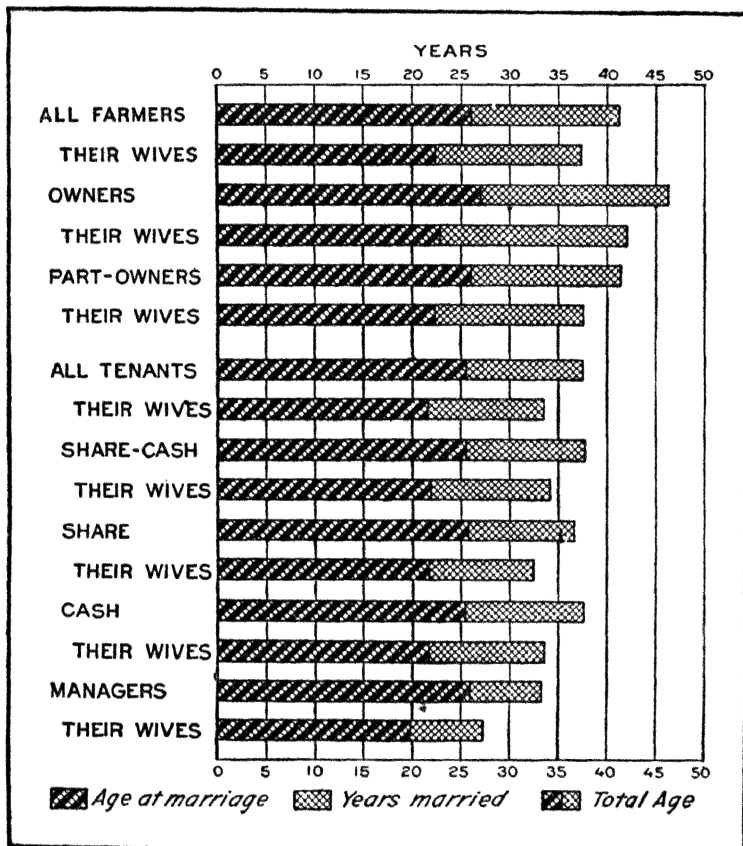


Chart 10: Horizontal bar chart showing age at marriage and at present of Nebraska farmers. Note the scale designations at both the top and bottom of the chart.

The bar chart shows the relation of the parts to each other better than it does the relation of the parts to the whole.

Another great advantage of the bar chart is the ease with which it adapts itself to pictogrammatic presentation. We have referred earlier in the chapter to the pictogram using soldiers of different sizes to represent the fighting strength of the different nations. Applied to the bar chart method, instead of having large soldiers and small soldiers, we would have in the place of the bars lines of soldiers of exactly the same size but of different number, each in his characteristic native costume. This will attract the attention quite as easily as the other method and will leave no question in the mind of the reader as to what is the element of comparison. Where this is impossible, the body of the chart itself may contain a picture or a diagram which will draw attention to the bars below. Such illustrative treatment, however, will be used only where necessary for propaganda purposes. The statistician does not need to have his charts thus decorated in order to understand them.

A pleasing variation of the bar chart is the bar chart and statistical table combined. The table is made in its usual form with the exception of a column at the right of sufficient width to contain bars representing the various classification at the left. In this manner the table is visualized and all of the advantage of both table and bar chart are obtained. For simple comparisons of size the horizontal bar chart is usually used, while for frequency distributions the vertical bar chart is the almost universal practice, except where the charts and frequency table are combined. For chronological series the vertical bar chart is the most commonly used, except when the bars are so few in number that a horizontal arrangement would look better.

Bar charts have the further advantage of lending themselves to subdivision and cross-hatching. If, for instance, we wish to show the change in proportion of urban to rural residents in the population by decades, it is possible for the

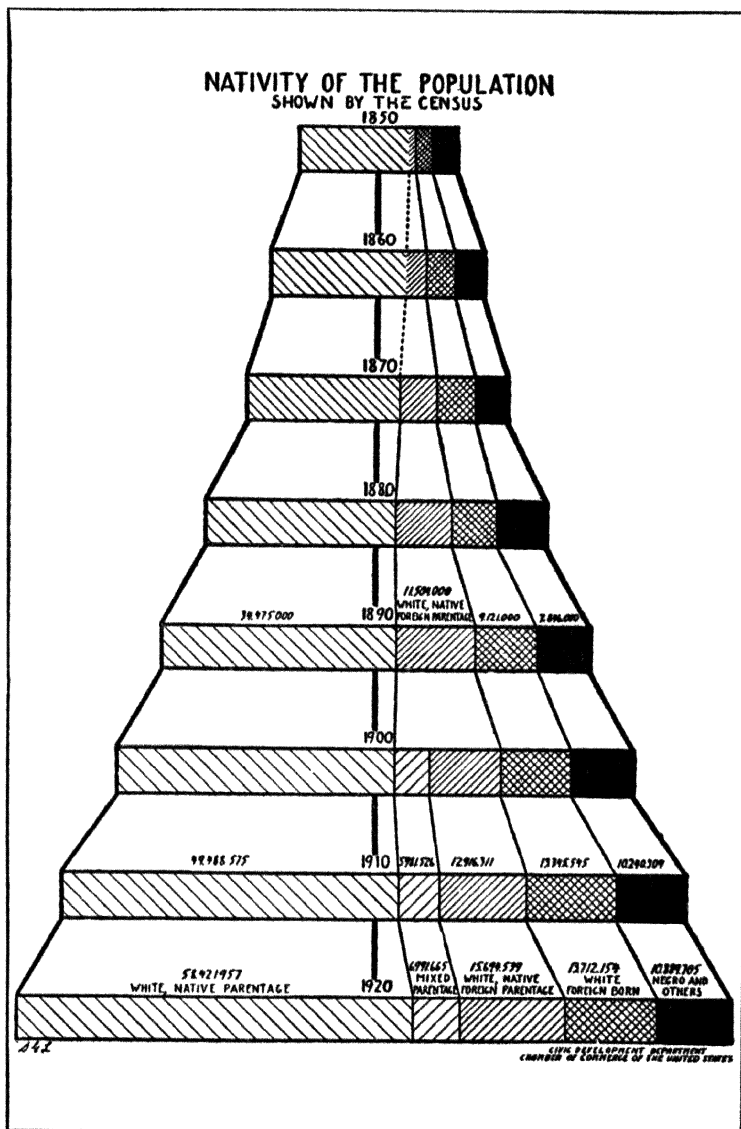


Chart 11: Horizontal bar chart arranged in perspective to give the effect of distance to earlier decades. The vertical axis is in the middle.

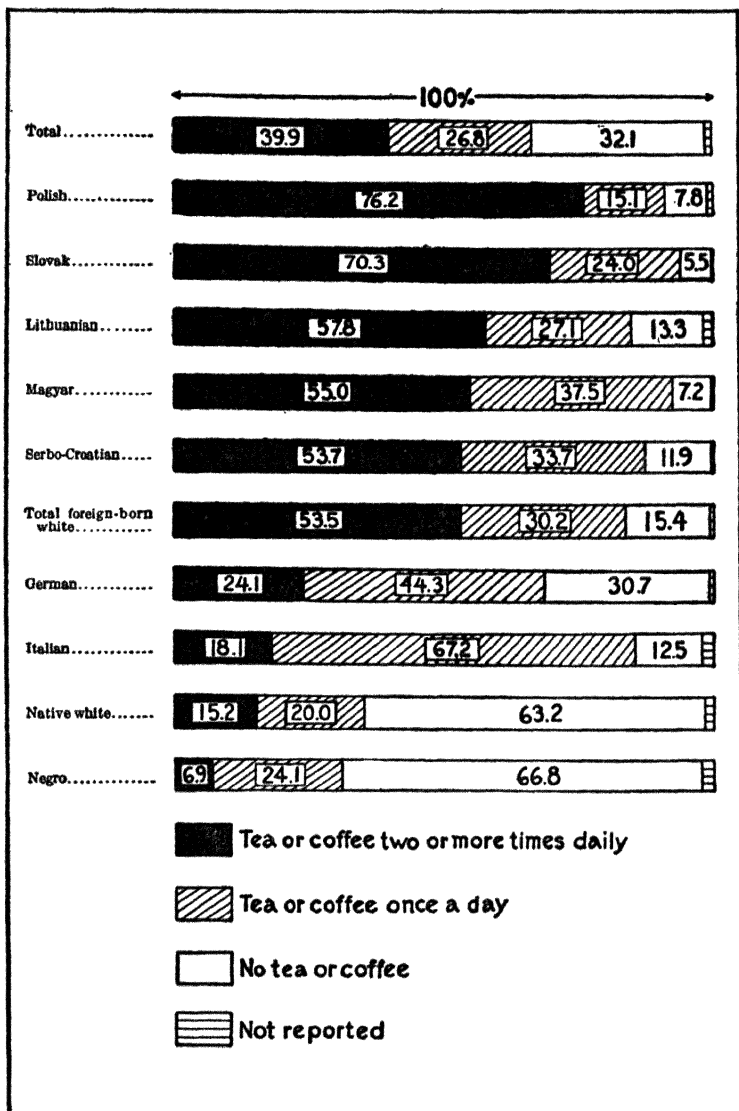


Chart 12: Horizontal bar chart showing the percentage of children two to seven years of age using tea or coffee by nationality of mother. A subdivided scale at the top with guide lines would have improved the chart.

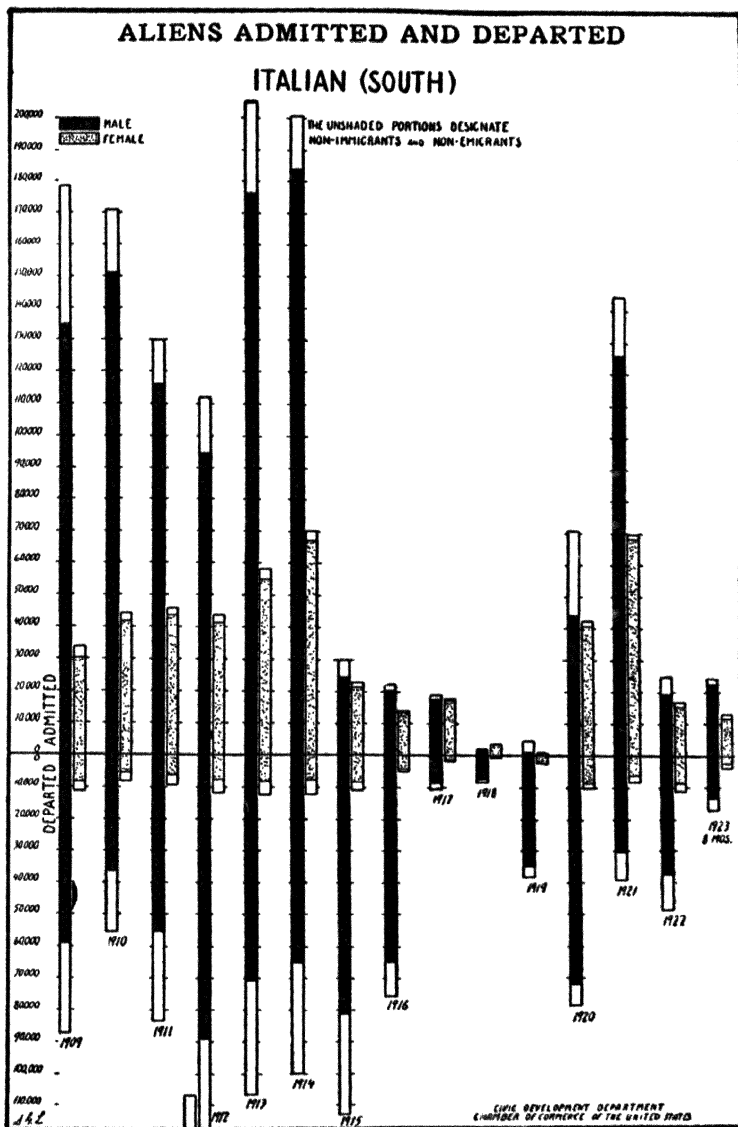


Chart 13: This chart illustrates the use of the zero line for the purpose of portraying positive and negative quantities.

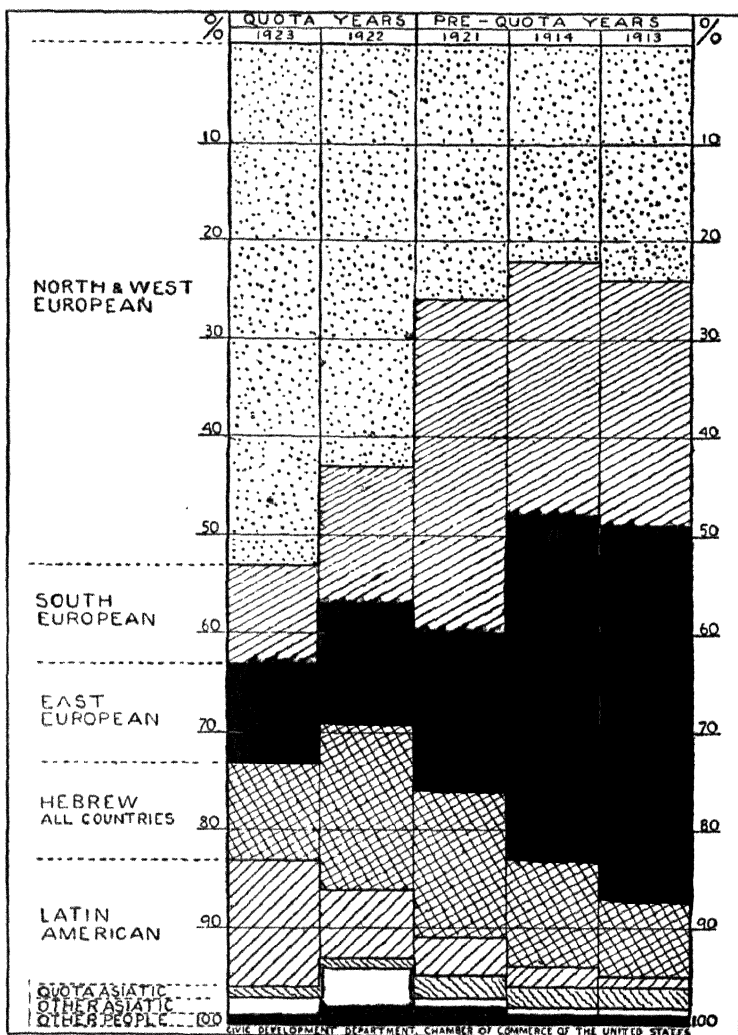


Chart 14: An excellent subject the possibilities of which are not fully realized due to rather lifeless cross-hatching. This chart has the opportunity of showing the steady decrease of immigration from Northern and Western Europe and the remarkable increase from Southern and Eastern Europe. Black is well chosen for the East European section, but the South European section should have been cross-hatched in heavy black lines in order to link the two together and to make them stand out in contrast to all the others.

	1890	1910	1920
1. PRESENT LAW PERMITTING 3% UPON NATIONALITY CENSUS OF 1910	11	11	11
2. PROPOSAL	11	11	11
3. "NET" 3% PROPOSAL WITH EMIGRATION OF 1923 ADDED TO PRESENT QUOTAS	11	11	11
4. PROPOSAL PERMITTING 2% UPON NATIONALITY CENSUS OF 1910	11	11	11
5. " "	11	11	11
6. " "	11	11	11
7. " "	11	11	11
8. " "	11	11	11
9. " "	11	11	11
10. " "	11	11	11
11. " "	11	11	11
12. " "	11	11	11
13. " "	11	11	11
14. " "	11	11	11
15. " "	11	11	11
16. " "	11	11	11
17. " "	11	11	11
18. " "	11	11	11
19. " "	11	11	11
20. " "	11	11	11
21. " "	11	11	11
22. " "	11	11	11
23. " "	11	11	11
24. " "	11	11	11
25. " "	11	11	11
26. " "	11	11	11
27. " "	11	11	11
28. " "	11	11	11
29. " "	11	11	11
30. " "	11	11	11
31. " "	11	11	11
32. " "	11	11	11
33. " "	11	11	11
34. " "	11	11	11
35. " "	11	11	11
36. " "	11	11	11
37. " "	11	11	11
38. " "	11	11	11
39. " "	11	11	11
40. " "	11	11	11
41. " "	11	11	11
42. " "	11	11	11
43. " "	11	11	11
44. " "	11	11	11
45. " "	11	11	11
46. " "	11	11	11
47. " "	11	11	11
48. " "	11	11	11
49. " "	11	11	11
50. " "	11	11	11
51. " "	11	11	11
52. " "	11	11	11
53. " "	11	11	11
54. " "	11	11	11
55. " "	11	11	11
56. " "	11	11	11
57. " "	11	11	11
58. " "	11	11	11
59. " "	11	11	11
60. " "	11	11	11
61. " "	11	11	11
62. " "	11	11	11
63. " "	11	11	11
64. " "	11	11	11
65. " "	11	11	11
66. " "	11	11	11
67. " "	11	11	11
68. " "	11	11	11
69. " "	11	11	11
70. " "	11	11	11
71. " "	11	11	11
72. " "	11	11	11
73. " "	11	11	11
74. " "	11	11	11
75. " "	11	11	11
76. " "	11	11	11
77. " "	11	11	11
78. " "	11	11	11
79. " "	11	11	11
80. " "	11	11	11
81. " "	11	11	11
82. " "	11	11	11
83. " "	11	11	11
84. " "	11	11	11
85. " "	11	11	11
86. " "	11	11	11
87. " "	11	11	11
88. " "	11	11	11
89. " "	11	11	11
90. " "	11	11	11
91. " "	11	11	11
92. " "	11	11	11
93. " "	11	11	11
94. " "	11	11	11
95. " "	11	11	11
96. " "	11	11	11
97. " "	11	11	11

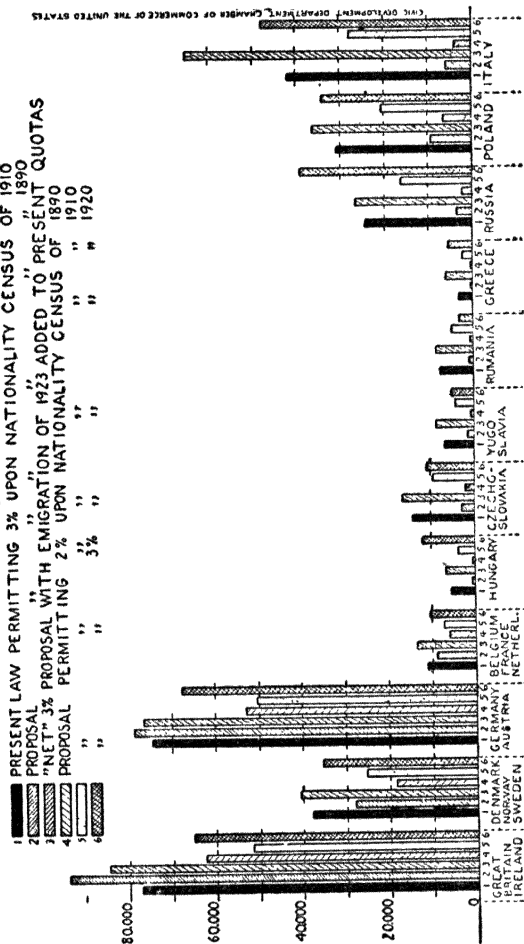
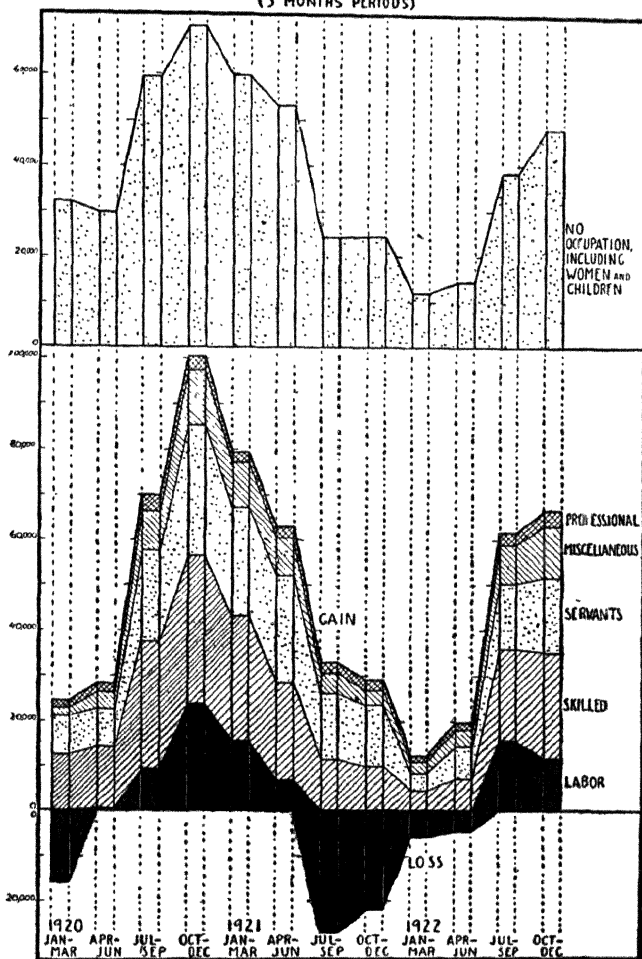


Chart 15: The above chart, while attractive to the eye, is very difficult of comparison due to the widely varying length of bars composing the groups, and due to the difficulty with which one compares the same element in each of the groups.

NET GAIN OR LOSS TO OCCUPATIONS RESULTING FROM ADMISSION OF IMMIGRANTS AND DEPARTURE OF EMIGRANTS (3 MONTHS PERIODS)



184 L.

CIVIL DEVELOPMENT DEPARTMENT
CHAMBER OF COMMERCE OF THE UNITED STATES

Chart 17: A variation of the subdivided bar chart.

complete bar to represent the total population and one portion of it to represent the urban and the other portion the rural population. Or if our study happens to be the standards of living, a larger number of divisions of the bar may be set aside for each of the elements composing the standard. Another manner of accomplishing the same thing is that of grouped narrower bars. But in either of these methods care must be taken not to make the chart so complicated as to produce confusion.

THE LINE GRAPH

As has already been noticed, *the line graph* is a perfectly normal evolution from, or refinement of, the vertical bar chart. In reality it stands for the same thing. Instead of erecting rectangles from a base line to given points representing the height of rise, these points are connected by lines, straight or curved, which really represent the same thing in a simplified manner. Either historical or frequency data can thus be represented, in the former case by means of dates, years, or decades along the base line and in the latter by means of designating the limits of the frequency classes. In either case the vertical axis will represent the frequency or the number of cases. Provided the same axes are used, it is possible to superimpose any number of graphs on the same chart within the limit set by the possibility of easily following them. Where the same base line is used, it is sometimes permissible to use two different vertical axes, one on the left and the other on the right-hand side of the chart representing different frequencies. In this case it is a good plan to have one of the vertical axes and the graph represented by that axis done in black and the other axis and its graph done in red, or light and heavy lines might be used. This prevents confusing the two. A frequency graph usually represents a sample plot, and the first step in making it is to connect by straight lines the points represented by each one of the classes in the sample. If the data represent a truly continuous series, we can get an idea what the com-

pleted universe would look like by ironing or smoothing out the angles. This is done by fitting a curved line to the points in question. There is no specific rule to follow in this ironing out of angles, except that care must be taken when deviating from the straight line to see to it that as much new area is added as has been excluded in other places. This may not give an exact picture of the universe, but it is a fairly close approximation.

The *cumulative frequencies* already discussed under tabulation can as easily, and in the same manner, be submitted to graphic presentation. They give us, however, quite a different type of graph from that of the simple frequency distribution. This graph is called the "ogive."

Percentages and other rates and ratios lend themselves to graphic presentation by means of circle and sector, bar chart, and line graph. One of the newer methods of graphic presentation showing proportional change is that of the ratio chart. By this method the logarithm of the number rather than the number itself is plotted. By ascertaining the logarithms of numbers, it is possible to plot such a curve on ordinary cross-section paper, but it is more desirable to use a specially prepared logarithmic paper, on which numbers can be entered directly and can be read off in terms, not of their logarithms, but of the original data. This method, however, is popularly so vaguely understood that it is of little use to any but statisticians.

Charts 18*a*, 18*b*, 18*c*, and 18*d*, based on the writer's investigation of the standard of living in North China, data for which are given in Table 15, illustrate the relationship existing between the bar chart, the line graph or frequency polygon, and the smoothed curve. The justification of smoothing lies in the fact that marked irregularities in straight-line graphs are probably due to the inadequacy of the sample. Considering Chart 18*b*, we can only conclude that there is no cogent reason for the marked fluctuation occurring at families of four or the less marked fluctuation

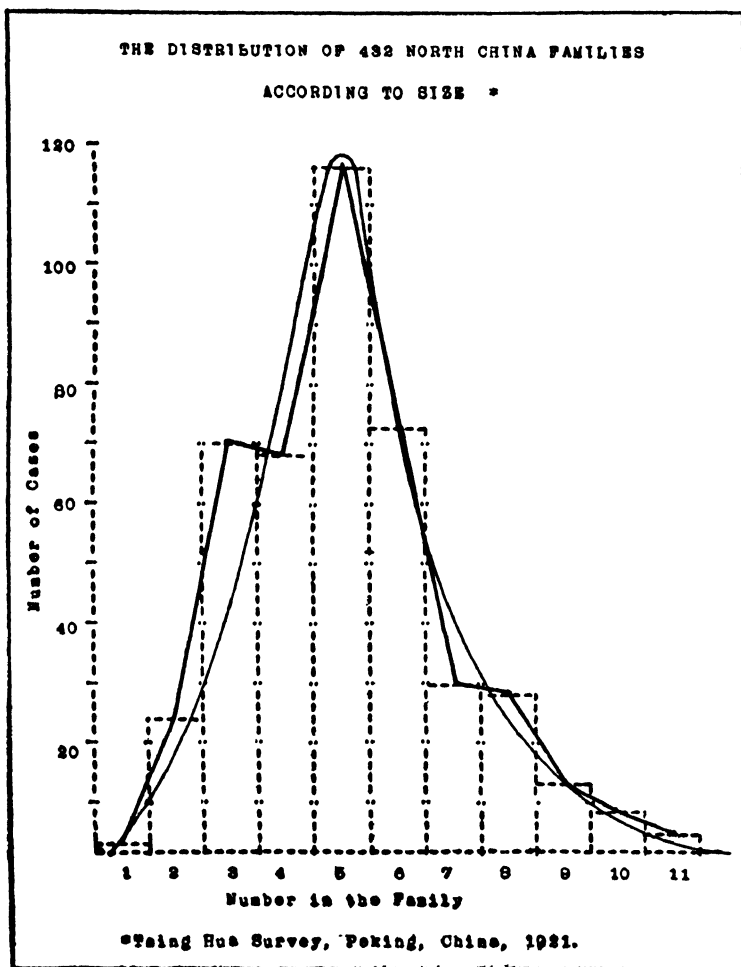


Chart 18a: A combination of the bar chart, straight-line graph or frequency polygon, and the smoothed curve.

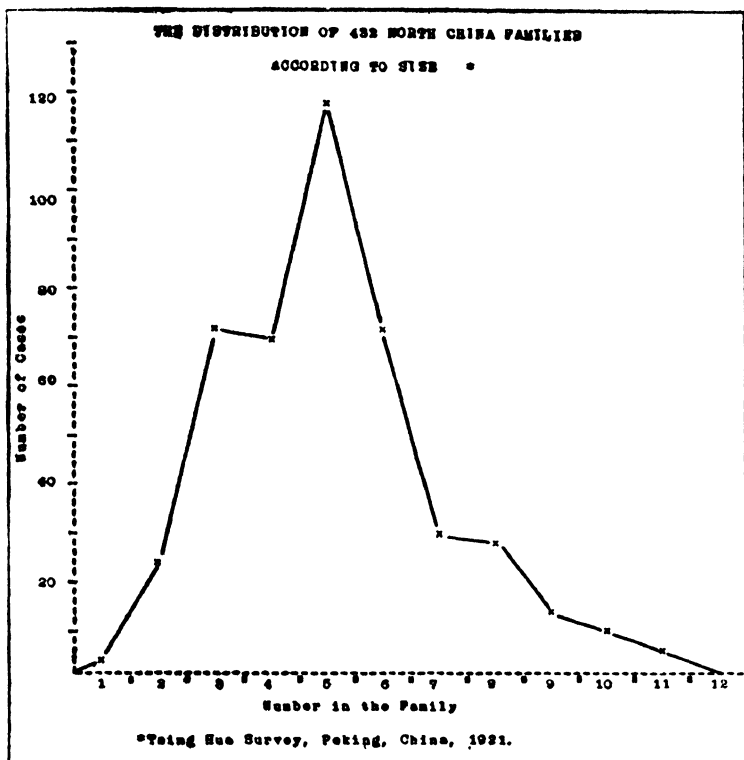


Chart 18b: The frequency polygon can be easily constructed from points without the aid of the bar chart or rectangular histogram used in the preceding chart (18a).

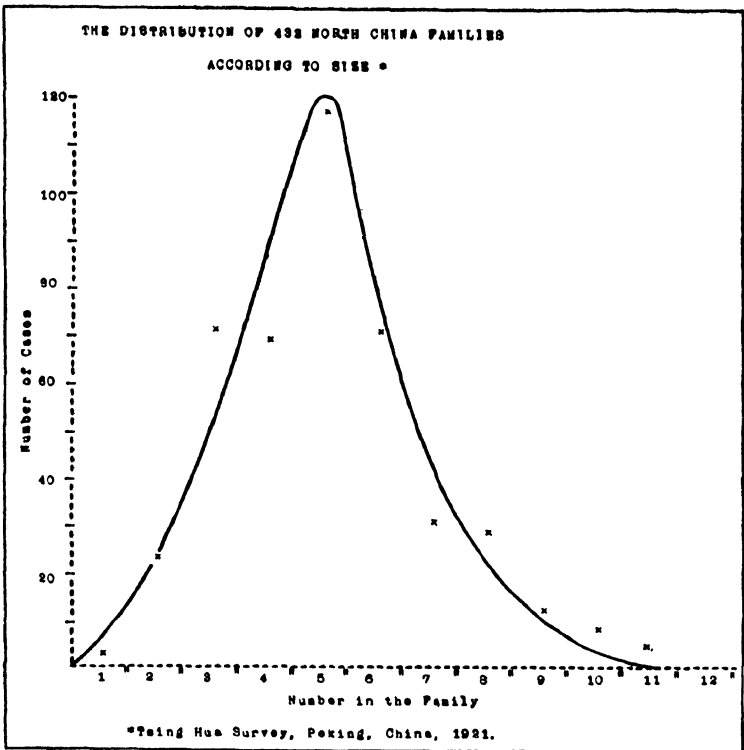


Chart 18c: The smoothed curve can also be drawn directly from points. Note the ease with which a chart can be set up directly with the aid of a typewriter. Charts 18a, 18b, and 18c represent discontinuous or discreet series and, as such, should be more sparingly smoothed.

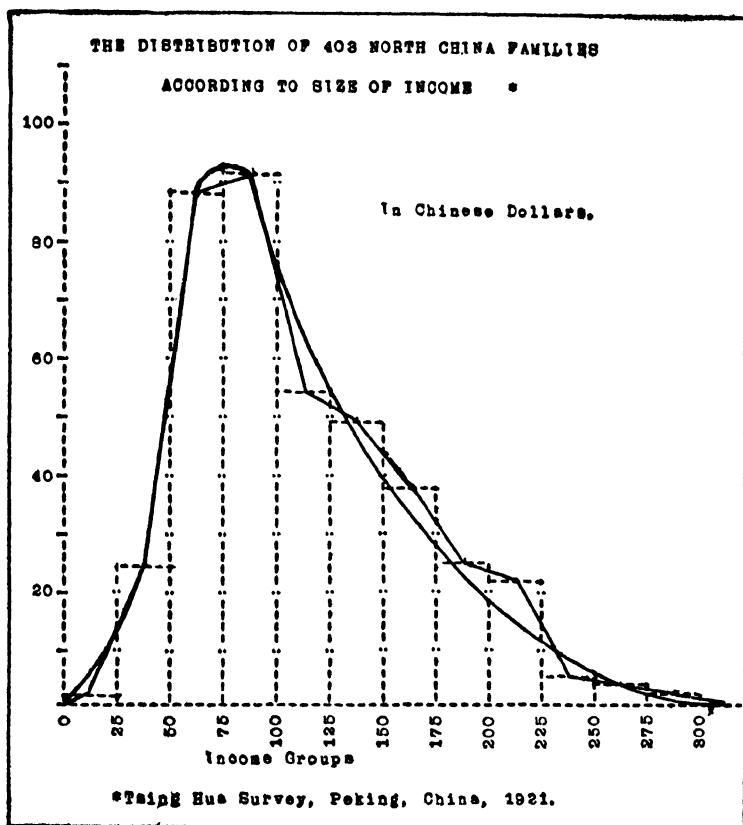


Chart 18d: The size of family represents a discontinuous series, but the above classification of income is continuous. Here greater freedom in smoothing may be indulged in.

at families of seven and eight. The fluctuation is more apt to be a statistical accident than an important social fact. Chart 18*c* shows us what more probably would have been the case had a larger sample been available. Chart 18*d* shows a higher degree of symmetry but with a marked skewness or leaning toward the left.

Note, also, in Chart 18*d*, that only the effective limits of

TABLE 15
THE DISTRIBUTION OF

412 NORTH CHINA FAMILIES
ACCORDING TO SIZE

491 NORTH CHINA FAMILIES
ACCORDING TO SIZE
OF INCOME

Tsing Hua Survey, China, 1921

Size of Family	Number of Cases	Size of Income in Chinese Dollars	Number of Cases
Total	432	Total	403
1	2	0- 24.9	2
2	23	25.0- 49.9	24
3	70	50.0- 74.9	89
4	69	75.0- 99.9	91
5	116	100.0-124.9	54
6	71	125.0-149.9	49
7	29	150.0-174.9	37
8	27	175.0-199.9	24
9	12	200.0-224.9	21
10	8	225.0-249.9	5
11	5	250.0-274.9	4
12	0	275.0-299.9	3

the frequency classes are expressed along the base line. Sometimes each bar is designated by recording the entire class interval, as 0-24.9, 25.0-49.9, and so on. This procedure is not only unnecessary, but adds to the complexity of the chart.

The first three charts, 18*a*, 18*b* and 18*c*, based on the size of the family, are discreet or discontinuous while the fourth, 18*d*, based on income, is a continuous series. Continuous series can be more freely smoothed than discontinuous ones, though smoothing is allowable in both cases.

CRITERIA OF GOOD GRAPHIC PRESENTATION

Most of the criteria applied to good tabular presentation already noted in a preceding chapter, apply equally well to charts and graphs. (1) Of primary importance is the designation or title which must provide adequate description without superfluity. (2) The data on which it is based must be present either within the chart itself or on an included or accompanying table. (3) Units of expression must be made clear on scale or in special key. (4) Where cross-hatching or different types of dots or other methods have been used, a key must be provided within the body of

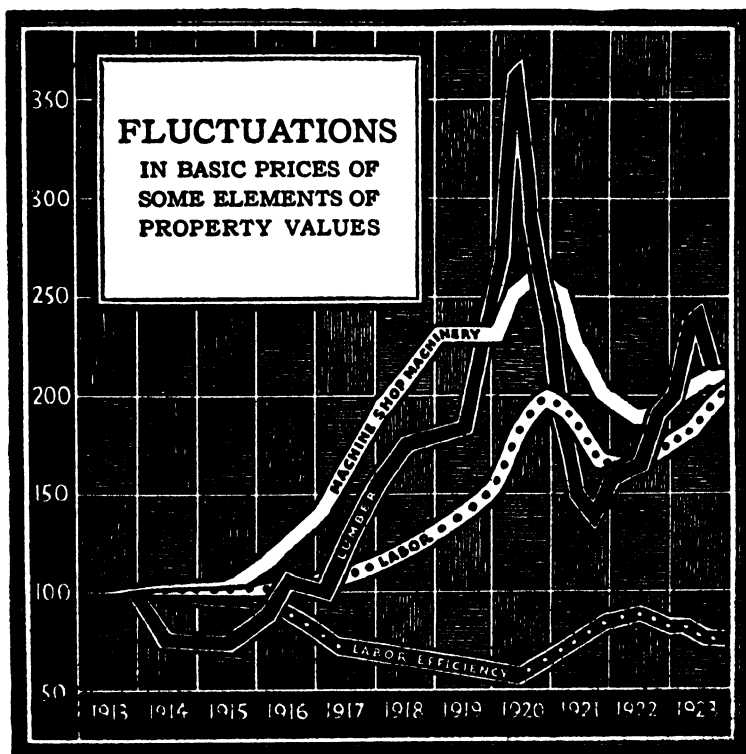


Chart 19: An unusual and effective treatment which arrests attention.

the chart which will show at a glance their respective meanings. (5) The mechanical form of the table should be adjusted to the size of the page which it is to occupy. Through the proper selection of lines, types, and colors, it should be given perspective. (6) When necessary to diverge from the horizontal, printing should be so arranged that the page can be read from the bottom or from the right-hand side. (7) Avoid the optical illusions which so easily result from unwise cross-hatching or from too complicated line graphs. (8) In bar charts and line graphs base lines and vertical and horizontal axes should be heavier than other rulings. (9) When more than one graph appears on the same chart,

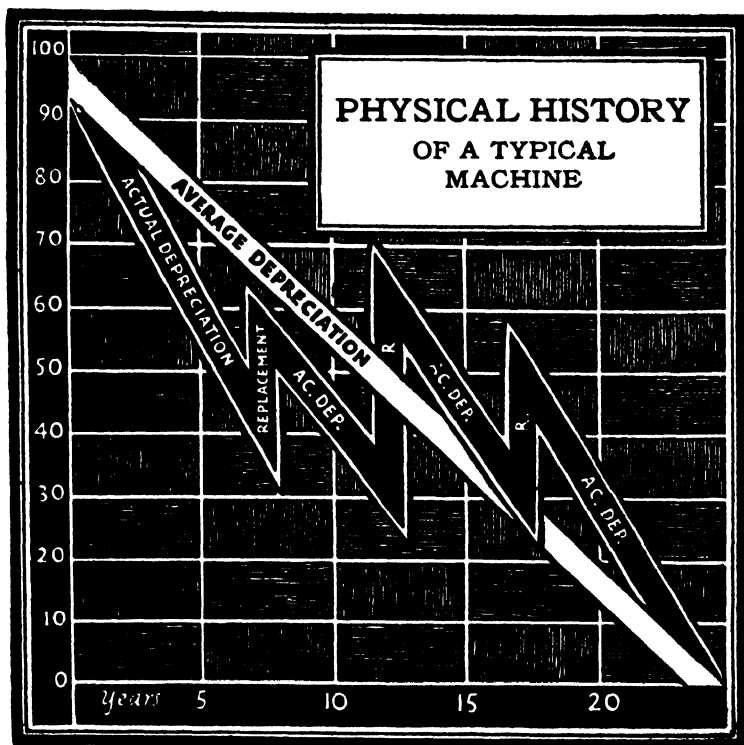


Chart 20: A chart which compels the reader to stop and figure it out.

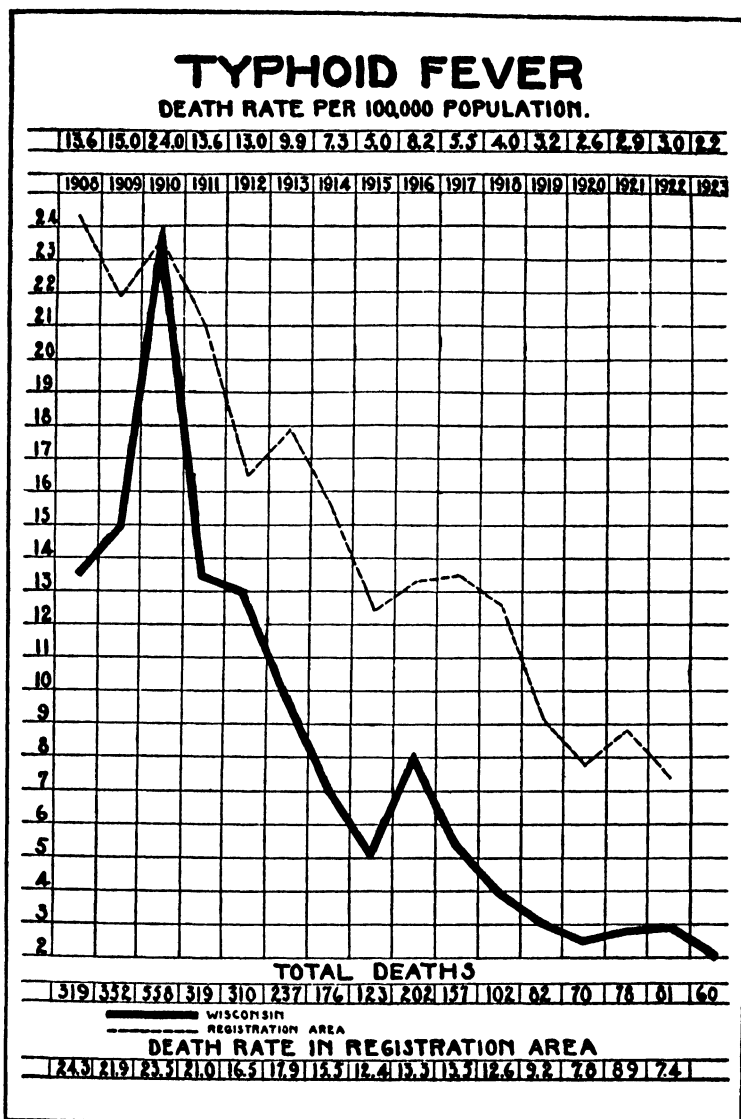


Chart 21: A unique and not entirely satisfactory treatment of the chronological line graph.

these may be indicated by lines of different colors, light or heavy lines, or by solid lines and broken lines of different interval. (10) Scales indicating the length of bars or the height of curves should be clearly indicated and conveniently located. For the horizontal bar chart this scale will be located at the top above the first bar; while at the left the significance of each bar will be indicated. In the vertical bar chart or in the line graph the scale should usually be at the left, though it may occur at the right when two scales are used or when the size of the table demands it. At the bottom along the horizontal axis will be placed either chronological values or class intervals. In chronological data bar designations are sometimes placed also at the top. (11) It is desirable but not always possible to show the zero point on scales. Where this is not possible it is a good practice to show by the use of a break in the scale where the interval occurs between the zero point and the point where the scale is resumed. (12) Since it is possible by the choice of disproportionate vertical and horizontal scales to distort the resulting graph, it is a good plan to have both in proportion. This, however, is not always possible and it is sometimes even desirable to magnify variations in order to make them stand out. (13) For the qualitative and geographical distributions the horizontal bar chart should be used. The vertical bar chart should be used for all frequency (quantitative) distributions and for chronological data where the periods are more than four in number.

The above general rules have been purposely reduced to the minimum since, as in the case of the table, each graph or chart represents a highly individual problem. A careful statistician will gain much information in regard to graphic presentation by the study of tables made by others and analysis in the light of the foregoing standards of good graphical presentation. Two valuable works on the subject, are *Graphic Methods for Presenting Facts*, by W. C. Brinton; and *Charts and Graphs*, by Karl G. Karsten.

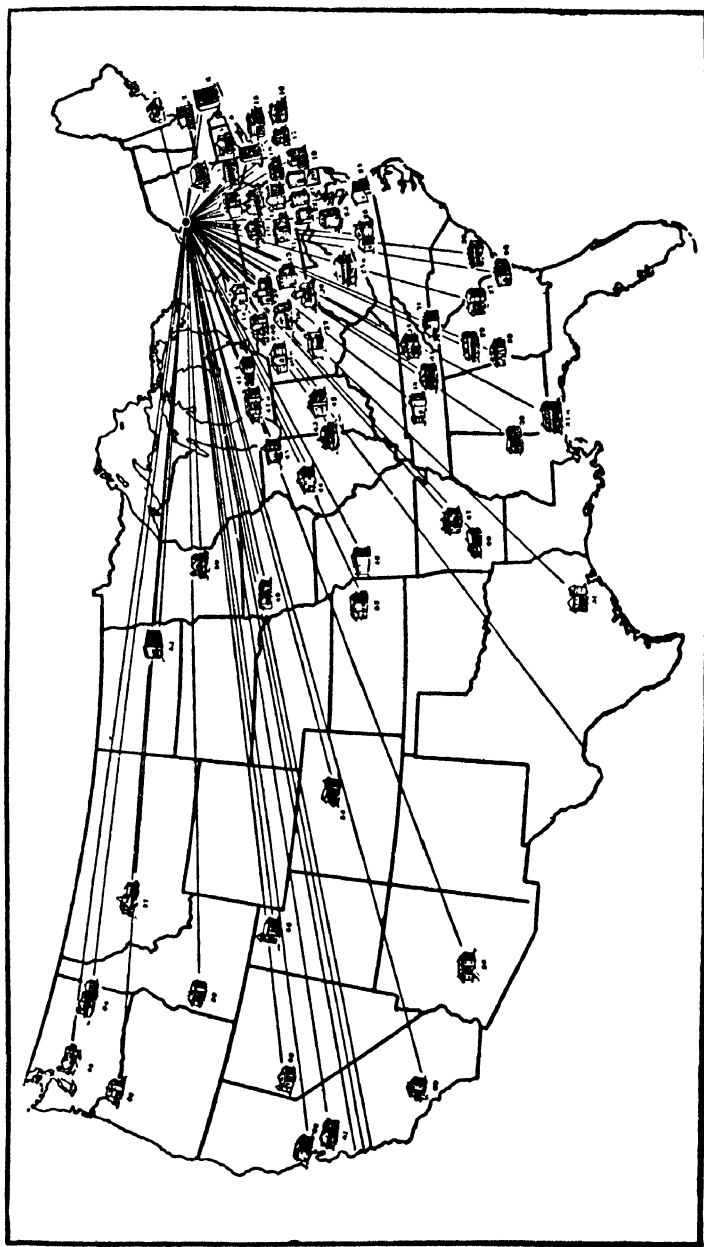


Chart 22: The pictogram and cartogram combined to show the distribution of Florence Crittenden Missions.

STANDARD RULES FOR GRAPHIC PRESENTATION¹

1. The general arrangement of a diagram should proceed from left to right.
2. Where possible, represent quantities by linear magnitudes, as areas or volumes are more likely to be misinterpreted.
3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.
4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.
5. The zero lines of the scales for a curve should be sharply distinguished from the other coordinate lines.
6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100% line or other line used as a basis of comparison.
7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not present the beginning or end of time.
8. When curves are drawn on logarithmic coordinates, the limiting lines of the diagram should each be at some power of 10 on the logarithmic scales.
9. It is advisable not to show any more coordinate lines than necessary to guide the eye in reading the diagram.
10. The curve lines of a diagram should be sharply distinguished from the ruling.
11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.
12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.
13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.
14. It is often desirable to include in the diagram the numerical data or formulas represented.
15. If numerical data are not included in the diagram, it is desirable to give the data in tabular form accompanying the diagram.

¹ Work of the Joint Committee on Standards for Graphic Presentation, W. C. Brinton, Chairman. *Quarterly Publication of the American Statistical Association*, Vol. XIV, December, 1915.

16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

17. The title of a diagram should be made as clear and complete as possible. Subtitles or descriptions should be added if necessary to insure clearness.

VII

NUMERICAL SUMMARIZATION

The meaning of numerical summarization. Types of numerical summarization. Averages: the arithmetic mean; the mode; the median. The characteristics and limitations of averages. The weighted average.

THE purpose of statistical tables and of various types of graphic presentation is to bring about order within given masses of statistical data and to present them in such manner that the relationship between parts can be observed. While accomplishing this purpose, however, these methods do not so easily lend themselves to the further purpose of the comparison of the separate groups of statistical data thus treated. We need, then, not only statistical methods, for the analysis of the relationship between parts, methods which give pictures representing elements composing the whole, but we also need methods which will reduce the whole to terms of one numerical value, in order that comparison between different groups of data may be facilitated. This is the aim of the types and averages of various sorts which are used by statisticians but so little used and understood by the general public. Statistics has been called the science of averages, but this statement is correct only if we understand the manner in which the latter term is used. The word average as popularly used and understood is merely a very loose generalization which covers a multitude of evils. The definition really means that on the average consecutive samples from within a given field will show a marked degree of similarity. We are continually hearing such expressions as, "on the average," "normal conditions," "per capita wealth," "a typical year," and so forth. These concepts and the numerical values in which they are usually expressed in-

dicating the need in the popular mind for some simple one-term expression by means of which comparisons can be made. The man of the street is not interested in our graphs and charts; indeed, he will probably view them with suspicion. But tell him that the average wage in a given industry is \$4 an hour, or that the general death-rate is but half of what it was 100 years ago, and he will be perfectly satisfied. Secrist well remarks that, "every average is a sort of fictitious substitute for the details which it replaces, serviceable when the conditions for which it stands are known, but deceptive when they are ignored." The seven main methods of summarization are: (a) averages, (b) measures of dispersion, (c) measures of skewness, (d) coefficients of correlation, (e) rates and ratios, (f) index numbers, and (g) probable error and normal distribution.

A. AVERAGES

1. *The arithmetic mean*, commonly called the average, is the most popularly used method of summarization. By the term is usually meant the most commonly observed quality or characteristic. When we say, for instance, that the average age of students on a given campus seems to be 18 years, we mean that students of that age are most commonly observed or encountered, but this most frequently observed or commonest characteristic in reality represents not the so-called average but the mode, which is obtained by quite a different mathematical process. The method of determining the arithmetic mean is simple and familiar to every one. It is merely that of adding up all of the items in a given distribution and dividing the total by the number of cases. This does not of necessity give us the size of the most frequently occurring item. If, for instance, we total the following five series and divide by the number of cases, which is the same for each, we will arrive at an identical arithmetic mean for each in spite of the fact that the series are widely different in nature and in the case of the third, fourth, and

TABLE 16
SAMPLE DISTRIBUTION OF NUMERICAL DATA FOR THE
ESTIMATION OF THE ARITHMETIC MEAN

	A	B	C	D	E
	10	51	42	35	15
	20	51	42	35	20
	30	53	44	35	35
	40	53	44	36	35
	50	55	44	36	36
	60	55	44	36	36
	70	57	45	37	36
	80	57	45	100	37
	90	59	100	100	100
	100	59	100	100	200
Total	550	550	550	550	550

The total in each case is 550. This, divided by 10, which is the number of cases in each instance, gives a common arithmetic mean of 55 for each.

fifth the mean does not represent the most commonly recurring item. This weakness of the arithmetic mean is well illustrated by the per capita concept. The per capita wealth in an industrial center may well be from two to five thousand dollars, but this tells little of the actual condition existing in the community, when we know that the community is made up of a great mass of poorly paid laborers and a few of the millionaire owners of the means of production. The advantages of the arithmetic mean are, in the main, that its computation is simple, and is well understood by the general public and that, for certain statistical purposes, we desire to give full weight to extreme items. It takes into consideration all sizes and gives weight to extreme variations without regard to distribution. It can be determined without knowledge regarding the individual items if the total and number of cases are known. Its main disadvantage lies in its popularity and in its careless use, not taking into consideration the fact that it does not represent the most common or frequently recurring item. The computation of the arithmetic mean may be reduced to the formula given at the top of the opposite page.

$$A = \frac{\Sigma m}{n}$$

A = The arithmetic mean

m = The size of the individual item

n = The total number of items

Σ = The sum of

The formula may be interpreted as follows: the arithmetic mean equals the sum of the individual items or the total divided by the number of items. The Greek letter Sigma (Σ), which appears in the formula, is the sign of summation which is usually used by mathematicians.

TABLE 17

SAMPLE DISTRIBUTION FOR THE ESTIMATION OF THE ARITHMETIC MEAN FROM A FREQUENCY TABLE

Class	Simple Frequency	Mid-Point of Class	Product of Frequency and Mid-Point
0.0- 4.9	1	2.5	2.5
5.0- 9.9	3	7.5	22.5
10.0-14.9	5	12.5	62.5
15.0-19.9	7	17.	122.5
20.0-24.9	5	22.5	112.5
25.0-29.9	3	27.5	82.5
30.0-34.9	1	32.5	32.5
Total	25		437.5

Dividing 437.5 by 25 we discover the arithmetic mean, as estimated from the table, to be 17.5. But from the actual items on which this table is based (page 116) we find the true total to be 423 and the true arithmetic mean to be 16.9.

The arithmetic mean, when based on a frequency distribution, is really an average of averages. We assume that the mid-point of the class will adequately represent the items of that class, but this is not necessarily the case. In the second (5.0-9.9) class, for example, we have three items or a frequency of three. We assume that the mid-point of the class, 7.5, is representative, but the actual items prove to be 5, 5, and 9, the average of which is 6.3, or more than one unit less.

The arithmetic mean, thus obtained by this process, usually approximates fairly closely the true arithmetic mean; its use, however, is justified only when its limitations are recognized.

2. *The mode* is the real concept which the "Mr. Average Man" has in mind when he speaks of the average. The mode is the most frequently recurring item or the most commonly observed size or characteristic. In the industrial center referred to on page 114 there will be a wide amount of difference between the arithmetic mean, which would include extremes of wealth and poverty, and the mode, which would ignore them entirely and indicate only the prosperity of the mass of the population. It would be well if the idea of the mode could be popularized, as it is in reality the most logical method of indicating the characteristics of a group. It

should, in fact, take the place of the arithmetic mean, commonly called the average for general purposes. The main difficulty is that it is not so easily derived. In order to determine it, it is necessary to know the size of each item and to arrange them in the form of a frequency table which will at once show where the massing of data or of items occurs.

The items listed in Table 18 might easily be added, divided by the total number, which is 25, and at once give us an average. This distribution, however, gives little indication as it stands as to where the mode lies. There is no particular evidence of concentration in its present form. Bringing these items together into the form of a frequency table with a class interval of 5, we get the frequency distribution which falls under "A" in Table 19. It will be noted now that the largest number of items are included

TABLE 18
SAMPLE DISTRIBUTION OF UNTABULATED ITEMS

4
5
5
9
10
11
12
13
13
15
16
17
17
17
18
19
20
21
22
23
24
25
27
29
31
423

within the limits of the fourth (15.0-19.9) class. Therefore, we know that the modal group is from 15 and 20 in size. Passing to distribution "B" we find the point of concentration to be between 10 and 15 and in distribution "C" to be between 15 and 25. Thus at a glance a statistical table shows us the point or points of concentration between the limits of which the mode lies in case a definite mode be present at all. It is often desirable, however, for us to know exactly what point, between these limits, may be considered as the modal point. For this purpose a formula, the theory of which we shall not attempt to discuss, has been evolved by the mathematician and which we can easily use for this purpose. This is called the method of *interpolation*.

TABLE 19

SAMPLE DISTRIBUTION FOR THE ESTIMATION OF THE MODE

CLASS	FREQUENCY DISTRIBUTIONS		
	A	B	C
0.0- 4.9	1	2	1
5.0- 9.9	3	5	2
10.0-14.9	5	7	3
15.0-19.9	7	4	4
20.0-24.9	5	3	7
25.0-29.9	3	2	5
30.0-34.9	1	1	2

$$Mo = L + \frac{Cf_2}{f_1 + f_2}$$

Mo = Mode

L = The lower limit of the modal class

C = The class interval

f_1 = The frequency of class next preceding the modal class

f_2 = The frequency of the class next following the modal class

A glance at frequency distribution "A" will demonstrate the fact that, since the frequencies are equally balanced on either side of the modal class, the mode must of necessity fall in the exact middle of the modal class, or at point 17.5.

Applying the formula to distribution "A" we obtain the following results:

$$15 + \frac{5 \times 5}{5 + 5} = 15 + \frac{25}{10} = 15 + 2.5 = 17.5 = \text{Mode}$$

Applying the formula to distribution "B" where there is no such balancing, we obtain the following results:

$$10 + \frac{5 \times 4}{5 + 4} = 10 + \frac{20}{9} = 10 + 2.22 = 12.22 = \text{Mode}$$

Another method of locating the mode can be demonstrated by turning to Charts 18*a*, 18*b*, 18*c*, and 18*d*. A glance at Table 15, on which these charts are based, will show that for the distribution which is based on the size of families the modal class is very definitely located at families of 5, with a frequency of 116, which is markedly larger than the frequencies on either side. The distribution based on the income of these families, however, presents a more complicated problem. Here we have two points of concentration, one of 89 and the other of 91, so close in size that we would scarcely be justified in choosing the slightly larger one as indicating the modal class. Turning now to Charts 18*b* and 18*c*, it will be noticed that the highest point of rise in the graph corresponds to this point of concentration in the table. In order to discover the actual mode, provided, of course, that the chart has been carefully enough drawn, it is necessary only to drop a perpendicular from the point of the highest rise on the chart to the base line below and read the size of the mode at the point of intersection. Turning to Chart 18*d* it will be noticed that it would be illogical to drop a perpendicular from the point of highest rise on the unsmoothed straight-line graph, but from the smoothed curve it is easily done and we would obtain a reading of a bit over 75 as the modal income of these 403 North China families. If we were using the method of interpolation in the data represented by Chart 18*d*, it would be a good plan to combine the frequency classes into class intervals of 10 instead of class intervals of 5. This would result in combining the two first classes which have frequencies of 2 and 24 into

one class of 26, and the next classes of 89 and 91 into a new modal class of 180 and the following two classes of 54 and 49 into one new class of 103. The result thus obtained by interpolation would more adequately represent the series than would be found by using 91 as the modal class and 89 as the preceding and 54 as the following frequencies.

The column of items on page 116, arranged in order according to size but not grouped into frequency classes, does not show a definite point of concentration which would adequately indicate the mode. Tabulating them with class intervals of two, three, or four fails to produce the desired result, but upon using a class interval of five a modal class definitely appears. The statistician will often be compelled to experiment with different groupings and different class intervals before he discovers the one which shows the mode to the best advantage. The primary value of the mode lies in the fact that it really represents what we are usually talking about when we use the word "average." It is based on frequencies rather than on individual sizes and does not take into account the extremes which are so apt to distort the size of the arithmetic mean. It is not a hypothetical value but rather that point at which most of the instances occur. It is characteristic, it is typical, and it is representative.

3. *The median.* The median is a third type of average. It is merely the size of the central or middle item in a given series. It does not seem at first thought as though this particular item would be as representative of the whole as are either the arithmetic mean or the mode, and yet we usually find it true that the median fairly closely approximates both of the others. The median is in reality a partition measure as are also the quartiles and the deciles, which will be discussed later. There are three methods of determining the size of the median. The first and simplest method can be used only when the entire series of items is present and in which they are arranged in order according to size. The middle item is ascertained by counting and its size repre-

sents the median sought. In the items ranged on page 116, for instance, the middle or thirteenth item happens to be 17. The second method is that of interpolation by formula in a frequency distribution.

TABLE 20

SAMPLE DISTRIBUTION FOR THE ESTIMATION OF THE MEDIAN

Class	Simple Frequency	Cumulative Frequency
0.0- 4.9	1	1
5.0- 9.9	3	4
10.0-14.9	5	9
15.0-19.9	7	16
20.0-24.9	5	21
25.0-29.9	3	24
30.0-34.9	1	25

In locating the median (the size of the middle item) it is first necessary to discover which one in numerical order is the middle item. In a series of three items the second item is obviously the middle one. In a series of four the middle one is a hypothetical point midway between the second and the third. For convenience where the number of items is large, the following formula may be used:

$$\frac{n+1}{2}$$

"*n*" represents the total number of items. By adding one and dividing by two we discover the numerical rank in sequence of the median item. In Table 18, on page 116, there is a total of 25 items. Accordingly 25 plus 1 divided by 2 equals 13. The thirteenth item in the series is thus the middle or the median item.

Having designated the median item our next task is to discover the class in which it falls and its estimated size.

The location of the class within which it falls is most easily done by making a cumulative frequency. We notice, for instance, in Table 20 that there is a total of 9 items up to the end of the third class and a total of 16 items up to the end of the fourth. Consequently, the thirteenth item

must fall within this fourth class, the limits of which are 15.0 and 19.9.

The exact size of the median or thirteenth item is discovered by the use of the following formula:

$$Mi = L + \frac{C(2i-1)}{2f}$$

Mi = The Median

L = The lower limit of the median class

C = The class interval

f = The frequency of the median class

i = The numerical order or rank of the median item in the class within which it falls

The only item in the above formula which may need further explanation is " i ," or the numerical order of the median item in the class within which it falls (the median class). With respect to the entire series the median item is the thirteenth item, but with respect to the class within which it falls, which item is it? This is easily determined by reference to the cumulative frequency. Since there are a total of nine items in the three classes preceding the median class which is the fourth, and since the thirteenth item is the median item, we need only to subtract 9 from 13 to determine its rank within the median class. Thus " i " is found to be 4 in Table 20, to which we are applying our formula, f is 7, C is 5, and L is 15.

Interpolating the above values from Table 20 we arrive at the following result:

$$15 + \frac{5(2 \times 4 - 1)}{2 \times 7} = 15 + \frac{35}{14} = 15 + 2.5 = 17.5 = \text{Median}$$

The reader will probably have noticed, as in the case of the use of the above frequency distribution for the determination of the mode, that the equal balancing of the frequencies on either side of the middle class makes it obvious that the median also can fall only at the exact mid-point of the class, which is 17.5.

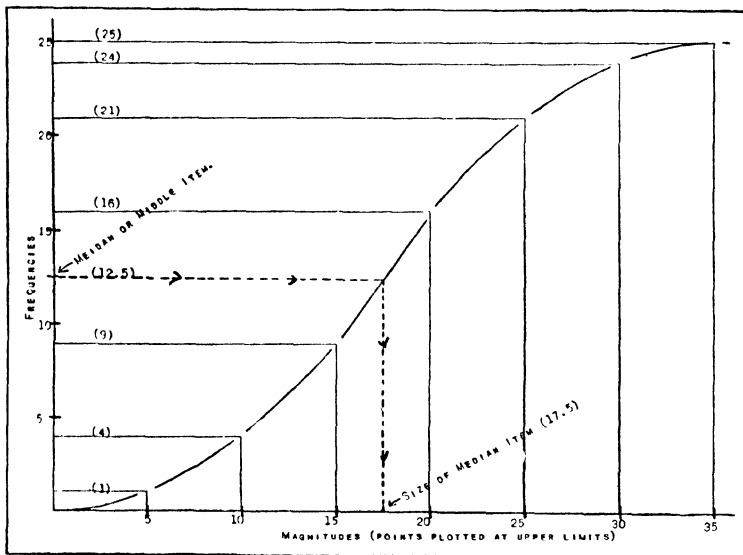


Chart 23: Cumulative frequency graph based on sample distribution for the purpose of locating the median.

The third method is location by means of a cumulative frequency graph and is illustrated by the above chart.

The median on the cumulative frequency graph is found by first locating the mid-point of the frequency distribution on the vertical axis. From this point project a line parallel to the base line till it meets the smoothed curve. From this second point drop a perpendicular to the base line and read the magnitude of the median directly from the scale. In this case the vertical axis locates the rank of the median and the base line or the horizontal axis its size.

Note that in this cumulative frequency graph the points are located at the upper limit of each class instead of at the mid-point.

In conclusion, it may be noted that the arithmetic mean takes into account to the largest extent the extremes of variation present in the distribution. The median gives them less emphasis due to the fact that merely their pres-

ence is taken into consideration and not their size; while the mode pays no attention to them whatsoever. In the case of an absolutely symmetrical or normal frequency distribution, balanced equally on either side—as in the case of the sample data, which we are using—the arithmetic mean, the median, and the mode will fall at exactly the same point. But, in case the distribution is not normal or symmetrical, being more heavily weighted on one side of the middle item than on the other, we will find the three items separated. First in order of size will occur the mode which, as we have said, pays no attention to the extremes of variation; second in size, either larger or smaller, will occur the median; and third, the arithmetic mean, which takes into largest account the extremes of variation.

THE WEIGHTED AVERAGE

A weighted average is one in which the elements composing it are given different values or numerical weights. For instance, a student in a given semester for five different courses may receive the following grades: 70, 80, 85, 85, and 95. The simple average for this group would be 83, but when we take into consideration that the first course was a 2-hour or credit course, the second a 3-hour, the third a 3-hour, the fourth a 3-hour, and the fifth a 5-hour course, we realize that the simple average of 83 does not really represent the group. We must then multiply each of the grades by the number of hours represented by the course. This gives us the following values: 140, 240, 255, 255, 475. The total number of hours represented by the courses is 16 and the total of the weighted grades is 1,365. Dividing this by 16 we get 85.3 as the weighted average of grades which the student has made during the semester. Similarly in the cost of living studies, where everything from the price of eggs to the price of automobiles is included and averaged together, it is necessary to weight each price on the basis of the total

amount of consumption per year. Thus, while the price of eggs may average 40 cents per dozen, we must multiply this 40 cents by the estimated number of dozens which will be used by the average family, and likewise with other commodities.

The frequency table represents in a way the weighted average principle, and the process of deriving an average from a table is similar to the process of weighting the grades; the mid-point of the group representing the size or magnitude of an item and the frequency representing the weight which must be used.

VIII

NUMERICAL SUMMARIZATION

(*Continued*)

Measures of dispersion: the range; average deviation; quartile deviation; and standard deviation. Coefficients of deviation. The measure of skewness. Coefficients of correlation. Rates. Index numbers. Probable error.

B. MEASURES OF DISPERSION

MEASURES of dispersion exist for the purpose of indicating either uniformity or its lack within a given group. The averages which we have just considered give us only a partial summarization of the group, while the numerical measure of dispersion helps to complete it. We are told that a certain king's guard was composed only of men who were 6 feet tall. This condition represents a high degree of uniformity or a very small amount of dispersion. If, on the other hand, we were to record the heights of the first hundred people that we passed on the street, our items would include everything from the height of the tiny tot to that of the adult who comes in "over size." Here, then, would be a marked lack of uniformity or a high degree of dispersion. In measuring social data it is as highly important that we should possess indications of uniformity as of average.

The range. The first indication of the amount of dispersion lies in the "range." Having assembled our items in order according to size, we note the difference between the sizes of the smallest and the largest items. In the king's guard above referred to, the shortest man may have been 5 feet and 11 inches and the tallest man 6 feet and 1 inch in height. The total range of dispersion thus estimated would be but 2 inches. In the case of the hundred people whose

heights we recorded, however, the smallest may have been a babe in arms of but 18 inches and the tallest a man 6 feet and 6 inches in height. In this case we notice a range of difference of 5 feet. We need, however, a finer method of measurement than this. It is found in average, standard, and quartile deviations.

Average deviation. The computation of the average deviation is very similar to that of the computation of the simple arithmetic mean and may be estimated on the basis of any average, though the median is most commonly used. The method of ascertaining the average *amount* of deviation within a series of items is first to find the difference between each item and the average used. Some of the deviations may be positive and some of them may be negative, but we are interested only in the fact that they are deviations and not in their direction. Next the deviations are added up and divided by the number of items. This is the average amount by which each of the items deviates from the average used. The computation of the average amount of deviation from a frequency table is slightly more complicated and is explained in Table 21.

TABLE 21

SAMPLE DISTRIBUTION FOR THE ESTIMATION OF THE AVERAGE AMOUNT OF DEVIATION FROM THE MEDIAN (17.5)

Class	Mid-Point of Class	Frequency	Deviation from the Median	Deviation Times Frequency
		<i>A</i>	<i>B</i>	<i>A</i> x <i>B</i>
0.0- 4.9	2.5	1	15	15
5.0- 9.9	7.5	3	10	30
10.0-14.9	12.5	5	5	25
15.0-19.9	17.5	7	0	0
20.0-24.9	22.5	5	5	25
25.0-29.9	27.5	3	10	30
30.0-34.9	32.5	1	15	15
Totals		25		140

Dividing 140 by 25 we get 5.6, which is the average amount by which each of the 25 items above represented deviates from the median. In the above illustration, since the three averages fall at the same point, it also represents deviation from the mode and the arithmetic mean.

The above process can be expressed in the following formula:

The average amount of deviation $= \frac{\Sigma d}{n}$ in which Σ = the sum of, d = the amount of the deviation of each item from the average used, and n = the number of items.

An amount of deviation, however, is not useful for comparative purposes. The average amount of deviation between students in the matter of grades would have to fall between the possible limits of the range, or 0 and 100, while the average amount of deviation with respect to their fathers' incomes would be represented by a very much larger amount. Obviously, these two average amounts of deviation would not readily lend themselves to comparative purposes and it is necessary for us to find another method by means of which these average amounts of deviation can be reduced to comparable terms. This is found in the *coefficient of average deviation* and is very simply computed by dividing the average amount of deviation by the average used. In Table 21, for example, the average amount of deviation is 5.6 and the median from which it was estimated is 17.5. Dividing 5.6 by 17.5 we get .32 as the coefficient of average deviation from the median. In the case of the grades and the paternal incomes above referred to, we can easily imagine, within a sample of a hundred cases, the average grade to have been 80 and the average amount of deviation to have been 8. In the case of the incomes, an average income of \$5,000, with an average amount of deviation of \$1,000, would be entirely possible. Now, average amounts of deviation of 8 and 1,000 cannot be compared, but by

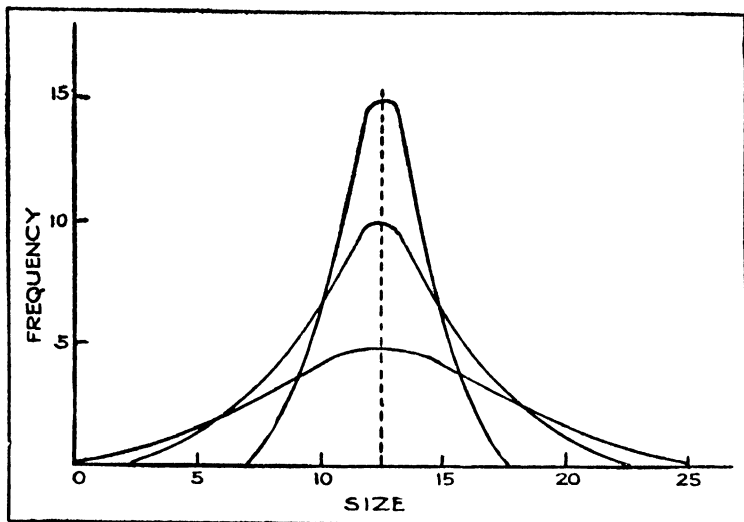


Chart 24: Three frequency distributions having the same arithmetic mean, mode, and median (17.5) but showing widely different dispersion.

reducing each to a coefficient, by dividing by the average used, we obtain results which are comparable. The resulting coefficients are .10 for the grades and .20 for the incomes, indicating that the average deviation for the latter is twice that of the former.

Since it is difficult for the average amount of deviation to be larger than the average itself, the coefficient of average deviation will usually be less than one unit in size.

Quartile deviation. Quartile deviation constitutes an interesting departure from the above method. Instead of finding the deviation of all items from the average used, in quartile deviation but two items of the entire series are employed. These two items, being the one which is one-quarter of the way up and the one which is three-quarters of the way up, are treated as though two items constituted an entire series. We noticed in our discussion of the mean that the formula, $\frac{n+1}{2}$ locates the position of the median

or the middle item. Similarly, the formula, $\frac{n+1}{4}$ will locate the first quartile (Q_1) and the formula $\frac{3(n+1)}{4}$ the third quartile (Q_3). The median is also the second quartile. Thus in the sample distribution which we have been using and which contains 25 items, the first quartile lies at 6.5 and the third quartile at 19.5. Having located these two items it is now possible through the use of the same formula which was used in the case of the median,

$$\left(L + \frac{C(2i-1)}{2f} \right)$$

to discover their size. This is the method of interpolation. This size, in the case of the first quartile, will be found to be 12 and the third quartile to be 23.

We now proceed as if we had a distribution composed of but two items. Naturally, the total amount of deviation is the total amount of difference between the two items, or 11. Dividing this by 2, since there are but two cases, we arrive at 5.5 as the quartile amount of deviation. In order to obtain a coefficient of quartile deviation, it is necessary for us now to divide this quartile amount of deviation by an average. Since there are but two items, 12 and 23, we will add them together, getting 35, and divide this by 2, or the number of cases, getting 17.5 as the result. We are now in a position to divide the quartile amount of deviation, which is 5.5, by our quartile average, which is 17.5, arriving at .32 as a result. This represents our quartile coefficient of deviation. It will be noticed that, by the two methods (average deviation and quartile deviation) we have secured an identical result. This is due to the fact that we are dealing with a normal or equally balanced distribution. In the case of an unbalanced distribution the results would not necessarily be the same, but they would be very similar.

Following is the detailed mathematical process by means of which the above results were obtained:

Numerical rank or location of the two quartiles:

$$Q_1 = \frac{n+1}{4} = \frac{25+1}{4} = 6.5$$

$$Q_3 = \frac{3(n+1)}{4} = \frac{3(25+1)}{4} = 19.5$$

Applying the formula which was used for the determination of the size of the median:

$$L + \frac{C(2i-1)}{2f}, \quad Q_1 = 12 \text{ and } Q_3 = 23$$

$$\text{Quartile Amount of Deviation} = \frac{Q_3 - Q_1}{2} = \frac{23 - 12}{2} = 5.5$$

$$\text{Quartile Average} = \frac{Q_1 + Q_3}{2} = \frac{12 + 23}{2} = 17.5$$

Quartile Coefficient of Deviation =

$$\frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_1 + Q_3}{2}} = \frac{Q_3 - Q_1}{Q_1 + Q_3} = \frac{23 - 12}{12 + 23} = \frac{11}{35} = .32$$

Standard deviation. Another type of deviation is known as the "Standard Deviation." The process of computation is similar except for the fact that each deviation is squared, and, after the sum of the squared deviations has been divided by the number of cases, the square root of the result is extracted. The effect of this squaring process is to give greater weight to extreme variations. It is used in the computation of the Pearsonian coefficient of correlation and of probable error. It may be mathematically expressed by means of the following formula:

$$\sigma \text{ or Amount of Standard Deviation} = \sqrt{\frac{\sum d^2}{n}}$$

The standard deviation, unlike the average deviation, usually is computed only from the arithmetic mean. The coefficient of standard deviation is obtained in the same manner by dividing the amount of standard deviation by the average used.

TABLE 22
SAMPLE DISTRIBUTION FOR THE COMPUTATION OF THE
STANDARD DEVIATION

Arithmetic Mean=17.5

Class	Mid-Point of Class	Frequency	Deviation from Arithmetic Mean	Deviation Squared	Squared Deviation Times Frequency
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
0.0- 4.4	2.5	1	-15	225	225
5.0- 9.9	7.5	3	-10	100	300
10.0-14.9	12.5	5	- 5	25	125
15.0-19.9	7.5	7	0	0	0
20.0-24.9	22.5	5	+ 5	25	125
25.0-29.9	27.5	3	+10	100	300
30.0-34.9	32.5	1	+15	225	225
Total	(Σ)	25			1300

$$\sigma \text{ or } SD = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{1300}{25}} = \sqrt{50.} = 7.07 = 7.1$$

Dividing 7.1, or the standard deviation, by 17.5, or the average from which it was computed, we obtain .40, which is the coefficient of standard deviation. Note that the average deviation for the same series was but .32. The difference is the result of squaring.

C. THE MEASURE OF SKEWNESS

The coefficient of skewness is a method of measuring asymmetry. A frequency distribution may have a wide range of dispersion and still retain absolute symmetry. Being graphed, it might show a normal curve equally balanced on either side but wide and with a tendency toward flatness. A skewed curve or unbalanced distribution, on the other hand, represents one in which the modal class lies to one side or the other of the mid-point of the range. It is a leaning curve. On the one side the rise will be gradual while on the other it will be more abrupt. Coefficients of skewness measure this asymmetry of distribution. As has

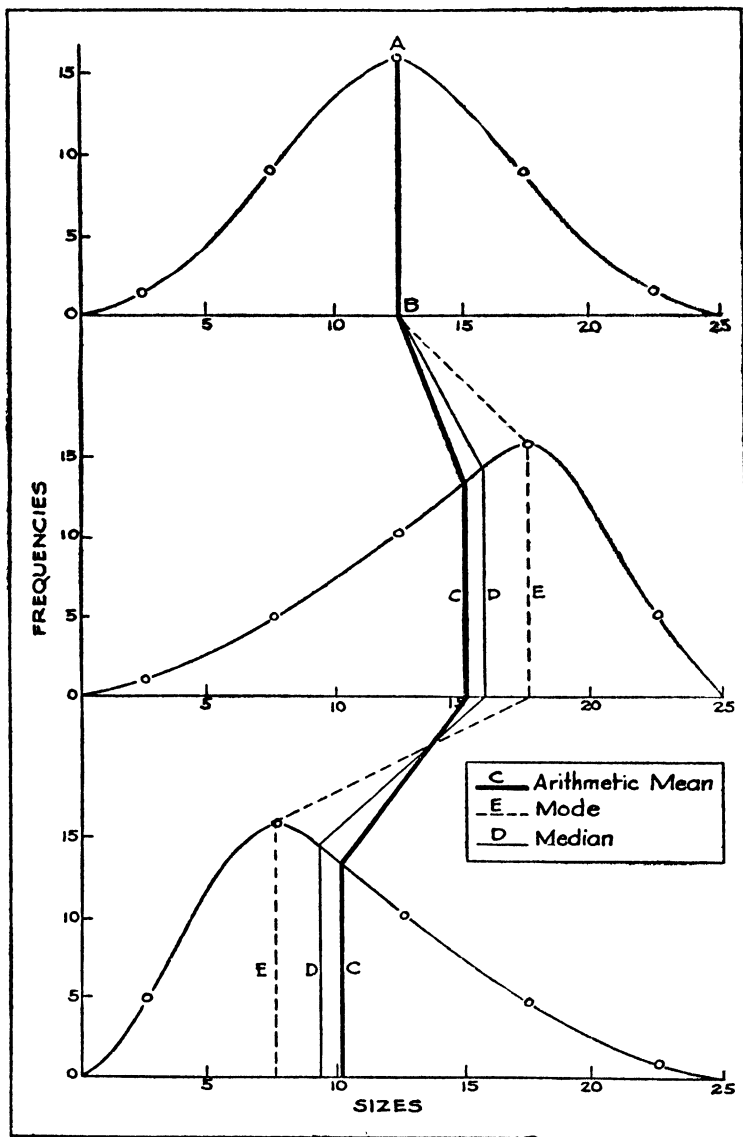


Chart 25: Chart showing the tendency of three averages to separate under conditions of skewness.

already been pointed out, the arithmetic mean, mode, and median in symmetrical graphs are located at the same point on the base line of a frequency graph, but as the graph becomes skewed they separate, the average due to the influence of extreme items moving either to the right or to the left, and the median, due to a lesser influence of these items, moving approximately two-thirds as far as the average moves. Thus the median will always be located between the mode and the arithmetic mean. Based on this fact the following formula has been evolved for the measurement of the coefficient of skewness (Chart 25).

$$\text{The Coefficient of Skewness} = \frac{\text{The Arithmetic Mean} - \text{The Mode}}{\text{The Average Deviation from the Mode}}$$

D. THE COEFFICIENT OF CORRELATION

The object of coefficients of correlation is to discover and to give numerical value to the relationships which exist between paired variables. It is easily observed that there is a definite relationship between height and weight, between the ages of husbands and wives, between the standard of living and the income. Of the various methods of showing this relationship we will here mention only the Pearsonian coefficient of correlation. Here we are dealing with paired items, such as the age of husband and wife, or the relationship between height and weight in the individual. Chronologically considered, we might compare birth-rates and death-rates by years or other periods of time. Between series of paired items correlation may be *direct* or *inverse*. Correlation is said to be direct when both series vary in the same direction. An example of this is the relationship between the ages of husband and wife. Wages also will tend to show a direct correlation to the cost of living, though lagging slightly behind in their movements. Correlation is said to be indirect or inverse when an increase in one is accompanied by a decrease in the other. This is illustrated by the relationship between supply and price. Other things being

equal, when supply is large, price is low, and vice-versa. Correlation also varies as to degree. In comparing the average price of wheat and the morbidity rate, for instance, we would not expect to find a high degree of correlation. Sometimes, however, a high degree of either direct or inverse correlation is observed where no reason for the same is present. The statistician must be careful not to draw sweeping conclusions from merely accidental correlations which may appear. We might easily discover, for instance, a high degree of correlation between the movie attendance and age at marriage, but this does not prove that a causal relation exists.

The following formula measures both the direction and the degree of correlation existing between series of paired items:

THE PEARSONIAN COEFFICIENT OF CORRELATIONS

$$\frac{\Sigma (X_1 \times X_2)}{n \times SD_1 \times SD_2}$$

Σ = The sum of

X_1 = The deviation from the average of the item in the first series

X_2 = The deviation from the average of the corresponding item in the second series

n = The number of paired items

SD_1 = The standard deviation of the first series

SD_2 = The standard deviation of the second series

The above formula might be read as follows: The sum of the products of the deviations of the paired items from the average used, divided by the number of items times the standard deviation of the first series, times the standard deviation of the second series.

It will be noticed from the above distributions that when series "A" and "B" (Table 23) are compared, a direct correlation is observed; that is, as "A" increases in size, "B" increases in size in direct ratio. Comparing "A" and "C" or "B" and "C," on the other hand, we observe a definite indi-

TABLE 23

SAMPLE PAIRED SERIES FOR THE COMPUTATION OF THE
PEARSONIAN COEFFICIENT OF CORRELATION

	A	B	C
	2	3	10
	4	5	8
	6	7	6
	8	9	4
	10	11	2
Arith. Mean	6	7	6

rect correlation. As the one increases in size, the other decreases. These hypothetical series have been purposely chosen to show perfect direct or perfect inverse correlation. Applying the above formula by the following method, let us observe what happens (Table 24).

TABLE 24

SHOWING THE COMPUTATION OF THE PEARSONIAN
COEFFICIENT OF CORRELATION

FIRST SERIES (A)			SECOND SERIES (B)			$X_1 \times X_2$
Item	X_1	X_1^2	Item	X_2	X_2^2	
1	2	3	4	5	6	7
2	-4	16	3	-4	16	16
4	-2	4	5	-2	4	4
6	0	0	7	0	0	0
8	2	4	9	2	4	4
10	4	16	11	4	16	16
Total (Σ)		40			40	40

In columns 1 and 4 we have the paired items. In columns 2 and 5 we have the deviations of the respective items from their arithmetic means. In column 7 we have the product of these paired deviations. The sum of these products gives us the numerator of our formula.

The number of paired items is 5. In columns 3 and 6 we have the squares of the deviations from the arithmetic mean. This is for the purpose of estimating the standard devia-

tion (SD) which is the square root of the sum of the squares of the deviations divided by the number of items. This process gives us the data for the denominator of our formula:

$$\frac{\Sigma (X_1 \times X_2)}{n \times SD_1 \times SD_2} = \frac{40}{5 \times \sqrt{\frac{40}{5}} \times \sqrt{\frac{40}{5}}} = \frac{40}{5 \times 2.83 \times 2.83} = \frac{40}{40} = +1$$

In the above Pearsonian coefficient of correlation $+1$ represents perfect and direct correlation, while -1 would represent perfect and inverse correlation. If a coefficient of less than .30 is obtained, we may conclude that correlation is present in such a small amount as to lack statistical significance. If the amount of correlation is .50 or more, we may consider it marked.

A most excellent study of correlation in the field of social data is that of Miss Thomas in her book, *The Social Aspects of the Business Cycle*. Here she attempts a correlation between the business cycle, on the one hand, and marriage, divorce, prostitution, birth, death, pauperism, alcoholism, crime, and emigration rates on the other. In one chart, for example, representing a period of time extending from 1854 to 1913, she presents three graphs representing the business cycle, the marriage rate, and the divorce rate for the various years of this period. A glance at the chart representing the divorce rate shows little or no similarity at all to the other two which do show marked resemblance. Reducing these to coefficients of correlation, she discovers between the business cycle and the marriage rate a coefficient of correlation of $+.67$ and between the divorce rate and the business cycle a coefficient of correlation of $+.02$. These coefficients would indicate, in the case of the marriage rate, a high degree of direct correlation, and in the divorce rate a negligible amount. We might expect, then, that the sociologists would be warned to look for the reasons for, or the causes of, this relationship in one case, and the lack of it in the other. This she does most effectively in her study.

E. RATES AND RATIOS

To be told that in the year 1900 in the state of Massachusetts there were 51,156 deaths and in 1910 there were 54,407 gives us little information until we take into consideration the change in size of the population during this period. While we have an absolute increase in the number of deaths, we also discover an increase in the size of the population. Reducing these two figures to rates, we find that the death-rate in 1900 was 18.2, whereas in 1910 it had materially reduced to 16.1.

The rate is merely a ratio between two numbers; sometimes a variable and a constant, and sometimes two variables. The variable is represented by the data in hand and the constant, or second variable, by the base chosen. It is a fraction. A per cent, or a decimal fraction, is merely a per one hundred ratio. The denominator or base of a statistical rate, however, may be any magnitude from one on. The per one hundred rate is merely the most commonly used.

The rate is of particular importance to the student of society, especially in the study of social change and national vitality. It is a numerical measure used for comparative purposes where a common base is possible. It is the simple process of giving fractions a common denominator.

Per capita rates. The simplest rate is the per capita rate. For instance, on January 1, 1919, there were approximately 12,000,000 telephones in use in the United States, or, roughly, 1 telephone to each 10 persons, which would be more exactly a per capita rate of .11. Various per capita rates are in use, such as per capita wealth, per capita consumption, per capita production, and so forth. This type of rate, then, is a rate based on unity. It is a per one rate.

Magnified rates. The death-rate above referred to for Massachusetts in 1900 on a per capita basis would be .0182. Since, manifestly, this rate results in numbers too small for practical purposes, we have found the general practice of

basing this rate on the per thousand basis more satisfactory. Rates may be on the bases of 1, 10, 100, 1,000, 10,000, or even 1,000,000 or 1,000,000,000. Population is not the only term, however, in which rates are expressed. Area, distance, time, and units are also used. For instance, in terms of automobiles we might speak of cars per square mile, gallons per 100 miles, miles per minute, and miles per gallon. The number of bases of this sort is unlimited.

Refined rates. The rates preceding are what may be called simple or crude rates. A rate is refined by doubling or by limiting the base, making it representative of a portion rather than the whole. A crude birth-rate would be births in relation to the total population. A refined birth-rate would be births in relation to the number of adult married women of child-bearing age. Death-rates may be refined by taking into consideration the composition as well as the size of the population. Since the death-rate varies with the age composition of the group, we would expect a higher rate to obtain in a population in which the young and old predominate than in a population composed, predominantly, of those in the prime of life. It was once demonstrated through the columns of a newspaper that it was healthier to go to war than to live in New York City. The basis of this statement was the fact that the death-rate among soldiers in the Spanish-American War was actually smaller than the normal death-rate in the city. This condition, obviously, would not have obtained if children and old folks had been included among the soldiers. An illustration of refinement by doubling the base would be the use of the ton-mile and the car-day bases.

Vital rates. The number and variety of rates which may be used by the sociologist is great. He has at his disposal many varieties of criminology rates, illiteracy rates, church attendant rates, and delinquency rates. Most important, perhaps, are those which have to do with the in-

crease and decrease of population, or vitality rates. The first of these is the rate of population increase, which may be on the basis of increase per year, increase per decade, or any other time element. This increase is often stated in terms of unity, but is more comparable as a per cent per year of increase. General crude death-rates, birth-rates, and marriage rates, are usually on the basis of one thousand per year. The accuracy of these vital rates depends upon the accuracy with which vital facts are reported to the government. Census reporters are capable of carelessness and doctors are capable of negligence in the matters of reporting births, deaths, diseases, and so forth. Not all of the states of the union are even in the birth and death compulsory registration area. In a recent study, made by the writer of some 25,000 birth certificates, covering a period of three months in a midwestern state, which is in the compulsory registration area, it was discovered that some counties in the state made the unenviable record of as high as 60% of inaccuracy or incompleteness in the birth certificates returned by doctors.

Again, certain localities in California, Florida, and Colorado might show exceptionally high death-rates due, not to the fact of an unhealthy climate or of unsanitary living conditions, but to the very favorableness of the environment of these communities which attracts the aged and the ailing who have come there either to regain health or to die. A corrected rate is therefore a rate which takes into consideration only *bona-fide* residents and not transients. The commonly used crude birth-rate is of less value than the more refined fecundity rate which divides births into legitimate, illegitimate, and still births, and divides the base into women of child-bearing age and married women of child-bearing age. The marriage rate and the wedding rate must be differentiated. The wedding rate will be but one-half of the marriage rate. Similarly, we would have to distinguish between the divorce rate and the broken family rate. Death-

rates may be made more specific by being related to age, group, sex, or cause of death. Modern insurance has been made possible because of the specific death-rates which have been worked out for each of the age periods of life. The death-rate graph, interestingly enough, represents the form of a U, high in infancy, lower during the middle period of life, and high again as old age approaches. The most useful death-rate is what may be called the *adjusted* death-rate, in which age adjustment is made for the composition of the population. For instance, the crude death-rate for the United States in 1900 was 17.55, but when the population had been adjusted on the basis of a hypothetically normal population, the death-rate was found to be 18.05, or .50 higher. This is probably due to the fact that our population was growing at that time, not only through genetic increase but through the influx of immigrants from Europe who were in the prime of life. Adjustment may also be made for race, occupation, and climate, but this is such a complicated and difficult process that only the bare beginning has been made in the standardization of the adjusted death-rate method. A morbidity rate is a sickness rate, and may be either general or specific. In the latter case, the basis is the particular disease from which individuals may suffer. The international list of the causes of sickness and death, of 1920, included more than 200 items. The *death-rate of infants* is the same as the specific death-rate of all individuals under one year of age. It is expressed in terms of per thousand of the total population. The *infant mortality rate*, however, is generally understood as being the relationship between the infant deaths per year and the number of births during the same year.

Finally we may mention increase rates; a general increase rate would include both births and accretions from the outside in the form of immigration. A natural or genetic increase rate, however, would show the relationship between births and deaths. This is sometimes called "excess" rate,

since it demonstrates the growth of the population by means of the excess of births over deaths. In the brief space of this study it is impossible to canvass the wide range of vital rates which are of interest and use to the sociologist. The foregoing, however, will serve to orient the field. For the most nearly complete presentation which has yet been made the reader is referred to Whipple's *Vital Statistics*.

Rates have been widely and effectively used in the field of social statistics. They are to the student of population movements what index numbers are to the economist. One word of warning must be given, however. Rates must not be used for comparative purposes unless their respective bases are really comparable.

F. INDEX NUMBERS

An index number is a percentage comparison of chronological variables by means of a common chronologically determined base. Rates are easily converted into index numbers but, unaltered, serve much the same purpose. They have been more widely used by the economist than by the sociologist. If we wish to study fluctuation in the prices of wheat, we must, first of all, collect the prices or the average prices of wheat for consecutive periods of time. The beginning of the period, sometimes the end, or some period between, which is for some reason strategical, is used for the base. For instance, the most common basis in recent studies is 1913. The reason for choosing 1913 is that it is the last year of normal economic conditions preceding the war. Having chosen the base year or the base period, the price of the commodity studied—wheat in this case—is taken as 100%, and prices in other years are designated in terms of this base or 100% year. If, for instance, the price of wheat in 1913 had happened to be 50 cents and the following year \$1 per bushel, their respective index numbers would be 100 and 200. The object of index numbers is to make series of

variables comparable in spite of differences in magnitude. In standard of living studies index numbers have been made of the fluctuations in the prices of the various, most representative articles of consumption. These are averaged together to produce a new index number representing the fluctuation of the whole, and this index of indexes gives a fairly satisfactory picture of the general movement of the whole. One reason why the sociologist has not made more use of index numbers is because of the wide use which he has made of rates. There is no reason, however, why we should not have indexes of social change as well as indexes of cost and price. Another index of a different nature is what Dr. Raymond Pearl has called the *Vital Index*. This is merely the ratio between births and deaths on the one hundred basis. If the index is exactly 100, the population is standing still; if it is larger than 100, it is growing; or if it is smaller than 100, it is decreasing in size. One probable factor in the situation which led to the recent immigration legislation was the realization of the fact that the vital index for whites of native parentage was less than one-half of that of whites whose parents were foreign born.

G. PROBABLE ERROR AND NORMAL DISTRIBUTION

Since samples at best but imperfectly represent universes, mathematicians have devised methods for the measurement of the probable error which is present in computations which are made on the basis of samples. But in the first place "errors must not be confused with blunders or mistakes. The latter arise from carelessness or incompetency in transcribing figures or reading values from a scale. The only remedy for these imperfections is the exercise of great care in the observation and handling of figures, and the use of checking devices. . . . There is no such thing as absolutely exact measurement. . . . Furthermore, measurements of the same object repeated with the

greatest care will not yield the same results." An error, then, "is the difference between the observed value and the most probable value. Therefore, error means a deviation, not a mistake."¹

A measure of error is a measure of the amount of departure from the normally expected frequency distribution. This normally expected frequency distribution has been variously called "The Normal Frequency Curve," "The Probability Curve," and "The Normal Curve of Error." It is a symmetrical, bell-shaped curve, the points on either side of which exactly correspond. Every frequency distribution represents either a straight line or a curve. In case each frequency class contains exactly the same number of items its graph will be a straight line which runs parallel to the horizontal axis or the base. If, on the other hand, the first and lowest class starts with a small frequency or a small number of items and each succeeding class is represented by a regularly increasing number of items, the resulting graph will also be a straight line which proceeds from the base line at a definite angle. Usually, however, no such arrangement of items is present. The lower and the higher classes are apt to contain the smaller frequencies or numbers of items, and the larger frequencies are apt to be located either exactly in the middle or skewed to one side or the other. In multi-modal data there may be more than one definite point of frequency concentration. In frequency distributions, however, which are based on purely chance occurrences or on physical measurements, the normal distribution above referred to is very apt to be approximated. In distributions of social data this symmetry of distribution either may or may not be present. It usually is present, but it is entirely possible that either the skewed or the multi-modal distribution may be normal to the particular data in hand. Care must be taken, then, not to carry too far into the field of the treatment of social data, reasoning which is

¹ Chaddock, *Principles and Methods of Statistics*, pp. 208-209.

based on the assumption of a normal frequency distribution.

In any curve representing a frequency distribution the median line divides the area included in the curve into two equal halves and similarly divides the frequencies. In like manner, the quartiles divide the frequencies into three portions, 25% of them being below the first quartile, 50% between the two quartiles, and 25% above the third quartile. From this we may conclude that there is a 50-50 chance that any item in the series will fall between the limits of the two quartile measurements. In the tossing of coins there is a similar 50-50 chance that either heads or tails will appear. If we toss once the chances are even that heads will be turned; if we toss twice we may get two heads or two tails, but the chances are that we will get one of each. In tossing four times a three to one ratio would not be surprising, but it would be very exceptional if we were to obtain that ratio as the result of a thousand tosses. In other words, the larger the number of trials or the larger the sample the more closely may we expect the results to approximate the normal probability. A single toss of the coin is not an isolated fact; it is related to that universe or family which is composed of all of the possible tosses which might be made. In like manner, the frequency distribution based on a sample of a given variable is not an isolated case but it is related to the universe of the variable in question and may be expected, within certain limits, to show the characteristics of that universe. Chaddock, for example, records an interesting experiment. The heights of 1,000 students were taken; from these 1,000 items 100 samples of 100 each were selected at random. Next the arithmetic mean of each of these samples was computed and it was discovered that the total range exhibited a dispersion of only $1\frac{1}{2}$ inches, a direct indication of the representativeness of the sample and the adequacy of its size.

It is the probable amount of this error due to the inability of any sample exactly to represent its universe that we wish

to measure. The concept is based upon the expectation that a normal frequency distribution represents the universe.

The basis of most probable error formulas is the decimal .6745 since this proportion of the standard deviation measured to the right and to the left of the central ordinate includes just one-half of the total area. The probable error of any distribution is .6745 of the standard deviation. As applied to the arithmetic mean it would be $.6745 \frac{SD}{\sqrt{n}}$

For instance, in Table 21, page 126, we present a hypothetical frequency distribution. The arithmetic mean for this distribution is found to be 17.5. The size of the sample or the total frequency is 25. The computation of the standard deviation by the method there explained is 7.1. Applying the above formula for the measurement of the probable error of the arithmetic mean we get the following results:

$$\begin{aligned} .6745 \times \frac{SD}{\sqrt{n}} &= .6745 \times \frac{7.1}{\sqrt{25}} = .6745 \times \frac{7.1}{5} = .6745 \times 1.42 \\ &= .957 = .96, \quad \text{or } P. E. \text{ mean} = \pm .96 \end{aligned}$$

This amount, .95, is the amount by which the average found, 17.5, is in probable error. $17.5 \pm .95$ means that this average of a sample of 25 cases may be expected to vary from the universe it represents within these limits. The chances are equal that the true average lies within the limits of $17.5 - .96$ or 16.54, and $17.5 + .96$ or 18.46. Likewise, another sample of 25 cases from the same universe would give an arithmetic mean which would have a "50-50" chance of falling between these limits.

Additional formulas for measuring the probable error of other statistical measures are available. This one, however, is sufficient to familiarize the reader with the term "*P. E.*" and the sign $\pm$ when he meets them. Ideally, the probable error of a universe would be zero and increases in amount as the size of the sample (and hence its adequacy) decreases.

IX

GRAPHIC AND TABULAR SUMMARIZATION

The moving average. The straight-line trend (line of least squares). Short-time oscillation. The correlation or double-entry table; the scatter diagram. The Lorenz curve. The ratio or logarithmic chart.

THE last two chapters have dealt with methods of characterizing groups for comparative purposes by means of numbers. The present chapter discusses methods whereby summarization may be further accomplished by graphic and tabular methods. Our discussion of this vast field will of necessity be brief. As yet its application to socio-statistical facts has been limited, but the economist has used it extensively, especially in the study of business cycles and the development of business barometers. There is little reason why waves of public opinion and recurring periods of emotionalism should not be studied in the same manner with resulting social welfare barometers.

A. THE MOVING AVERAGE

The moving average is a refinement of the numerical averages which we have already considered. It is a progressive average and is used only in connection with time series to show progressive average movement. It is used to the best advantage where a definitely recurring periodic increase and decrease in a given variable is observed. There is, for instance, a definite relationship existing between certain types of crime and the season of the year. Temperature and humidity seem to affect the individual proclivity in this direction. Crime, in general, may be on the increase or on the decrease due to quite different causes, but whether

on the increase or on the decrease this periodic seasonal fluctuation is a constant factor. Crime for a given community, plotted as a graph in the form of historical series by months or by seasons, will show a definite wave-like movement. Similarly, the domestic consumption of electricity in

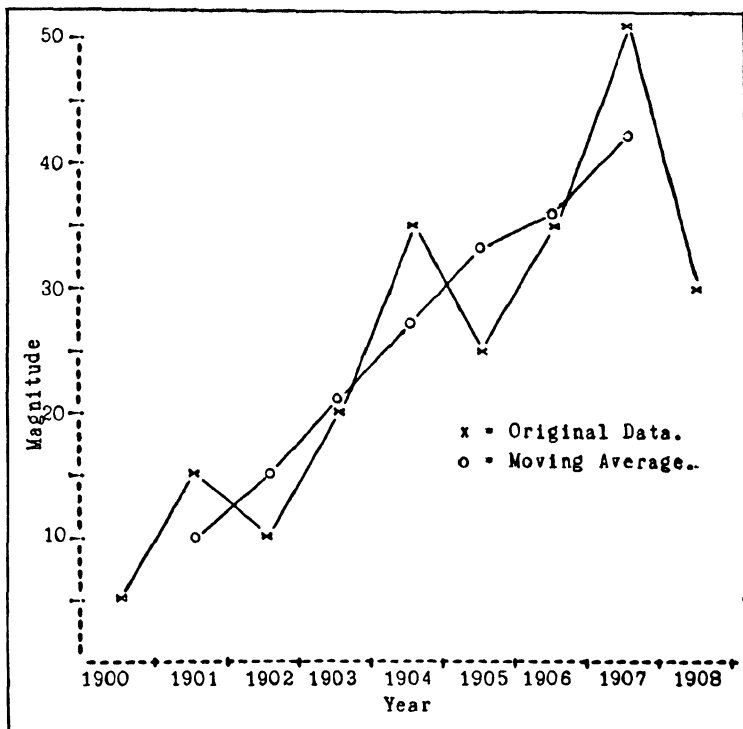


Chart 26: The sample series and its moving average.

a given community will be affected by the length of the day, more in winter and less in summer. Having plotted the actual numerical data, by lesser divisions of a given period of time, we first observe the average length of time which elapses between the high points of the curve or between the low points of the curve. In case the data are plotted by months, we can well imagine a wave-like movement which

TABLE 25
DETERMINATION OF THE MOVING AVERAGE

Year	Data	Moving Average 3-Year Cycle
1900	5	
1901	15	10
1902	10	15
1903	20	21.6
1904	35	26.6
1905	25	32.6
1906	35	36.6
1907	50	42.6
1908	40	

would require 1 year or 12 months to complete. The crest of the wave might, for example, as in the case of the consumption of eggs, be in the summer-time and the trough in the winter. On the basis of a given series plotted by months for a period of years in which the periodic movement is 12 months in length, we will derive a moving average by adding the first 12 months together and dividing by 12. This average for the 12 months will be plotted at the mid-point of the year. Next, the first January will be dropped off and the second January added on, giving us a new series of 12 months which is one month in advance of the first series or extending from February 1 to January 31. In like manner these 12 months are averaged and the new average plotted at the mid-point of this year, or one month in advance of the last point plotted. Continuing in this manner, dropping off each time the first month of the series and adding on the next month, we will locate a series of points which will approximately bisect the curve or graph represented by the data of the individual months of the entire series. This curve will show the general trend of the series without all of the seasonal fluctuations which are present in the original data but not entirely eliminating them.

The moving average is used primarily for the purpose of eliminating seasonal (or monthly) fluctuations. It can also be used to show cyclical (or yearly) trends but not to the

same advantage since periodicity in this case is not apt to be as regular. For example, the yearly immigration into the United States, reduced to a moving average for a 10-year period, very effectively shows the general trend of immigration without all of the violent fluctuations which are present when the original data are plotted. A further refinement of this method is, first, the reduction of the original data to index numbers, plotting these, and from these, plotting the moving average. Percentage series and rates may be treated in the same manner.

B. THE STRAIGHT-LINE TREND

While the moving average gives us a good general picture of the trend of the data concerned, it is capable of entirely eliminating periodic variations only when these periods of variation are of identical length. The moving average also has the disadvantage of resulting in a curve or graph which does not represent the entire length of the series. In the illustration which we have used in connection with our discussion of the moving average, the first point in the moving average was at the mid-point of the first three-year period or two years up from the beginning of the data. Likewise the last point plotted is two years from the end of the period. The moving average, further, shows us the actual trend of increase or decrease. We are now in need of a further method which by means of a straight line for the entire period will indicate not actual but normal growth.

This is accomplished by two methods: the first is that of mere inspection and fitting to a graph a straight line which seems to indicate its general trend. In doing this, exceptional and violent fluctuations would be disregarded and the straight line would be fitted to that portion of the curve which seems to represent normal conditions or movements. For all practical purposes this is often satisfactory, but a more exact mathematical method is that of the line of least squares. By this is meant a series of points resulting in a

straight line, in which the sum of the squares of deviations of the actual data from this line are reduced to a minimum. This is quite easily accomplished by locating two points and drawing a straight line through them. The first point will be the point representing the arithmetic mean of the data

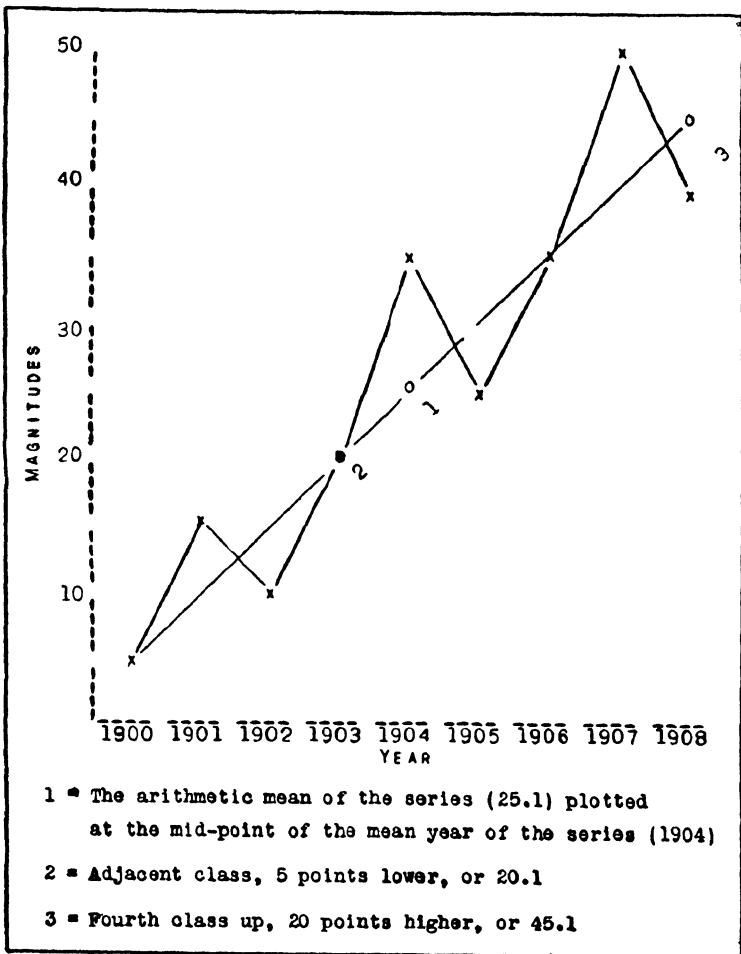


Chart 27: The sample series and its line of least squares.
(Straight-line trend.)

for the entire period. It will be used as the mid-point of the series. The formula for this point is the $\frac{\Sigma m}{n}$ in which m represents the original data and n the number of cases or the number of time periods represented. The second point is represented by the formula the $\frac{\Sigma xm}{\Sigma x^2}$. In this x represents the deviation of any given year from the median year. Thus in a 5-year period the third year will be the median year and the first year will represent a deviation of -2 , the second year a deviation of -1 , the third a deviation of 0 , the fourth again of $+1$ and the fifth of $+2$. x^2 is the square of the respective deviations and xm is the product of the deviation for a given year and the data represented by that year. The numerical result thus obtained represents the rise on one side and the fall on the other, for the two years imme-

TABLE 26
DETERMINATION OF THE LINE OF LEAST SQUARES FROM
A SAMPLE DISTRIBUTION

Year	Deviation from the Median Year	Deviation Squared	Magnitude	Magnitude Times Deviation
66	x	x^2	m	$x \times m$
1900	-4	16	5	- 20
1901	-3	9	15	- 45
1902	-2	4	10	- 20
1903	-1	1	20	- 20
1904	0	0	35	0
1905	1	1	25	25
1906	2	4	35	70
1907	3	9	50	150
1908	4	16	40	160
Total (Σ)		60	235	- 105 + 405 + 300
$\frac{\Sigma m}{n} = \frac{235}{9} = 25.1$		$\frac{\Sigma xm}{\Sigma x^2} = \frac{300}{60} = 5$		

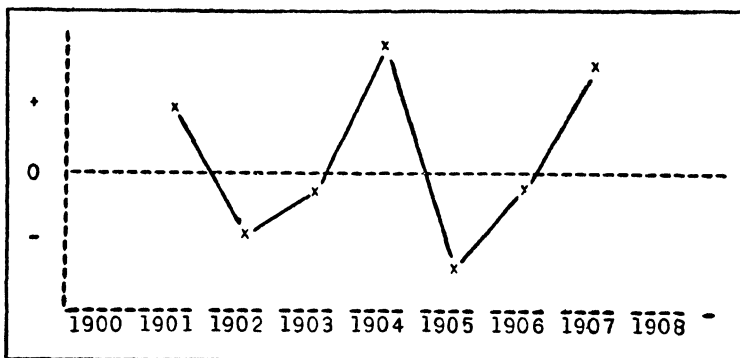


Chart 28: The same sample series showing the cyclical variation only. The general upward trend has been eliminated.

diately adjacent to the median year. While these points are sufficient to locate the straight line, the value of any given year can be easily ascertained by multiplying this amount by the deviation for the year and adding to the arithmetic mean which we have taken as the mid-point of our straight line period. Other and more highly complicated methods have been evolved by the economist for accomplishing this general result in connection with economic data. Special and similar technique will have to be developed ultimately by the sociologist for the treatment of his data.

The first point in the line of least squares will be 25.1, which will be the point plotted for the median year, or 1904. Five, as above derived is the amount by which the two points adjacent vary above or below this median point. The exact location of the 1900 point will be $25.1 + (5 \times 4) = 45.1$. Five is multiplied by four in this step because 1900 is four years up from the median year.

C. SHORT-TIME OSCILLATION

Where it is desired to show, not the general trend, but only the seasonal or cyclical variation or short-time oscillation, the actual deviation of each item from the moving

TABLE 27
DETERMINATION OF SHORT-TIME OSCILLATION FROM
A SAMPLE DISTRIBUTION

Year	Magnitude	Moving Average	Deviation from M. A.
1900	5		
1901	15	10	+5
1902	10	15	-5
1903	20	21.6	-1.6
1904	35	26.6	+8.4
1905	25	32.6	-7.6
1906	35	36.6	-1.6
1907	50	42.6	+7.4
1908	40		

average or line of least squares is plotted above and below a centrally located horizontal "o" line. Minus deviations are plotted below the line and positive deviations above.

D. THE CORRELATION OR DOUBLE-ENTRY TABLE

The correlation which we have already discussed, from the numerical standpoint, may also be expressed either in tabular or in graphic form. The double-entry table is merely a two cross-classification table in which paired items are entered in terms of both captions and stubs. The resulting distribution of items over the face of the table merely indicates in a fairly general manner whether or not correlation is present. If the items are equally scattered over the face of the entire table, there is no evidence whatsoever of correlation. But, if the result is a diagonal grouping from the upper left-hand to the lower right-hand corner of table, or from the lower left-hand to the upper right-hand corner, we may be sure that correlation is present. Perfect correlation would be represented by a straight line of diagonal items without scatteration on either side. Table 28, which is based on the ages of 175 pairs of grandfathers and grandmothers of university students, is self-explanatory. Since the band of concentration is toward the lower left-hand corner, direct correlation is indicated. Sometimes, however, as

table. Dots are placed in the squares in place of numbers and the trend is thus observed. Scatter diagrams are of two types, in the first an individual dot is located for each pair of items, whereas in the second each dot represents an average.

The scatter diagram shown in Chart 29 is arranged from the lower left-hand corner. The concentration of dots

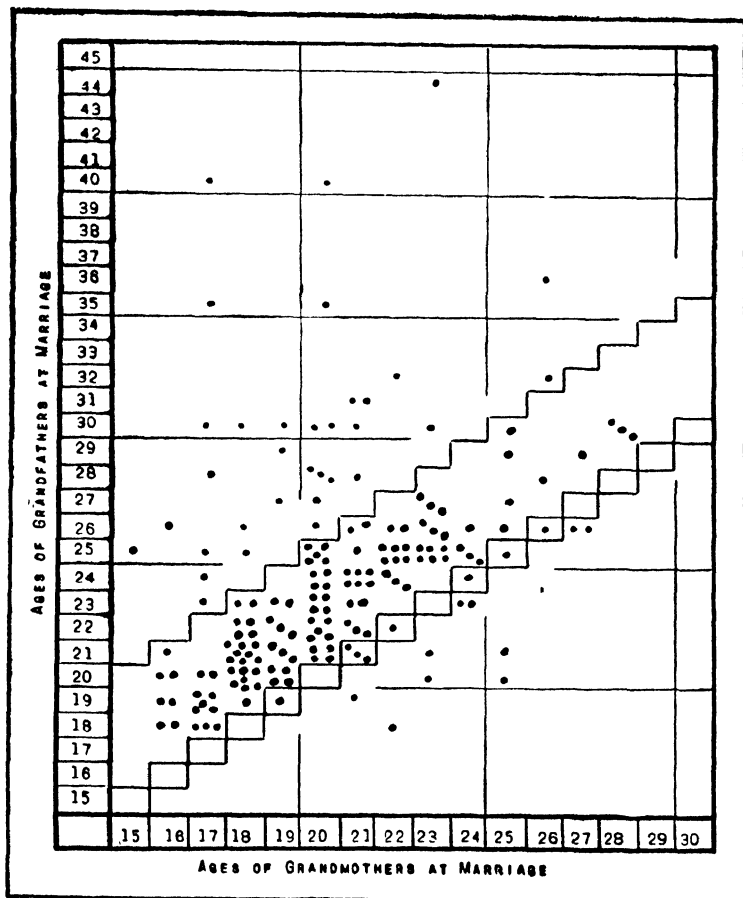


Chart 29: Scatter diagram based on the correlation table shown on page 154.

in a band extending upward and toward the right indicates direct correlation between pairs. The diagonal line of squares are those in which both parties were the same age at marriage. All dots below this line of squares represent cases in which the woman was older than the man; and all those above, those cases in which the man was older than his wife. The latter condition markedly predominates. The stair line seems to locate the upper limit of the correlative massing. In all cases above this line the husband is at least 6 years older than his wife. Correlation is here definitely present but is reduced by the large number of scattered cases above the stair line.

E. THE LORENZ CURVE

The purpose of the *Lorenz Graph* is similar to that of the coefficient of dispersion. It graphically represents the amount of deviation by which a distribution departs from equal distribution. It is of particular value in showing the unequal distribution of wealth, income, and so forth. It is another double-entry system in which points, which have previously been reduced to a cumulative percentage basis, are located on a chart. On this type of chart, absolutely equal distribution is represented by a straight line which extends at an angle of 45 degrees from the point of origin. All

TABLE 29
SAMPLE DISTRIBUTION FOR THE ESTIMATION
OF THE LORENZ CURVE

CLASS	CUMULATIVE PERCENTAGES		
	Of the Simple Frequency	Of the Total Size	
		Series <i>N</i>	Series <i>X</i>
0.-19.9	20%	20%	3%
20.0-39.9	40	40	8
40.0-59.9	60	60	20
60.0-79.9	80	80	40
80.0-99.9	100	100	100

graphs plotted on this chart will have the same point of origin and the same point of termination. Deviation is indicated by the amount of curve by which they deviate from this straight line.

It is obvious that if 20% of the population is in possession of 20% of the wealth, 40% of 40%, 60% of 60%,

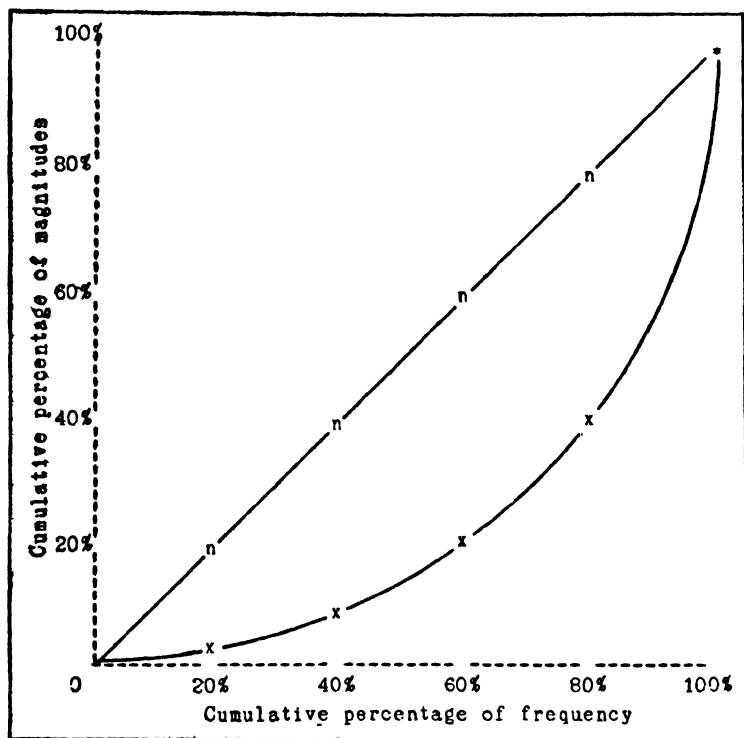


Chart 30: Lorenz curve, based on sample distributions.

and 80% of 80%, on the graph, we will have points plotted which will produce a straight line. But if 20% of the population is in possession of but 3% of the wealth, 40% in possession of 8%, 60% possessing 20%, and 80% possessing 40%, then we will have a curve, the second curve on the

chart, which represents a large amount of deviation from equal distribution.

F. THE RATIO OR LOGARITHMIC CHART

The ratio or logarithmic chart, already mentioned in the chapter on Graphic Presentation, is also a method of summarization in that it represents proportional rather than absolute change. It is not necessary either to understand or to use logarithms in the preparation of ratio charts since

TABLE 30
SAMPLE DATA FOR THE ILLUSTRATION OF THE RATIO CHART

Arithmetic Increase	Geometric Increase (100%)	Constant Increase (10%)	Increase by Increasing Percentage
A	B	C	D
5	1	10	4.6
10	2	11	4.7
15	4	12.1	5.0
20	8	13.3	5.6
25	16	14.6	6.3
30	32	16.1	7.5
35	64	17.7	9.5
40	128	19.5	13.5
45	256	21.4	19.0
50	512	23.6	37.0

Notice the widely differing results obtained by subjecting the above data to arithmetic and logarithmic treatment. On the ratio chart (page 159) line "A" represents a decreasing ratio, line "D" an increasing ratio, and "B" and "C," constant ratios of increase. In the case of the latter two lines a comparison of the amounts of vertical rise for the same time period compares also their rates of increase.

logarithmically ruled cross-section paper is available for that purpose. The advantage of the use of this paper is that while the graph shows proportional change it is numerically read in terms of the original data. The ratio chart shows proportional changes between years and makes easy the comparison of proportional change between periods of

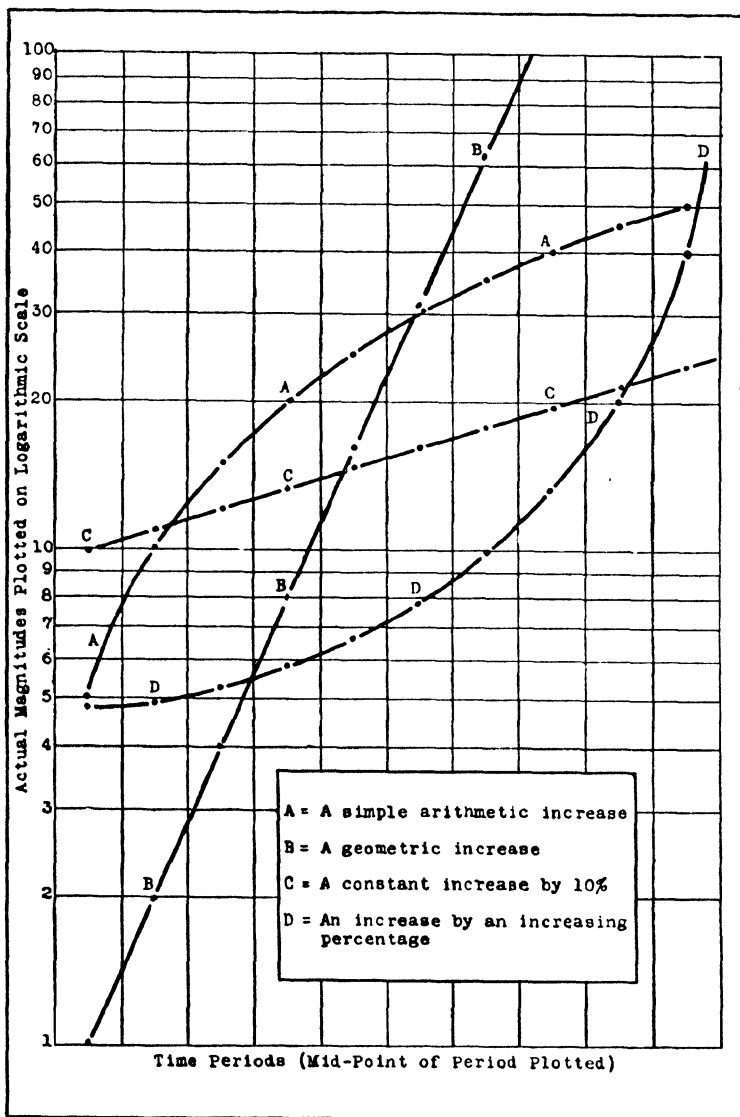


Chart 31: Sample distributions A, B, C, and D (Table 30, page 158). Magnitudes plotted on logarithmic scale. Compare with Chart 32.

time. The ratio chart is a telescoping chart. The amount of vertical space on the chart is the same for 10 units, 100 units, 1,000 units, and so on, indefinitely. Likewise, the

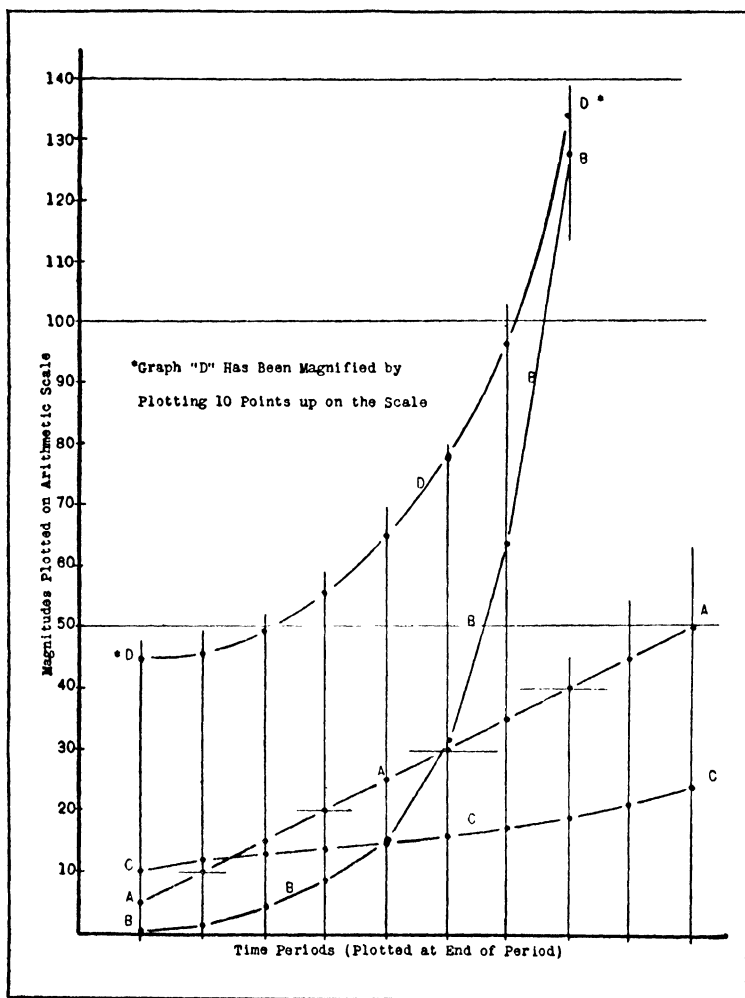


Chart 32: Sample distributions A, B, C, and D (Table 30, page 158). Magnitudes plotted on arithmetic scale. Compare with Chart 31.

first unit in a series of 10 is *ten* times the size of the last unit.¹

The ratio chart has no zero line. A straight line (of necessity diagonal) represents a constant proportional rate of change. On ordinary cross-section paper this line would be represented by an upward trending curve. If, on a ratio chart, a graph is an upward trending curve, the rate of change is increasing. If, on the other hand, a graph has a downward trend, the rate of change is decreasing. Thus it may be concluded that equal slopes will indicate equal rates of change whether or not the two graphs coincide on the chart. This makes it possible to compare the rates of change in two series of variables which differ widely in magnitude. For purposes of comparison it is often desirable to bring graphs on the ratio chart into as close proximity as possible. This is done by partially disregarding the vertical scale. The unit 3 on the vertical logarithmic scale, for example, may denote a value of 30 for one series and 30,000,000 for the other. Still the resulting graphs show us comparable proportional change, and given the unit of expression for each series, we can easily read the original numerical values of the points plotted. This is called vertical shifting.

Charts 31 and 32, plotted on the logarithmic and the arithmetic bases, illustrate the process.

¹ Simply stated, a logarithm is the power to which 10 must be raised to obtain a given number. The logarithm of 10 is 1, of 100 is 2, of 1,000 is 3, and of 10,000 is 4. Likewise the log of 1 is 0, of 5 is 0.69990, of 20 is 1.3010, and of 50 is 1.6990.

INDEX

INDEX

A			
Accuracy, statistical	22 ff.	Complete cross-classification tables	64
Adding machine, use of	24	Continuous and cross-section classifications	36
Adjusted rates	140	Continuous series	25
Age of statistics	4	Correlation	
Angular graphs or frequency polygon	90 ff.	coefficient of (Pearsonian)	134
Arithmetic mean	113 ff.	double-entry table	153
Approximations, scientific treatment of	23	scatter diagram	155
Area charts or pictograms	78	Course in statistics,	
Array or ranged data	65	suggestions regarding	iii
Aschenwall	4	Criteria of	
Asymmetrical curve or skewness	131	good graphic presentation	105 ff.
Averages	112 ff.	good tabular presentation	53 ff.
arithmetic mean	113	Cross-hatching	83
average deviation	126	Cumulative and compensating errors	26
median	119	Cumulative frequency	99
mode	116	Cumulative frequency table	69
moving average	146		
science of	112		
B		D	
Bailey's analysis of the field of social statistics	15	Decimal point, the location of	27
Bar charts	86	Definition of statistics	7, 22
Beginnings of statistics	4	Derivation of the term "statistics"	4
Bias, personal	28, 35 ff.	Deviation, measures of	125 ff.
Booth	13	average	126
Boxes, captions, and stubs	58	quartile	128
Brinton, W. C.	108, 110	range	125
Business statistics	17	standard	130
C		Discontinuous series or discreet variation	25
Captions	58	Dispersion	125 ff.
Cartogram of statistical map	82	Dot designations	82
Census	4, 5, 16	Dot or scatter diagram	155
David's	5	Double entry or correlation table	153
early	4, 5		
United States	23, 30, 33, 37, 60, 68	E	
Chaddock	22, 143, 144	Editing	28
Chance	142	Educational statistics	20
<i>see also</i> Probable error		Engel	35
Chapin	7	Enumerators	
Charity and statistics	12	investigation by	35, 30
Charts	77 ff.	selection of	44
bar	86	Error	
cartogram	82	definition of	23
circle and sector (pie)	84	probable	142
criteria of	105	sources of	26
line graphs	78	Exactitude	29
pictogram	78 ff.	Excess rates	140
types of	16, 88, 99		
China	16, 88, 99	F	
Chronicles 1, 21	5	Fictitious accuracy	26
Chronological series	66	Field of socio-statistical fact	11 ff.
Circle and sector or pie charts	84	Frequency distribution	68
Class limits and intervals	67	Function of statistics	8
Classification			
bases of	66	G	
types of	64	Gathering the data	30 ff.
Code	55	Gaunt, John of	5
Coefficients of correlation	133	General and special purpose tables	70, 71
dispersion	125 ff.	Geographic series	66
		Government and statistics	7
		Graphic and tabular summarization	146
		Graphic presentation	
		criteria of	105

standard rules for types of	110	Philosophical method	3
	77 ff.	Pictogrammatic presentation	78
		Pin maps	84
Halley, Edmund	5	Political Arithmetic	4
Histogram or frequency graph	98 ff.	Preliminary survey and plans	34
Historical or chronological series	66	Primary or original statistical investigation	34
History of statistics	4	Probability curve	143
Horizontal bar chart	86 ff.	Probable error	142
		Psychology and statistics	9
		Purpose of statistics	6
I			
Importance of statistics, causes of present	6	Qualitative series	66
Increase rates	140	Quantitative series	67
Index numbers	141	Quartile deviation	128
Inertia of large numbers	31	Questionnaire	35
Infant mortality rate	140	Questions, schedule	36
Insurance	5		
Intellectual integrity	10		
Interpolation	117 ff.	R	
Investigation	35	Random selection	31, 32
		Range	125
Jerome	7	Rates	137 ff.
		adjusted	140
		corrected	139
Karsten, K. G.	108	crude	139
King, W. I.	38, 77	increase	140
		magnified	137
		per capita	137
		refined	138
		vital	138
Law of samples	31	Ratio or logarithmic chart	99, 158
Least squares, line of	149	Relative nature of statistics	22
Lettering, choice of	59	Requirements for success in statistical practice	10, 22 ff., 29
Limits of classes	67	Round numbers, the use of	24
Limits of statistics	8	Ruling, choice of	59
Line graphs	98		
Lines, choice of types of	59		
Logarithmic or ratio charts	99		
Lorenz curve	156		
		S	
M		Sampling process	30-32
Magnified rates	137	Scatter diagram	155
Mail questionnaires	35	Schedule	36 ff.
Maps, statistical	82	amateur examples of	45 ff.
Mathematical statistics	20	form	37
Mayo-Smith, analysis of the field of socio-statistical fact	13	instructions	40
Mean, arithmetic	113	letter	38
Mechanical aids to statistical work	23	of the National Tuberculosis Association	39
Mechanical tabulation	71	Seasonal variation	146
Median	119	Secondary sources	33
Medical statistics	19	Secrist	17, 64
Mode	116	Secrist's analysis of business statistics	17
Moving average	146	Secular trend or line of least squares	149
Multiplication, mechanical	24	Series, continuous and discontinuous	25
		Series, types of	66
N		Short-time oscillation (trend)	152
Normal distribution	142	Simple tables	64
Numerical summarization	112 ff., 125 ff.	Skewness, the measure of	131
		Slide-rule, the use of	24
		Smoothing	99
Ogive, or cumulative frequency	99	Social Bookkeeping	7
One-man or single survey	35	Social progress and statistics	6
		Social sciences and statistics	3, 28
		Social statistics	8
		Sources of error	26
Partial or special purpose cross-classification	71	Squares, line of least	149
Pearl, Raymond	19, 142	Standard deviation	130
Pearsonian coefficient of correlation	134	Standard of accuracy	28
Per capita rates	137	Statistical fact, field of	11 ff.
Personal investigation	35	Straight-line trend	140
		Stub	5

Summarization				U	
graphic and tabular		146 ff.	Unit		
numerical	112 ff.,	125 ff.	of expression		58
Sussmilch, Johann Peter		9	statistical		58
	T			V	
Tables			Vertical bar chart		86 ff.
location of		63	Vital index		142
mechanical form		58	Vital rates		138
types of	53 ff.,	64	Vital statistics		19
Tabular aids to statistical work		24		W	
Tabulation		53 ff.			
Tallying key		55	Ward, Lester F.		6
Tallying, methods of		55	Weighted average		123
Tally sheet		55	West, Mathematical Statistics		20
Theory of probability	142 ff.		Whipple	19, 31, 33,	141
Titles		57	Whipple's analysis of vital statistics		19
Totals			Whitney, Jessamine S.	40,	71
accuracy in		27	Work-sheet		54
location of		62			

