TEXT FLY WITHIN THE BOOK ONLY

UNIVERSAL
LIBRARY

OU_160680

UNIVERSAL
LIBRARY

STATISTICAL ANALYSIS

STATISTICAL ANALYSIS

EDWARD C. BRYANT

Professor of Statistics
Head, Department of Statistics
University of Wyoming

McGRAW-HILL BOOK COMPANY, INC.
New York Toronto London
1960

STATISTICAL ANALYSIS

Copyright © 1960 by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers. *Library of Congress Catalog Card Number* 59-15045

II

08615 THE MAPLE PRESS COMPANY, YORK, PA.

To Virginia

PREFACE

This book presents the subject matter of statistical methods with emphasis on managerial decision making. An effort has been made to bridge the gap between the predominantly descriptive statistics textbook and the highly mathematical book in "decision making."

With increasing emphasis on quantification of management and decision making based on probability, it is essential that the management student have some careful training in statistical inference. This book is designed to fulfill this need for the student who does not have the opportunity to prepare himself to study decision making from the mathematical approach. Also, it will be found useful as a first course for students who may later have this opportunity.

The first seven chapters constitute a teaching unit which ordinarily can be completed in one semester. These chapters must be studied in sequence. Chapters 8 through 13 present special topics which can be studied in any order desired. It is assumed that the instructor will choose topics from these chapters to complete the year's course in statistics.

Emphasis throughout is on brevity of text and carefully chosen examples and exercises which illustrate the point of the discussion. Exercises are inserted in the text at the point where they seem to have the most meaning to the student. They should be thought through, if not worked completely, as the text is read.

The subject matter of the first seven chapters has served as the foundation course in statistics for sophomore business students at the University of Wyoming. Later chapters have been taught to sophomore classes on an experimental basis. Reception has been enthusiastic because the statistics introduced here "does something."

No mathematical sophistication beyond college algebra is required. For the most part, methods are presented rather than theory, although

a few sections on theory are included for the students who wish to know "why." These sections are marked and can be skipped without loss in continuity.

It would be impossible to give credit for the existence of this book to all those who deserve it. However, I am particularly indebted to R. F. White for his critical review of Chapter 8, to Timon A. Walther and Richard L. Beatty for their use of earlier drafts in the classroom, and to L. M. Giessinger for detailed checking of the manuscript. I am indebted to Professor Sir Ronald A. Fisher, Cambridge, to Dr. Frank Yates, Rothamsted, and to Messrs. Oliver & Boyd, Ltd., Edinburgh, for permission to reprint Table II from their book *Statistical Tables for Biological, Agricultural, and Medical Research.*

Edward C. Bryant

CONTENTS

Preface vii

1. Introduction 1
2. Probability 13
3. Statistical Inference I 29
4. Statistical Description 37
5. Statistical Inference II 67
6. Analysis of Frequencies 100
7. Simple Linear Regression and Correlation 113
8. Analysis of Variance and Industrial Experimentation 136
9. Analysis of Time Series 170
10. Multiple Regression 198
11. Sampling 225
12. Statistical Quality Control 244
13. Operating-level Problems 264

Appendix 285

Index 297

STATISTICAL ANALYSIS

1. INTRODUCTION

1-1. STATISTICS IN MANAGEMENT

The business manager, by the very nature of his job, must make decisions. The function of statistics is to help him make such decisions. It can be of assistance to him in two ways. First, statistics can present the manager with numerical facts (or approximations to facts, for seldom are quantities known exactly). Second, statistics in many cases can attach a probability to the risk of making a wrong decision. This can be an extremely important function—not because it eliminates risk, for it doesn't—but because it permits management to evaluate the magnitude of the risk in the light of possible gains to be achieved.

In the first place, it must be realized that statistics is not the only basis for managerial decision. When management is called upon for a decision, it relies on its knowledge of a great many factors, such as the international situation, Federal and local politics, weather, actual and anticipated style changes, personnel, and public relations. A manufacturer of automatic washing machines may be making and selling a clothes dryer at a loss. His statistician may predict accurately that this operation will continue to lose money during the next year. In spite of this prediction, the manufacturer may decide to continue production of the losing line because of increased sales of washers created by offering the public a "matched pair." Nevertheless, the statistical information furnished to management forms an important part of the basis for decision.

1-2. STATISTICS AS NUMERICAL FACTS

It was indicated above that one of the functions of statistics is to present management with quantitative facts (or approximations thereto) which may assist in making a managerial decision. This activity can be

subdivided further into (1) the recording and summarizing of numerical facts which are known or which can be determined and (2) the estimation of quantities from examination of samples.

The first activity includes recording, summarizing, and presenting in tables or charts historical data such as gross sales, number of employees, and units produced. It is primarily a clerical job and does not form a part of the subject matter to be discussed in this book.

Unfortunately for the statistical profession, this is the sort of activity which forms the popular concept of statistics. We hear of "statistics" of football games, and the like. In this book, the word "data" is used where such meanings are implied. The word "statistics" is reserved for the science and art which form the subject matter of this book and for the plural of "statistic," which is defined below. Similarly, the fellow who records yardage gained, penalties, and so forth at a football game is not referred to as a statistician. True enough, he is doing statistical work, but would you call a laboratory technician a physician?

We turn our attention then to the second activity listed above, that of estimating unknown quantities by examination of samples. Here it seems advisable to introduce a minimum of statistical terminology relating to sampling.

The whole collection of objects, individuals, or things under consideration is called a *universe*. For example, one might consider a universe of accounts receivable on a particular date or a universe consisting of 1 day's production of gismos. A universe may be finite (as in the above examples) or it may be infinite. For example, the universe composed of the tosses of a coin may be boundless, and therefore infinite. Sometimes the word *population* is used interchangeably with *universe*.

The elements of a universe may have one or more characteristics associated with them. These characteristics are called *variables*. For example, in a universe of accounts receivable some associated variables might be the amount of the account, purchases during the past 10 days, and the status of the account as current or delinquent. The first two variables are numerical, and the third is qualitative. The third can be made numerical, or quantitative, by assigning, say, the value 0 to "delinquent" and the value 1 to "current."

A *sample* is a segment of a universe, the segment ordinarily being selected in a particular way. For example, if a company makes 10,000 individual sales during a month, we may consider the amount of each sale as the variable and the total collection of 10,000 sales as the population. Suppose we wish to find out what percentage of these sales has been made to restaurants and hotels. We do not wish to take the time to examine 10,000 sales slips, so we examine a sample of 500. The method of drawing the sample is discussed in the next section. Suppose that we

find the following data upon examination of the sample:

	Restaurants and hotels	Others	Total
Number of sales	75	425	500
Amount of sales	$10,000	$80,000	$90,000

We can compute the following statistics (a *statistic* is a computation applied to sample data):

1. Estimated percentage of sales slips made out to restaurants and hotels = $\frac{75}{500} \times 100 = 15$ per cent.
2. Estimated percentage of total sales made to restaurants and hotels = 10,000/90,000 × 100 = 11.1 per cent.

Now ask yourself whether you really read the above figures and computations. If you can honestly answer yes, you are indeed an unusual person. Furthermore, you will have little difficulty with this book. If you skipped over the figures, now is the time to break a habit which is very detrimental to the learning of statistics. Let's go back and have another look:

How many sales slips are in the population?
How many are in the sample?
How many of the sample sales slips were made out to restaurants and hotels?
What is the average sale to restaurants and hotels?
What is the average sale to others?
Do you understand the computation of the percentages?

Now let us proceed. The boss is not particularly interested in what the percentages are for the 500; he wants to know the percentages for the 10,000. These unknown figures relating to the population are called *parameters*. Ordinarily, parameters are not known but can be estimated. So you estimate them for your boss. You say something like this: "I estimate that 1,500 sales, or 15 per cent, were made to restaurants and hotels and that these sales amounted to $200,000, or 11.1 per cent of the total." (See whether you can figure out where the $200,000 came from. Don't proceed until you know.)

You know that your figures are subject to error since you examined only one-twentieth of the population. Therefore, you might want to hedge a bit and say that between 12 and 18 per cent of the total number of sales were made to restaurants and hotels and that they accounted for

between 7 and 15 per cent of the total sales. You could still be wrong, but the chances are small that you would be. (If you know the origin of these figures, you shouldn't be reading this book.)

The statement of the results in terms of limits of error, or *confidence* limits, is one of the principal achievements of statistics. The techniques for establishing these limits are described in a later chapter. The term "confidence" refers to your own confidence that in the long run your statement will prove correct. The boss cannot expect you to be right every time, but he may be justified in expecting you to be right, say, 95 per cent of the time.

1-3. RANDOM SAMPLING

By *sample* we do not mean just any old chunk of a population. When the word "sample" is used in this book, it will mean "random sample," or perhaps a sample which is essentially random, that is, one which may be expected to have the same characteristics as a random sample.

One of the characteristics of a random sample is that all elements (units, individuals, or numbers) in the population have an equal opportunity of being included in the sample. If one wishes to sample customer opinion in a department store, he would not obtain a random sample of customers by talking only with those in the hardware department. Certain customers might never frequent the hardware department; hence they would have no opportunity of being included in the sample.

How does one sample at random? There are several alternatives. Nearly everyone is familiar with some of them. The boys who break a window with a baseball and who draw lots to see who will retrieve the ball are employing a random procedure. Drawing a number from a hat is one of the commonest randomization techniques.

To return to our hypothetical sample of 500 sales slips, one might shuffle the entire 10,000 (imagine the consternation of the accounting department at this suggestion!) and then select 500 of these. The disadvantages of this technique are obvious.

Since the sales tickets are numbered serially, it may be possible to draw numbers at random which identify them. This process is facilitated by use of a table of random numbers, such as Table 1-1. To visualize how such a table is constructed, assume that 10 cardboard disks are numbered 0, 1, . . . , 9 and placed in a bowl. They are stirred thoroughly, and one is selected. Suppose it is 6. It is then the first digit appearing in Table 1-1. It is replaced in the bowl and, after thorough mixing, another disk is drawn. It is number 4 and is the second digit appearing in Table 1-1, and so forth. There are more efficient ways of constructing the table, however. The above process is described only for the purpose of illus-

Table 1-1. Random Digits

6433	2582	0820	1460	6606	7143	9158	5114	9491	8063
3465	7348	5774	3821	6216	2148	1221	5895	7942	9971
9601	9189	0141	1377	3467	7971	0811	8309	0504	4606
2364	3260	1430	9505	3146	4815	9732	3447	7705	4532
7304	9292	4580	8160	7144	8073	8476	1896	6661	1285
3764	5460	6385	9045	7170	5831	4668	9386	3979	1116
0251	3139	4201	0578	2172	6876	4347	4288	1514	9985
2031	0919	7613	1535	1610	7491	3255	4014	3614	5599
6398	1374	1904	7490	3941	0284	5817	1630	4629	6773
0911	3930	0324	8151	3365	6685	0566	5047	8471	6166
5052	5023	3045	3433	6365	7310	5073	5416	2332	0922
9225	3984	4659	4642	7260	1383	7625	7512	8547	7343
3100	7916	9757	8869	5307	2691	0786	2701	0102	5745
4598	0065	4257	6557	4638	8418	7398	9790	5074	8018
5956	7285	0480	1411	7766	3377	5023	0227	8047	1887
9360	1041	2094	4212	2623	2384	6422	5374	0651	8673
8796	9974	1913	8309	4943	9423	9143	4683	4436	8413
7071	8254	6825	3020	9000	4673	6129	0176	3670	4836
7336	4451	5868	6559	5344	0714	1856	0451	7855	5998
1660	0222	2005	0215	2370	2687	3039	7953	1960	6579
7506	1020	8718	9665	1892	8245	7249	6023	4602	4227
5500	8237	6203	6829	5325	5784	8720	5053	6347	1112
4255	6894	8093	9191	5011	0452	6199	0009	8086	5170
5764	9837	6780	7490	5412	4869	6950	4183	8671	4008
3609	1368	9129	7113	3099	1887	0544	6415	9148	4381
7218	5939	4932	5465	6648	6365	4179	9266	9803	5572
6854	5911	1495	4940	4630	4514	0942	7218	7382	2145
4403	4263	4755	5451	8251	2652	6207	4841	3528	7665
2978	4381	2205	9638	6946	7126	9039	9194	6676	4396
1072	2292	4428	4934	8183	7385	3236	7748	4488	1351
6488	6568	9530	8316	7709	9022	8041	5564	6667	5329
9263	7756	6300	6793	7769	3099	3606	2468	2574	5230
0357	3493	0385	4451	4313	3024	8243	4920	3523	9644
5372	9351	8393	6023	2811	1744	2306	7083	4330	7278
6570	2866	7565	7871	9490	9050	4454	3475	8319	2972
8596	8251	0336	8119	1966	9115	4202	7785	5269	5941
4177	0092	4207	7386	9891	1149	3429	7062	4622	8415
8438	4892	2089	5509	2054	9024	1213	5791	2543	7863
5820	6287	7484	0339	8585	0968	3675	2440	4000	5148
7721	3804	9520	6184	9152	1853	8640	3601	5606	7218

trating the nature of the digits drawn. The grouping of digits in Table 1-1 is just a convenience for following rows and columns.

Large tables of random digits are available and may be found in nearly any statistical office.* They completely eliminate the necessity for the "numbers in a hat" technique.

To use the tables, select any page, close your eyes, and place a pencil point on the page. You may poke holes in the desk blotter and upset the inkwell in the process, but eventually you will hit the page. Suppose your pencil is pointing at the fourth row and the second column of four digits in Table 1-1. The starting number is then 3,260. Which of the 10,000 sales slips does this number identify? Assume that the first ticket is numbered 30,001 and the last, 40,000. It is clear that the four digits will identify the ticket uniquely. The first ticket included in the sample is number 33,260. (The first digit is ignored, since all numbers are in the 30,000 group.) Moving down the column, you find that the next number is 9,292, which identifies number 39,292. When you reach the bottom of the page, take the next column and continue until 500 numbers identifying 500 different sales tickets have been drawn. Ignore duplicated numbers, since the sample should not include any ticket twice.

Sometimes we cannot use all the numbers drawn. Suppose we wish to draw a random number between 0,001 and 4,260. If we draw a number greater than 4,260 (say, 7,240), we ignore the number and draw until a number in the range 0,001 to 4,260 is drawn.

If we are selecting three-digit numbers from the table, we use only the last three digits in the four-digit random numbers. Or, suppose we wish to select or not select individuals with equal probability. We may use only the last digit, letting an even digit represent "select" and an odd digit "don't select." If we wish to select with probability $\frac{1}{4}$, we may let the digits 01 through 25 represent "select" and 26 through 00 represent "don't select."

Again, suppose we wish to draw a number at random between 1 and 63,492, inclusive. Five columns of digits are required. We can consider two adjacent columns of four digits each as making up an eight-digit number. These eight-digit numbers are drawn at random and the first three digits ignored in identifying the random number to be used. For example, suppose that in Table 1-1 we combine columns 1 and 2, 3 and 4, 5 and 6, 7 and 8, 9 and 10. Suppose also that the random starting place we have chosen is row 6 of the combined columns 7 and 8. The eight-digit number is 46,689,386. Ignoring the first three digits we have 89,386, which lies outside the range of 1 to 63,492. We proceed down the double column and obtain 74,288, which also is too large. The next number,

* In particular, the Rand Corporation has constructed a table of 1 million such digits.

54,014, lies within the desired range and identifies our first random selection.

Although the table of random digits is an extremely valuable device for selecting many samples, the *systematic* sample is more efficient for some sampling problems. If, for example, the sample is to consist of one-twentieth of the population, every twentieth ticket may be selected. The starting point should be chosen at random from the first 20 tickets. The first ticket is 30,001 and the twentieth is 30,020, so a two-digit random number is drawn in the interval 01 to 20 to identify the starting point. Suppose it is 13. Then the sample tickets are numbered 30,013, 30,033, 30,053, and so forth. The systematic sample should have the characteristics of a random sample if there is no systematic pattern in the arrangement of the tickets (perhaps a reasonable assumption in this case).

It appears that a good bit of attention has been given to the topic of randomness, but it is a much-misunderstood concept, albeit a very important one.

1-4. RISKS IN DECISION

From the point of view of management, perhaps the most important idea in all of statistics is that each managerial decision involves two types of risk and that it is usually impossible to avoid both.

Consider a decision to accept or to reject a policy, idea, or procedure—in other words, a decision to "do something." If we do something when we should have done it, we have made a correct decision. Similarly, if we don't do it, and shouldn't have done it, we also have made a correct decision. But consider the other two cases: we may do something when we shouldn't or may not do something when we should. Both acts represent errors, and it is impossible (generally) to avoid the risk of one or the other. We can avoid the error of doing something when we shouldn't by never doing anything. But this doesn't avoid the error of not doing something when we should—in fact, it aggravates it.

To illustrate, let us consider the decision to hire a prospective employee. If he would make a good employee and we don't hire him, we err; if he would make a poor employee and we hire him, we also make an error. In such cases management often seeks a rule of procedure to govern its decision. It may, for example, give a standardized test and hire the prospective employee only if he scores 75 or above.

It must be clear that the efficiency of such a scheme in reducing errors of decision is closely related to the effectiveness of the test as an indication of ability and to the effectiveness of 75 as a criterion score. It is inconceivable that a test could ever be constructed so that everyone scoring 75 or over would be "successful" and everyone scoring below that point

would be "unsuccessful." In the first place, there are all shades of success and failure. Furthermore, a capable person may, for innumerable reasons, fail to obtain a high score on a particular test. The converse obviously is true also. In a long sequence of individual tests a certain portion of "unqualified" individuals will obtain scores of 75 or above. Similarly, a certain portion of "qualified" persons will obtain scores below 75.

The proportion of unqualified persons receiving a score of 75 or above represents the risk that an unsatisfactory employee will be hired. The proportion of qualified persons receiving a score below 75 represents the risk that a good prospective employee will be turned away. The relative seriousness of these errors in decision depends upon the need for filling the vacancy, the availability of prospective employees, and the cost of interviewing and testing.

Suppose that we raise the criterion from 75 to 80. Clearly we shall reduce the probability of hiring unsatisfactory employees, but we shall increase the probability of turning away good prospects. This might be a wise policy if there is a plentiful supply of workers, but it might be very unwise if there is a tight labor market.

The point of this discussion is that there are two classes of errors in decision and that it is almost always impossible to avoid both of them. One of the principal functions of statistics, as far as management is concerned, is to evaluate the risks involved in a particular decision rule. This is another way of saying that statistics is used in establishing decision rules.

1-5. AN EXAMPLE OF A DECISION RULE

Mr. Long lives five blocks from the subway station. In the morning he walks to the station and catches the train to his office. In the evening he rides the train to his station and walks home. He finds that on a number of occasions he must walk home in the rain. If he has his raincoat with him, he doesn't mind the rain, but it costs him a pressing bill if his clothes get wet.

Mr. Long has an orderly mind, but no imagination. (We'll see why later.) He searches through a mass of weather records and finds out that if the wind is from the east in the morning, the chances are 3 out of 4 that it will be raining in the afternoon. On the other hand, if the wind is in any other direction the chances are only 1 out of 10 that it will be raining in the afternoon. He decides, therefore, that if the wind is from the east he will take his raincoat. Otherwise he will not.

If the wind is from the east, then, he runs a risk of 1 out of 4 of carrying his raincoat unnecessarily. If the wind is not from the east he runs a risk of 1 out of 10 of getting wet. Note that he can avoid taking his raincoat

unnecessarily by never taking his raincoat, but he is sure to get wet every time it rains. Or, he can avoid getting wet by always carrying his raincoat, but he does not want to be bothered with a raincoat unnecessarily.

Now, suppose that it costs Mr. Long \$1 to have his suit pressed and also that it costs him 10 cents per day in depreciation to carry his raincoat. Suppose also that in a period of 24 days there are 4 days in which the wind is from the east. If Mr. Long follows the rule he has chosen, what is his "expected raincoat cost"? Clearly, he will carry his raincoat 4 days (the days in which the wind is from the east), so his depreciation cost will be 40 cents. On the remaining 20 days he can expect to be rained on twice ($\frac{1}{10}$ of the time), so his pressing bill will be \$2. His total expected cost is then \$2.40.

Now let us examine another decision rule. Suppose he always takes a raincoat. Then he has no pressing bill, but his depreciation bill is \$2.40. This rule is just as good. Or, suppose he decides never to take a raincoat. Then he may expect to get wet 2 days when the wind is not from the east and 3 days ($\frac{3}{4} \times 4$) when the wind is from the east. His expected pressing bill is then \$5, a greater expense than he would incur with either of the previous plans. For any particular situation the best rule will depend upon the cost of depreciation, the cost of pressing, and the number of days of east wind.

It is important to note here that when we speak of expected cost, we are basing our expectation on the expected number, or average number, of days of rain. In any particular 24-day period there might be more or less than 5 days of rain, but *on the average* we should expect 5. This, too, is typical of a management decision rule, or policy. It is designed to yield a satisfactory result on the average, but not necessarily each time that it is employed.

We said earlier that Mr. Long had no imagination. If he possessed imagination, we should expect him to buy two raincoats and to keep one at the office. Then, if it was raining when he left the office, he would take his raincoat home with him, returning it the next day. Since it is expected to rain only 5 days out of the 24, his expected cost is then 50 cents—a substantial saving. From the standpoint of illustration, however, it is better not to attribute imagination to Mr. Long.

EXERCISE 1-1. Consider Mr. Long's case again and suppose that the cost of carrying his raincoat (in depreciation) is only 5 cents. What is the expected cost of each of the three policies proposed?

EXERCISE 1-2. Assume the same depreciation cost as in Exercise 1-1 but suppose that the wind is from the east for 8 days out of 24. What is the expected cost of each of the three proposed policies? Note that there is nothing wrong with "pricing out" the fractional day, because in a long sequence of such 24-day periods it will rain, on the average, this fraction of the time.

1-6. MISUSES OF STATISTICS

Most of us have heard such comments as "Figures don't lie, but liars figure"; "There are liars, damn liars, and statisticians"; and "If all the statisticians in the world were laid end to end—it would be a good thing!" These derogatory expressions have arisen largely because of the misuse of statistics by untrained practitioners of the art of statistics rather than by statisticians themselves. These people stand in the same position with respect to the professional statistician as the witch doctor does to the trained medical doctor. Such misuse may reflect either ignorance or an intent to deceive. In any case, the public is so often exposed to such "statistics" that attention will be given to some obvious misuses here. Some of the principal classes, with illustrations, are given below.

Arithmetic ignorance. "Traffic fatalities increased from 15 to 45, an increase of 300 per cent." The *increase* is 200 per cent, not 300.

Spurious accuracy. "The average age at which girls in Hopeville begin to use lipstick is 11.897 years." This pins it down to about a third of a day! Even with a sample of 1,000, one could not come within several days of the average age. In any case, such information is probably of use to advertising departments of cosmetics companies as a guide to proper use of advertising media. For such purposes, accuracy to within half a year is certainly adequate.

Careless generalization. "A class in Psychology 301, after listening to a recording of Brahms' *Symphony No. 4 in E Minor* and Beethoven's *Symphony No. 3 in E Flat Major*, indicated the following preferences: Brahms, 32; Beethoven, 12. This shows clearly that people prefer Brahms to Beethoven." There are at least two major fallacies in the generalization: (1) the assumption that these two symphonies are representative of the works of the two composers; (2) the assumption that the psychology class is representative of "people." Might the order in which the symphonies were played be important? Might the recordings themselves have had an effect? How about the artistry of the orchestras?

The careless generalization is one of the greatest sources of error in the interpretation of statistics. Many cases are not so obvious as the one given here. Suppose one wishes to find out whether a new worktable design will increase production in the assembly of complex instruments. A simple approach might be to experiment and find out. But the novelty itself, rather than the design, may be the stimulus which motivates increased production. If so, a large investment in new equipment may result in money wasted.

Improper assignment of cause. One of the difficulties with social science research is inability to experiment, except within narrow limits. One generally must observe society as it functions under a myriad of

uncontrolled influences, and from these observations attempt to find some pattern of behavior. It is easy to be led astray, as the following example illustrates. Suppose a research worker wishes to find out whether radio advertising increases sales. He convinces management of the importance of the project, and the company puts on an intensive radio advertising program for a month. At the end of this period he observes that sales have increased over the level of the previous month. It would be easy to conclude that the increase was *due to* radio advertising. Not necessarily so! Sales in general may have been up. The seasonal pattern of sales may have been on the upswing. A major competitor may have been closed for alterations. Maybe there were more Saturdays in this month than last. Any number of things may have contributed to the rise in sales, and it is not appropriate to attribute the increase to radio advertising without careful investigation of other factors. Even then, some doubt will exist, for the business situation never remains stable, and the phrase "other things being equal" seldom is indicative of a true state of affairs.

The dishonest survey. The question of ethics cuts across all disciplines, but in statistics it appears to be a critical matter, because of the faith that many people have in figures. A published figure is not a fact per se, although many people apparently believe so. A survey showing that 70 per cent of the students on a campus were carrying brand X cigarettes on a given day may have no significance if on the previous day the brand X representative was busy handing out free samples!

The above are only a few of the improprieties which may crop up in statistical work. A complete cataloguing of them would be out of place here, but it is hoped that these few comments will stimulate the student to analyze statistical statements carefully.

EXERCISE 1-3. Criticize each of the following statements: (*a*) Women prefer Choko cigarettes. A survey in Denver revealed that 55 per cent of women prefer Chokos. (*b*) A new cure for arthritis has been found. Four patients treated with the new wonder drug are completely recovered. (*c*) Production of refrigerators decreased over 120 per cent last year.

EXERCISE 1-4. Read the definition of "random sample" again and describe how you would draw a random sample of students on your campus. Note that some definition of the population may be necessary.

EXERCISE 1-5. Describe how you would organize a study to determine whether barbecue relish should be shelved with the pickles or with the ketchup and mustard in a chain grocery store.

EXERCISE 1-6. A new keyboard arrangement has been developed for the typewriter. The inventor claims that it is superior because it enables a person to type faster. How would you organize an experiment to find out whether the new arrangement is superior? Pay particular attention to the persons chosen for the experiment. How about the factor of prior training?

EXERCISE 1-7. A foreman wishes to assign an operator to a machine at random. He has 4 operators and 4 machines. He decides (without looking at his watch) that if the second hand is between 0 and 15 he will choose A, if between 15 and 30 he will choose B, and so on. Is this a satisfactory random procedure?

EXERCISE 1-8. A one-to-one correspondence can be drawn between decimal digits and binary digits as follows:

Decimal digit:	0	1	2	3	4	5	6	7	8	9
Binary digit:	0	1	10	11	100	101	110	111	1,000	1,001

Explain how one could draw a number at random between 1 and 12 by tossing a coin, letting heads equal 1 and tails 0.

EXERCISE 1-9. Refer to the table of the Poisson distribution in the Appendix. Could one use the last digits of these numbers as random digits? Why or why not?

EXERCISE 1-10. Suppose you adopt this decision rule in taking a true-false examination: Toss a coin; heads equals true, tails equals false. (*a*) What is your expected score if you get one point for each correct answer? (*b*) What is your expected score if double the point value is subtracted for each wrong answer? (*c*) Suppose you decide to mark the longest statements "true" and the shortest statements "false." Will the same probability distribution hold? What is the difference, if any, in the two decision rules?

2. PROBABILITY

2-1. INTRODUCTION

All of us are familiar in general with probability concepts. We hear that it probably will rain tomorrow or that the odds are 2 to 1 that Swayback will win the fourth race. Since a major portion of statistics deals with probability concepts, it is important for us to refine our ideas about probability so that we can use them for problem solving. This is not to say that a philosophical discussion of probability is in order. In fact, we shall avoid such discussions; we shall simply admit the existence of probability and develop certain rules of manipulation.

2-2. DEFINITIONS

We need some preliminary definitions before beginning our discussion on the rules of manipulation. First, we must distinguish between *a priori* and *empirical* probability. An a priori probability, as the name indicates, is one which can be determined prior to any experimentation or trial. For example, we say that the probability of obtaining heads in tossing a coin is $\frac{1}{2}$. We do not toss the coin to find this out. We simply observe that there are two faces to the coin, only one of which is heads, and that the probability must therefore be $\frac{1}{2}$. To formalize these ideas, let S be the set of possible occurrences and assume that S can be divided into two classes, the elements of one class having a property A and the elements of the other class not having the property A. Then the probability of A, written $P(A)$, is

$$P(A) = \frac{N(A)}{N(S)} \tag{2-1}$$

where $N(A)$ is the number of elements in A and $N(S)$ is the number of elements in S. In our illustration above, S consists of the two faces,

heads and tails; hence $N(S) = 2$. A is the property "heads," so that $N(A) = 1$. Therefore

$$P(\text{heads}) = \tfrac{1}{2}$$

Note that we ignore the event that the coin may come to rest on an edge. Note also that we do not argue whether the probability is in fact $\frac{1}{2}$. We simply assert that by definition it has this value, and to justify our position we may assert that we are dealing with a "fair coin," that is, one which is carefully balanced so that the probability is in fact $\frac{1}{2}$. But here we are involved in circular reasoning. The truth of the matter is that we cannot prove that the probability is $\frac{1}{2}$ by experimentation. Neither can we disprove it, in the strict sense.

The handling of this a priori probability may seem pretty arbitrary, but this is a characteristic of such probabilities. Actually, we all have had experience with this sort of arbitrariness in our high school geometry courses. We deal with points, straight lines, angles, and so forth. But one cannot construct a point or a straight line. He can only construct rough representations of these concepts. Similarly, with probability it is easy to conceive of a probability of, say, $\frac{2}{3}$, but it is impossible to construct a physical representation of this exact probability. Perhaps it is more precise to say that, if one constructs a representation of a given probability, it is impossible to prove that it is exact.

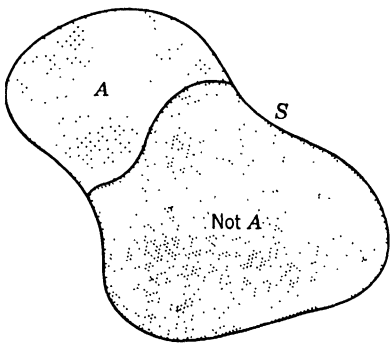

FIG. 2-1. Example of area as the measure of a set

We seem to have belabored the point, but an understanding of a priori probability is important in many statistical methods.

EXERCISE 2-1. Suppose we are examining a lot of 1,000 diodes, of which 20 are defective. Let A be the property that the diode is defective. (*a*) What is S? (*b*) What is $N(S)$? (*c*) What is $N(A)$? (*d*) What is $P(A)$?

Examination of the above illustrations, as well as Eq. (2-1), shows that we are limited seriously in using this as our only definition of probability. The limitation arises because we must be able to count the elements with the property A and the elements in S. Consider the configuration shown in Fig. 2-1. Clearly there are infinitely many points in A and infinitely many points in S, so that $N(A)/N(S)$ is undefined. We need some measure other than number of elements to apply to A and to S. Assuming that we can find such a measure, we can redefine the probability of A

to be

$$P(A) = \frac{M(A)}{M(S)} \tag{2-2}$$

where $M(A)$ is the measure of the set A and $M(S)$ the measure of the set S. If we let our measure be the area in this case, we have the probability of A equal to the area of A divided by the area of S. It should be obvious that in some cases length would be a suitable measure and in other cases the number of elements might serve as a measure. That is, referring again to Eq. (2-1), if we let the number of elements serve as our measure, we can write $N(A)$ as $M(A)$ and $N(S)$ an $M(S)$, so that Eq. (2-2) can serve as a perfectly general definition of probability.

EXERCISE 2-2. A machine operator is busy 80 per cent of the time. What is the probability that at a randomly selected instant he will be idle?

So far we have considered only a priori probabilities, that is, probabilities that are determined by means of logic and that need not be verified by experimentation. An *empirical* probability is one which is estimated on the basis of experimentation or observation. These probabilities are quite important in the application of statistics to management, for in many cases a probability cannot be determined exactly but can be estimated. Empirical probabilities are discussed in a later chapter.

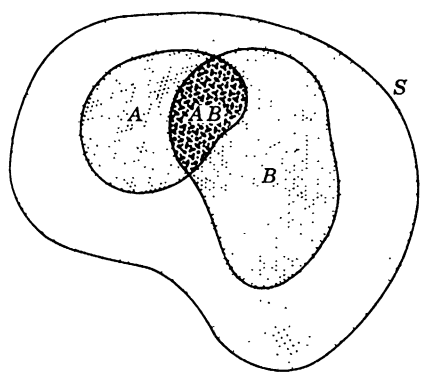

FIG. 2-2. Overlapping subsets of the set S

2-3. RULES OF MANIPULATION

We shall be concerned first with a rule for the addition of probabilities. Suppose that the elements of a set S can be classified as having the characteristic A, having the characteristic B, having both A and B, or having neither A nor B. We can picture this situation by the schema of Fig. 2-2. The probability of A or B may be expressed as follows:

$$P(A \text{ or } B) = \frac{M(A) + M(B) - M(AB)}{M(S)} = P(A) + P(B) - P(AB) \tag{2-3}$$

This expression follows from the definition that $P(A \text{ or } B)$ is the measure of the elements which are either A or B divided by the measure of the whole set S. It is to be noted that the set of elements having the property

A includes the elements which are both A and B, so that $M(A) + M(B)$ includes this overlapped set twice. This explains the necessity for subtracting $M(AB)$. Equation (2-3) can be extended. With three classes, A, B, and C,

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) \quad (2\text{-}4)$$

EXERCISE 2-3. Why?

A special case arises when the classes do not overlap. In this case

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) \quad (2\text{-}5)$$

and we say that A, B, and C are mutually exclusive categories. In general, we can say that, if $P_1, P_2, P_3, \ldots, P_r$ are the separate probabilities of r *mutually exclusive* events, the probability that one of these events will happen in a single trial is the sum of the separate probabilities.

EXERCISE 2-4. What is the probability of obtaining an odd number in casting a single six-sided die? Which of the above formulas did you use? Note that you can obtain the answer by using the definitional formula (2-2) as well as Eq. (2-5).

EXERCISE 2-5. In a lot of 2,000 manufactured items, 100 have defective finishes and 200 have defective mechanisms. (*a*) What else do you need to know to answer the question, how many are defective? (*b*) If 80 items have only defective finishes (their mechanisms are all right), then what is the probability that an item drawn at random from the 2,000 will be defective (either defect)?

Next, we need to know something about *conditional probability*, by which we mean the probability that an event will happen once another has occurred. It will be helpful to refer again to Fig. 2-2. We may use the symbol $P(B|A)$ to denote the probability that an element has the characteristic B if it has the characteristic A. By reference to Fig. 2-2, we see that

$$P(B|A) = \frac{M(AB)}{M(A)} = \frac{M(AB)/M(S)}{M(A)/M(S)} = \frac{P(AB)}{P(A)} \quad (2\text{-}6)$$

In a lot of 10 items, 4 are defective. Suppose that 1 item is drawn at random and found to be defective. What is the probability that the next item drawn at random will be defective? Using the notation of Eq. (2-6), let A represent the property that the first item drawn is defective and let B be the property that the second item drawn is defective. Then $P(AB)$ is the probability that the first 2 drawn are defective. Solving Eq. (2-6) for $P(AB)$, we have

$$P(AB) = P(A)P(B|A) = \tfrac{4}{10} \times \tfrac{3}{9} = \tfrac{12}{90} = \tfrac{2}{15} \quad (2\text{-}7)$$

Let C be the property that the third item drawn is defective. Then

$$P(ABC) = P(A)P(B|A)P(C|AB) = \tfrac{4}{10} \times \tfrac{3}{9} \times \tfrac{2}{8} = \tfrac{1}{30}$$

This relationship, which establishes a rule for the multiplication of probabilities, can be extended indefinitely. A change in notation is helpful for the generalization. Let P_1 be the probability of event 1, $P_{2,1}$ the probability of event 2 once 1 has occurred, $P_{3,12}$ the probability of event 3 once 1 and 2 have occurred, and so forth. Then the probability that r events will occur in the sequence specified is

$$P(1, 2, 3, \ldots, r) = P_1 \times P_{2,1} \times P_{3,12} \times \cdots \times P_{r,12\ldots(r-1)} \quad (2\text{-}8)$$

EXERCISE 2-6. Refer to the numerical illustration above. (*a*) What is the probability that the 4 defective items will be drawn on the first 4 trials? (*b*) What is the probability that 4 good items will be drawn in 4 trials? (*c*) What is the probability that the first 2 will be good and the second 2 bad? (*d*) What is the probability that the first 2 will be bad and the second 2 good?

We say that two events, A and B, are *independent* if $P(B|A) = P(B)$ or if $P(A|B) = P(A)$. (The statements are equivalent.) In the case of independent events, the probability that r events will occur in a specified sequence is simply the product of the r probabilities:

$$P(1, 2, 3, \ldots, r) = P_1 \times P_2 \times P_3 \times \cdots \times P_r \quad (2\text{-}9)$$

In the case of the lot of 10 items of which 4 are defective, suppose that one replaces the first item drawn before drawing the second. It is then clear that $P(B) = P(B|A)$, and we say that the drawings are independent.

EXERCISE 2-7. Answer the questions of Exercise 2-6 assuming that the items are replaced in the lot of 10 before each succeeding draw.

2-4. PERMUTATIONS AND COMBINATIONS

The actual counting of the ways in which an event may occur often becomes tedious without the aid of some mathematical tools. It is with this in mind that we discuss permutations and combinations.

A *permutation* is an arrangement of things. When the order of arrangement is changed, a new permutation results. For example, consider the 3 letters A, B, C. They may be arranged in the following ways:

ABC	BAC	CAB
ACB	BCA	CBA

We conclude, then, that there are 6 permutations of 3 things when they are all considered together. Writing out all the possible permutations would become impossible if many items were involved; hence it is neces-

sary to generalize the procedure into a formula. We shall consider the problem of determining the number of permutations of n things when all n are included in each arrangement. There are n ways that the first position in the permutation can be filled. After selecting one of these n ways, there are $n - 1$ ways that the second position can be filled. For each of the n ways that the first position can be filled, there are $n - 1$ ways that the second position can be filled; hence there are $n(n - 1)$ ways that the two positions can be filled, and so on. We see, then, that there are $n!$ (n factorial) ways of arranging n things. We may express this symbolically as follows:

$$ {}_nP_n = n(n - 1)\ (n - 2) \cdot \cdot \cdot [n - (n - 1)] = n! \qquad (2\text{-}10) $$

or, in words, the number of permutations of n things taken n at a time equals n factorial. For example, in the above illustration we have 3 letters A, B, and C, so $n = 3$. Then the total number of ways we can arrange 3 letters when all 3 are taken at one time is $3! = 3 \cdot 2 \cdot 1 = 6$, which checks with the listing above.

A word about factorials is in order. Since we shall be working with factorials of whole numbers only, we see that

$$ n! = n(n - 1)(n - 2) \cdot \cdot \cdot (1) \qquad (2\text{-}11) $$

For example,

$$ 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120 $$

In the division of factorial numbers we can sometimes cancel out portions of the factorial numbers in order to reduce work. For example,

$$ \frac{8!}{5!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{5!} = 8 \cdot 7 \cdot 6 = 336 $$

Now let us return to the consideration of the number of arrangements above. Suppose we do not wish to consider all n things together. We wish to consider the permutations of n things considered r at a time, where r is equal to or less than n. We have then

$$ {}_nP_r = n(n - 1)(n - 2) \cdot \cdot \cdot (n - r + 1) \qquad (2\text{-}12) $$

For sake of simplicity in computation, we generally write this as

$$ {}_nP_r = \frac{n!}{(n - r)!} \qquad (2\text{-}13) $$

By the cancellation rule demonstrated above, we see that Eq. (2-13) is the same expression as Eq. (2-12). To illustrate, suppose we wish to compute the number of permutations of 8 things taken 3 at a time. There are 8 ways to fill the first space, 7 ways to fill the second space, and 6 ways

to fill the third space. Altogether there are $8 \cdot 7 \cdot 6 = 336$ ways to fill the 3 spaces. Let us apply formula (2-13):

$$ {}_nP_r = \frac{n!}{(n-r)!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 336 $$

EXERCISE 2-8. It takes 3 men doing jobs 1, 2, and 3 to assemble a mechanism, and there are 9 men working on the assembly floor. We want to pick the best assembly crew of 3 men out of the 9. How many such possible assembly crews are there?

Sometimes we have duplicated items to consider. We may eliminate duplicated permutations by division. For example, suppose we consider the letters which make up the word "STATISTICS." We note that S appears 3 times, T appears 3 times, and I appears 2 times. If we compute the total permutations of these letters, without duplications, we must take out the duplications due to S's, T's, and I's, as follows:

$$ \frac{{}_{10}P_{10}}{{}_3P_3 \, {}_3P_3 \, {}_2P_2} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 50{,}400 $$

This figure represents the total number of different 10-letter "words," whether or not they make sense, that can be constructed from the 10 letters making up the word "STATISTICS."

We have defined a permutation as an arrangement of things. A *combination* is a grouping of things, the order of which is not considered. There are 6 permutations of 3 things taken all at a time, but only 1 combination. Suppose, though, that we wish to compute the combinations of 3 things taken 2 at a time. Considering the 3 letters A, B, and C, we could have

AB AC or BC

Hence there are 3 such combinations. To reduce this procedure to a formula we note that, for every combination of r things, there are $r!$ permutations. That is,

$$ {}_nP_r = r! \, {}_nC_r $$

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!} \qquad (2\text{-}14) $$

If we wish to solve for the combinations of 10 things taken 4 at a time, we write

$$ {}_{10}C_4 = \frac{10!}{4!6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6!}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 6!} = 210 $$

We note that the number of combinations of n things taken r at a time is exactly equal to the number of combinations of n things taken $n - r$ at

a time, because

$$ {}_nC_{n-r} = \frac{n!}{(n-r)![n-(n-r)]!} = \frac{n!}{(n-r)!r!} $$

which is equal to formula (2-14) above.

EXERCISE 2-9. Refer to Exercise 2-8. If the jobs are all alike, how many possible different assembly crews could one try out?

EXERCISE 2-10. You are a member of a committee of 5 from which a subcommittee of 3 will be chosen at random to present a complaint to the boss. What is the probability that you will be chosen?

Hint: (*a*) Find the total number of different committees of 3 that can be chosen. (*b*) Find the number that will include you. (Your inclusion is definite, so you need to find the ways that 2 others can be added.) Divide (*b*) by (*a*). *Ans.* $\frac{6}{10}$.

2-5. PROBABILITY DISTRIBUTIONS

Sometimes we are interested not only in the probability that a particular event will occur but in the distribution of probabilities over the whole range of possible events. Consider, for example, a perfectly balanced six-sided die with faces numbered 1, 2, 3, 4, 5, and 6. If we let y represent the variable number on the face of the die, we can write

$$ \begin{aligned} P(y = 1) &= \tfrac{1}{6} \\ P(y = 2) &= \tfrac{1}{6} \\ &\cdots\cdots\cdots \\ P(y = 6) &= \tfrac{1}{6} \end{aligned} $$

All the probabilities are the same, because of our assumptions about the perfection of the die. The listing of the possible values for y and their associated probabilities is called a *probability distribution*. We can avoid writing all the separate probabilities by using the formula

$$ P(y) = \tfrac{1}{6} \qquad (y = 1, 2, \ldots, 6) \tag{2-15} $$

This says the same thing as the complete listing.

It will be noted that the probabilities, when summed over all the possible values for y, add to 1 (unity). This is a characteristic of all probability distributions. It simply means that all possible eventualities are accounted for. All probabilities must be nonnegative.

EXERCISE 2-11. Assume that the probability distribution is $P(y) = y/15$ ($y = 1, 2, \ldots, 5$). (*a*) What is the probability of obtaining a 2? (*b*) a 3? (*c*) a 5? (*d*) What is the probability of obtaining a value less than 4? (*e*) greater than 2? (*f*) Do the probabilities add to 1?

There are certain probability distributions which are used extensively by statisticians. Some of the more important ones from the standpoint of management decisions are presented in the following sections.

2-6. BINOMIAL PROBABILITY

Let us consider a variable x such that $x = 1$ with probability p and $x = 0$ with probability $1 - p$. It is clear then that x must equal either 0 or 1, and we can express the probability of x as

$$P(x) = p^x(1 - p)^{1-x} \tag{2-16}$$

If we let $x = 0$, we have

$$P(0) = p^0(1 - p)^{1-0} = 1 - p$$

remembering that any number to the zero power is equal to 1. This checks with our definition above. If we let $x = 1$, we have

$$P(1) = p^1(1 - p)^{1-1} = p$$

which also checks with the definition of x above. We call (2-16) the *binomial distribution for a single trial.*

For example, suppose $x = 1$ if we obtain heads in tossing a coin and $x = 0$ if we obtain tails. Then the probability that $x = 1$ is

$$P(1) = p^1(1 - p)^{1-1} = p \qquad (= \tfrac{1}{2} \text{ by assumption})$$

Now suppose we conduct n trials with the same binomial variable. We shall let y equal the number of "successes" in the n trials. It is clear that y may take on any value from 0 to n, inclusive. That is, it is possible that we shall not obtain any successes in n trials, and it is possible also that every trial will result in a success, in which case $y = n$. It is also clear that our new variable y is simply the sum of the x variable over the n trials. Now, let us proceed to find the probability that y equals any value from 0 to n, inclusive.

From what has been discussed earlier we can see that if $p_1, p_2, \ldots, p_n$ are the separate probabilities of success of n independent events, the probability that they will all fail on a given occasion is

$$(1 - p_1)(1 - p_2)(1 - p_3) \cdots (1 - p_n) \tag{2-17}$$

This follows directly from the multiplication rule. The probability that the first y will succeed and the remaining $n - y$ will fail is

$$p_1 \times p_2 \times \cdots \times p_y \times (1 - p_{y+1}) \times \cdots \times (1 - p_n) \tag{2-18}$$

For example, if we wish to compute the probability that in rolling a six-sided die 4 times the first 2 rolls will yield aces and the second 2 will fail

to yield aces, we compute, according to formula (2-18),

$$\frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} = \frac{25}{1{,}296}$$

If we let $1 - p = q$ and note that in this illustration all the probabilities of success are identical and also that the probabilities of failure, q, are alike, we can write a formula for this procedure as follows:

$$p^y q^{n-y}$$

where y is the number of successes and $n - y$ is the number of failures. Note that the order in which the successes and failures occur is not important; it is essential only that there be y successes and $n - y$ failures. Now there are ${}_nC_y$ ways that one may have y successes out of n trials; hence the probability of having *exactly* y successes in n trials may be written as

$$P(y) = {}_nC_y p^y q^{n-y} \tag{2-19}$$

This is the probability of y in which we are interested. The probability, then, of obtaining exactly 2 aces in 4 throws of a die is

$${}_4C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2 = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} \times \frac{25}{1{,}296} = \frac{150}{1{,}296}$$

The probability of obtaining at least y successes in n trials is the summation of the probability of obtaining exactly y successes plus exactly $y + 1$ successes, and so forth, up to n successes. This follows from the addition rule. We may write this probability as

$$P = {}_nC_y p^y q^{n-y} + {}_nC_{y+1} p^{y+1} q^{n-y-1} + \cdots + {}_nC_n p^n q^0$$

That is, the probability is just a sum of terms, each of which is computed by formula (2-19).

Now, if we add together all the probabilities (i.e., of 0 successes, 1 success, 2 successes, and so on, up to n successes), we obtain a formula for the binomial expansion* as follows:

$$(q + p)^n = {}_nC_0 p^0 q^n + {}_nC_1 p^1 q^{n-1} + {}_nC_2 p^2 q^{n-2} + \cdots + {}_nC_r p^r q^{n-r} + \cdots + {}_nC_n p^n q^0 \tag{2-20}$$

* The formula for expanding the binomial is usually written

$$(a + b)^n = a^n + na^{n-1}b + \frac{n(n-1)a^{n-2}b^2}{2!} + \cdots + \frac{n(n-1)\cdots(n-r+1)}{r!} a^{n-r}b^r + \cdots + b^n$$

It is immediately apparent that this expression is identical to formula (2-19) when n is a positive integer.

This summation is obviously equal to 1 since $q + p = 1$. Note that in ${}_nC_0$ and ${}_nC_n$ we have 0! in the denominator. The quantity 0! is equal to 1.

Suppose we wish to compute the probabilities of obtaining 0 aces, 1 ace, 2 aces, 3 aces, 4 aces, and 5 aces in the simultaneous cast of 5 dice. We compute

$$\left(\tfrac{5}{6} + \tfrac{1}{6}\right)^5 = {}_5C_0\left(\tfrac{1}{6}\right)^0\left(\tfrac{5}{6}\right)^5 + {}_5C_1\left(\tfrac{1}{6}\right)^1\left(\tfrac{5}{6}\right)^4 + {}_5C_2\left(\tfrac{1}{6}\right)^2\left(\tfrac{5}{6}\right)^3 + {}_5C_3\left(\tfrac{1}{6}\right)^3\left(\tfrac{5}{6}\right)^2$$
$$+ {}_5C_4\left(\tfrac{1}{6}\right)^4\left(\tfrac{5}{6}\right)^1 + {}_5C_5\left(\tfrac{1}{6}\right)^5\left(\tfrac{5}{6}\right)^0 = \frac{3{,}125}{7{,}776} + \frac{3{,}125}{7{,}776} + \frac{1{,}250}{7{,}776}$$
$$+ \frac{250}{7{,}776} + \frac{25}{7{,}776} + \frac{1}{7{,}776} = \frac{7{,}776}{7{,}776} = 1$$

The probability of obtaining 0 aces is then 3,125/7,776; the probability of obtaining 1 ace is 3,125/7,776; and the probability of obtaining all 5 aces is 1/7,776. Note that the number of successes (aces) is specified by the exponent on p in each term. The plotting of the distribution in Fig. 2-3 shows its asymmetrical (skewed) nature. If we had computed the probabilities of obtaining various numbers of heads in tossing 10 coins, we should observe that the distribution is symmetrical, since $p = q = \frac{1}{2}$. An arbitrary degree of skewness can be imparted to a distribution by changing the values of p and q, so long as $p + q = 1$.

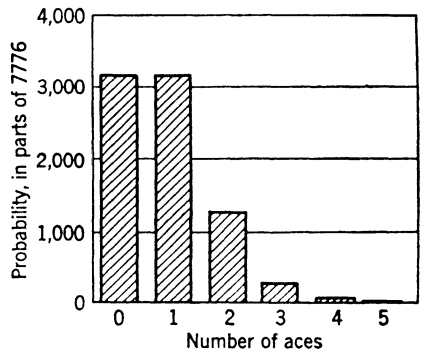

FIG. 2-3. Probability distribution for number of aces in casting 5 dice

In summary, we can recognize a binomial distribution by the following facts:

1. The events can be classified into only two categories (success, failure; yes, no; male, female; etc.).
2. The probability of a success remains constant from one trial to another. That is, the events are independent.

If we have a single trial, the probability of x may be expressed by

$$P(x) = p^x q^{1-x}$$

where x can equal 0 or 1, depending upon whether we fail or succeed. If we have a sum of binomial trials, the probability of exactly y successes in n trials is

$$P(y) = {}_nC_y p^y q^{n-y}$$

EXERCISE 2-12. If, on the average, 10 per cent of the items produced are

defective, what is the probability that in a sample of 5 items there will be (*a*) 0 defectives? (*b*) 5 defectives? (*c*) more than 3 defectives?

2-7. THE HYPERGEOMETRIC DISTRIBUTION

In the preceding section it was assumed that the trials were independent. That is, the probability remained constant from one binomial trial to the next. Sometimes this is not the case. For example, consider an urn containing 5 white marbles and 10 black marbles. What is the probability of obtaining exactly 2 white and 3 black marbles in drawing 5 marbles *without replacement* from the urn? It is clear that, if we draw *with replacement*, the trials are independent and we can use the distribution of the binomial sum y to find the answer. It is, in fact,

$$ {}_5C_2(\tfrac{1}{3})^2(\tfrac{2}{3})^3 $$

since the probability of obtaining a white marble remains constant at $\frac{1}{3}$ from one draw to the next.

However, when we draw *without replacement*, the probability changes after each draw. We obtain our probability as follows:

1. Find the total number of ways we can draw 5 marbles out of 15 $({}_{15}C_5)$.
2. Find the total ways we can draw 2 white marbles out of 5 $({}_5C_2)$.
3. Find the total ways we can draw 3 black marbles out of 10 $({}_{10}C_3)$.
4. Multiply (2) by (3) to find the total ways in which we can be successful $({}_5C_2\ {}_{10}C_3)$.
5. Divide the total ways of being successful by the total ways of drawing 5 marbles [(4) divided by (1)] to find the probability.

That is,

$$ P(2 \text{ white, } 3 \text{ black}) = \frac{{}_5C_2\ {}_{10}C_3}{{}_{15}C_5} = 0.40 \qquad \text{(approximately)} $$

In general, if we let

A = number in the population having a given characteristic
B = number in the population having another characteristic
a = number of A found in n trials
b = number of B found in n trials

then

$$ P(a \text{ of } A,\ b \text{ of } B) = \frac{{}_AC_a\ {}_BC_b}{{}_{A+B}C_{a+b}} \tag{2-21} $$

This is called the *hypergeometric distribution.*

EXERCISE 2-13. On a board of directors composed of 10 members, 6 favor a policy change and 4 oppose it. (*a*) What is the probability that a randomly chosen subcommittee of 5 will contain exactly 3 who favor the change? (*b*)

exactly 4? (*c*) exactly 5? (*d*) What then is the probability that a subcommittee of 5 will vote in favor of the change?

2-8. THE POISSON DISTRIBUTION

If n is very large and p is very small, so that np is a small positive number, then

$$P(y) = \frac{e^{-np}(np)^y}{y!} \tag{2-22}$$

gives a close approximation to binomial probabilities. The distribution, called the *Poisson distribution*, is most useful, however, when neither n nor p is known but their product np is known or can be approximated. Let us see how this situation might arise. Suppose the paint shop in a factory is concerned with the quality of its work. Quality is measured by the number of defects in an enameled surface. There may be no defects in a surface or there may be 1 defect, 2 defects, or more. On the average, there are m defects. This average number of defects corresponds to the product np above. Note that it would be impossible to use the product np, since n would represent the total number of opportunities for defects—something which is not defined in this case. Since we have replaced np by the single parameter m, we may write the formula for the Poisson distribution as

$$P(y) = \frac{e^{-m}m^y}{y!} \tag{2-23}$$

Let us consider a specific example. Suppose in our paint shop we find on the average that there are 2 defects per surface inspected. Then the probability of finding a surface without any defects is

$$P(0) = \frac{(2.718)^{-2}\,2^0}{0!} = 0.135$$

That is, about 13.5 per cent of all surfaces examined will have no defects The probability that there will be 1 defect is

$$P(1) = \frac{(2.718)^{-2}\,2^1}{1!} = 0.271$$

Probabilities of other numbers of defects can be found in similar fashion. A table of the Poisson distribution for selected values of m is given in Appendix Table V. The table is designed to illustrate the use of such tables rather than to cover all the values of m in which one might be interested. Extensive tables of the Poisson distribution have been computed, however.*

* For example, E. C. Molina, *Poisson's Exponential Binomial Limit*, D. Van Nostrand Company, Inc., Princeton, N.J., 1949.

2-9. CONTINUOUS PROBABILITY DISTRIBUTIONS

In the previous sections we have discussed discrete probability distributions, that is, distributions in which only isolated values of the variable can occur.* For example, in tossing a coin, only two possible values can occur. In casting a die only six values are possible. Many of the variables with which the statistician works are not discrete in character. Take, for example, the length of time that it takes to correct a stoppage in an automatic machine. Here, time is the variable and is continuous. That is, given any two times, say, now and $\frac{1}{2}$ sec from now, there are infinitely many different time values between these two.

Suppose we let Y be the length of time it takes to reduce a stoppage. How do we answer a question such as, what is the probability that $Y = 10$ sec? Before we give a quick answer to this question, suppose we consider the question, what is the probability that $Y = 10.57$ sec? or 9.99998? or 10.000001? It becomes clear that 10 is an exact length of time. So are the other numbers listed above. We begin to wonder whether there is any answer to our first question.

We shall recast our question as follows: What is the probability that Y is equal to or greater than 9.5 sec and less than 10.5 sec? Suppose we assume some value for this probability, say, p_1, and write†

$$P(9.5 \leq Y < 10.5) = p_1$$

Now it may be reasonable to suppose that, if a short time interval is cut in half, the probability will be cut approximately in half. At least we shall make this assumption and shall write

$$P(9.75 \leq Y < 10.25) = \frac{p_1}{2}$$

We cut the time interval in half again and write

$$P(9.875 \leq Y < 10.125) = \frac{p_1}{4}$$

and so forth. Each time we cut the interval in half we divide the probability by 2. It is clear that, as the interval converges on the figure 10, the probability approaches 0. We must conclude then that the question, what is the probability that $Y = 10$ sec? can only be answered by zero.

This is distressing to some people, but should not be. All it means is that, when we work with a continuous variable, we must apply our prob-

* A set of isolated values is one such that, given any two values of the set, there exists an interval between them which contains no other value of the set.

† The symbols $a < b$ are read "a is less than b," and $a \leq b$ is read "a is less than or equal to b." Similarly, $a > b$ is read "a is greater than b."

ability statements to intervals of that variable rather than to specific values. This means that our probability distributions must be stated differently also. We have seen that it is impossible to list all the possible values for the variable along with their associated probabilities, as we did for discrete variables at the beginning of the previous section. Instead, we shall express continuous probabilities in functional form. For example, we might write

$$f(y) = 0.05e^{-0.05y} \qquad (y \geq 0) \tag{2-24}$$

where e is the mathematical constant equal to approximately 2.718 and y is the continuous variable. The variable y might, for instance, represent

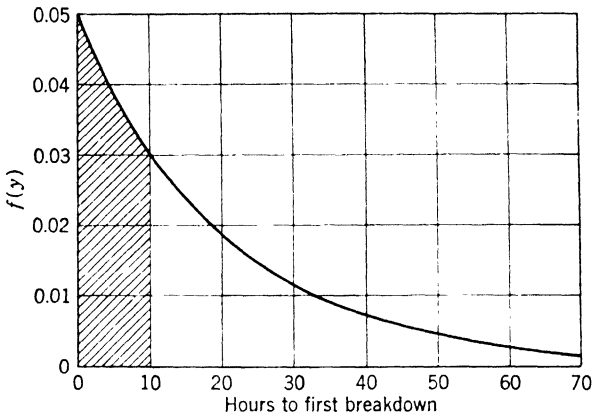

FIG. 2-4 Probability distribution for time to first breakdown of a complex mechanism

the time to the first breakdown of a complex mechanism. A graph of the distribution is shown in Fig. 2-4.

It should be noted that Eq. (2-24) is a special case of a general equation

$$f(y) = \alpha e^{-\alpha y} \tag{2-25}$$

where α is a parameter of the probability distribution. The value assigned to the parameter in the previous illustration is 0.05. In general, $1/\alpha$ is the average value of the y variable. Thus, in the numerical illustration above, the average time to first breakdown is $1/0.05$, or 20 hr.

Suppose we wish to know the probability that a mechanism will last less than 10 hr before the first breakdown. Using the area under the curve as our measure, the required probability is the area to the left of the point 10 hr (since the area under the whole curve is equal to 1). This area is shaded in the diagram. It can be found by direct computation using the integral calculus or, what is more practical for management students, can be read from a table of probabilities. More will be said about this later. In our particular illustration the required probability

is 0.39. In similar fashion the probability that y will lie in any other interval can be found.

EXERCISE 2-14. If an employee loafs 30 per cent of the time, what is the probability that he will be caught at it if the boss checks 4 times? It is easier to compute 1 minus the probability that he won't be caught.

EXERCISE 2-15. What is the probability of drawing 2 hearts and 3 spades from a deck of 52 cards (*a*) with replacement? (*b*) without replacement?

EXERCISE 2-16. What is the probability of drawing 2 red and 3 black cards from a deck of 52 cards (*a*) with replacement? (*b*) without replacement?

EXERCISE 2-17. Assume that heads = 1 and tails = 0 on a fair coin. Assume that the coin is tossed and that a six-sided die is cast simultaneously. What is the probability that (*a*) a sum of 1 is obtained? (*b*) a sum of 2? (*c*) a sum of 7?

EXERCISE 2-18. In the above illustration, assume that the die is cast only if heads appears on the coin. What is the probability of obtaining (*a*) a sum of 0? (*b*) a sum of 1? (*c*) a sum of 2?

EXERCISE 2-19. A committee is composed of 4 men and 6 women. A subcommittee of 5 is chosen at random. What is the probability that it is composed (*a*) of 2 men and 3 women? (*b*) of 4 men and 1 woman?

EXERCISE 2-20. What is the probability of obtaining each face in casting a die 6 times?

EXERCISE 2-21. In manufacturing bottles it is found that 5 per cent are defective. What is the probability that in a sample of 10 bottles there are (*a*) 0 defectives? (*b*) 3 defectives? (*c*) more than 4 defectives?

EXERCISE 2-22. (*a*) In how many ways can 4 people be seated at a bridge table? (*b*) If it is specified that A and B are to be partners, in how many ways can A, B, C, and D be seated?

EXERCISE 2-23. (*a*) If the probability is $\frac{1}{2}$ that A will pass a course, $\frac{1}{3}$ that B will pass, and $\frac{3}{4}$ that C will pass, what is the probability that all will pass? (*b*) none will pass? (*c*) only 1 will pass? (*d*) What assumption did you make about independence?

EXERCISE 2-24. Four people are shooting at a target. The probability that each will hit the target is $\frac{1}{2}$. What is the probability that, if all 4 shoot, the target will be hit? Note that you cannot add the probabilities. It may be simpler first to compute the probability that none will hit the target.

EXERCISE 2-25. (*a*) What is the probability of obtaining 2 men and 4 women by drawing 6 people at random from a group of 5 women and 5 men? (*b*) What is the approximate probability of drawing 2 men and 4 women from 500 men and 500 women? (*c*) Why the difference in answers?

EXERCISE 2-26. (*a*) What is the approximate probability of drawing 2 bad units in a sample of 400 from a lot of 10,000 items containing 200 bad units? Use the Poisson distribution. (*b*) What is the probability of obtaining more than 2 bad units?

EXERCISE 2-27. It has been found that 10 per cent of simple assemblies have bent covers and won't work. If 30 per cent have bent covers, what is the probability that an assembly with a bent cover won't work?

3. STATISTICAL INFERENCE I

3-1. INTRODUCTION

Mr. Guilfoil, before catching the commuter train home one evening, bought a bag of unshelled peanuts and the evening paper. When he had seated himself, he placed the bag of peanuts on his lap, folded his paper to expose the sports page, and began to read. As he read, he ate peanuts. Being absorbed in the paper, he did not examine each peanut, but popped the nuts into his mouth as rapidly as he could shell them with one hand. When he finished the sports page, he laid the paper aside, removed a nut from the bag, opened it, and examined the contents. The nut meats were wormy. That is, no worm was present, but evidence that one had been there was indisputable. Mr. Guilfoil examined another, with the same results. He examined three more. All showed the same evidence of worminess. Question: If you were Mr. Guilfoil, would you feel fairly certain that you had eaten some wormy peanuts? The logical process leading from sample analysis to a conclusion about a population is called *statistical inference.* It is the very meat of the subject matter of statistics

3-2. THE STATISTICAL HYPOTHESIS

A *statistical hypothesis* is some assertion about a population which may, or may not, be capable of verification. It is easier, as a general rule, to discredit a hypothesis than to verify it; hence the statistician is constantly seeking hypotheses which he can reject, rather than substantiate.

Consider this illustration. Mr. Himes is selling bluegrass seed. You want to buy some. Mr. Himes indicates a barrel of seed and says, "There are no weed seeds in this bluegrass." This is a hypothesis about the population (the barrel of seed). You are skeptical and wish to gather some evidence about the truth of the hypothesis.

You decide to withdraw a pinch of the seed and examine it. You do

so and find no weed seeds. Have you proved the hypothesis? Obviously not. You return the seed to the barrel, stir the contents vigorously, and examine another pinch. Same results. Have you verified the hypothesis? It is pretty clear that repeating this process until either you or the seeds are worn out can never *prove* the hypothesis that there are no weed seeds in the lot.

However, suppose that examination of the first pinch, or of any other for that matter, discloses some weed seeds. Has the hypothesis been disproved? Clearly it has.

Ordinarily, a hypothesis cannot be disproved so clearly. This makes it necessary for us to accumulate evidence about a hypothesis until we have an overwhelming case against it. Then we can reject the hypothesis with small probability of being wrong.

Mr. McQueedy, director of personnel, says, "I can tell whether or not a prospective employee has a college degree as soon as he enters the room." You wish to test him. Assume that job applicants are equally divided between college graduates and others. (If 80 per cent were college graduates, Mr. McQueedy could be right 80 per cent of the time, on the average, by judging each applicant to be a college graduate.)

What constitutes a suitable hypothesis for the test? Suppose we say that Mr. McQueedy can *always* tell. Then any error rejects the hypothesis. We decide, however, that this is unfair and that what Mr. McQueedy really means is that he can tell more than half of the time. Therefore, we set up the hypothesis that he does no better by relying on his judgment than he could do by tossing a coin—heads = graduate, tails = nongraduate. We call this the *null hypothesis.* Mr. McQueedy hopes that we can reject this hypothesis.

We start our sequence of tests. Mr. McQueedy is right on the first trial. Do we reject our hypothesis that he cannot distinguish between graduates and nongraduates? Probably not. After all, his probability of guessing correctly is $\frac{1}{2}$. But he is also correct on the second trial. Now, the probability is $\frac{1}{4}$ that he could guess correctly twice in a row ($\frac{1}{2}$ on the first trial times $\frac{1}{2}$ on the second trial, by the multiplication rule of probability presented in the previous chapter). Suppose we are not yet convinced and give him 8 more trials. Will we reject our hypothesis that he can't distinguish and accept his that he can? Let's have a look at the probabilities before answering. The probability that he would guess correctly 10 times in a row is

$$ {}_{10}C_{10}p^{10}q^0 = \frac{10!}{0!10!}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0 = \frac{1}{1{,}024} $$

Since the chances are less than 1 in 1,000 of guessing 10 correctly in a row, we attribute his success to something else, namely, to a real ability to

make the correct choice. We raise our opinion of Mr. McQueedy accordingly.

3-3. ERRORS IN DECISIONS

Suppose Mr. McQueedy feels that he can make the correct choice $\frac{3}{4}$ of the time and we feel that he can choose correctly only $\frac{1}{2}$ of the time. That is, we think that $p = \frac{1}{2}$, and Mr. McQueedy believes that $p = \frac{3}{4}$. We decide to test the hypothesis that $p = \frac{1}{2}$ and, if we reject it, to accept Mr. McQueedy's hypothesis that $p = \frac{3}{4}$.

A little reflection shows that there are two types of error that we can make. They are:

Type I. Rejection of a true hypothesis. That is, our hypothesis that $p = \frac{1}{2}$ is correct, but we reject it in favor of Mr. McQueedy's, $p = \frac{3}{4}$.

Type II. Acceptance of a false hypothesis. That is, Mr. McQueedy is actually correct, but we accept our own hypothesis that $p = \frac{1}{2}$.

Let α equal the probability of the Type I error and let β equal the probability of the Type II error. Suppose we make 5 trials and adopt this rule of decision: Reject the hypothesis that $p = \frac{1}{2}$ if all 5 are chosen correctly; otherwise accept. In terms of the alternative hypothesis (Mr. McQueedy's), this means that we shall accept his if all 5 are chosen correctly—otherwise we shall accept ours. Computation of α and β will help us evaluate the decision rule.

Remember that α is the probability of rejecting our hypothesis that $p = \frac{1}{2}$ *if this hypothesis is, in fact, true.* Hence α is the probability of getting 5 successes out of 5 trials if $p = \frac{1}{2}$. Using Eq. (2-19), we have

$$\alpha = {}_5C_5\left(\tfrac{1}{2}\right)^5\left(\tfrac{1}{2}\right)^0 = \tfrac{1}{32}$$

That is, we should expect Mr. McQueedy to guess all 5 correctly about once out of 32 tries if he can't actually make any distinction, that is, if $p = \frac{1}{2}$. The probability of accepting our hypothesis, H_0, when the alternative, H_1, is true is β. That is, it is the probability of 0 or 1 or 2 or 3 or 4 successes *if p actually is equal to* $\frac{3}{4}$.

$$\beta = {}_5C_0\left(\tfrac{3}{4}\right)^0\left(\tfrac{1}{4}\right)^5 + {}_5C_1\left(\tfrac{3}{4}\right)^1\left(\tfrac{1}{4}\right)^4 + {}_5C_2\left(\tfrac{3}{4}\right)^2\left(\tfrac{1}{4}\right)^3 + {}_5C_3\left(\tfrac{3}{4}\right)^3\left(\tfrac{1}{4}\right)^2 + {}_5C_4\left(\tfrac{3}{4}\right)^4\left(\tfrac{1}{4}\right)^1$$

by Eq. (2-19) and the addition rule. But these are all the terms of the binomial expansion except the last, so it is easier to compute

$$\begin{aligned}\beta &= 1 - {}_5C_5\left(\tfrac{3}{4}\right)^5\left(\tfrac{1}{4}\right)^0 \\ &= 1 - \frac{3^5}{4^5} = 1 - \frac{243}{1{,}024} = \frac{781}{1{,}024}\end{aligned}$$

Note that in computing α we use our hypothetical value of p, that is, $p = \frac{1}{2}$. In computing β, we use the alternative hypothesis, $p = \frac{3}{4}$. Our risk of accepting the false hypothesis that $p = \frac{1}{2}$ is 781/1,024 if Mr. McQueedy can actually make the correct choice 75 per cent of the time. We may feel that this probability is excessively large.

Is it possible to change the decision rule to reduce this error? Suppose we try this rule: Reject the hypothesis $p = \frac{1}{2}$ if Mr. McQueedy makes 4 or 5 correct choices; otherwise accept.

$$\alpha = {}_5C_4(\tfrac{1}{2})^4(\tfrac{1}{2})^1 + {}_5C_5(\tfrac{1}{2})^5(\tfrac{1}{2})^0 = \tfrac{6}{32}$$

$$\beta = 1 - {}_5C_4(\tfrac{3}{4})^4(\tfrac{1}{4})^1 - {}_5C_5(\tfrac{3}{4})^5(\tfrac{1}{4})^0 = \frac{376}{1{,}024}$$

Be sure to verify these figures before proceeding. We have substantially reduced the probability of accepting a false hypothesis (β is reduced from 781/1,024 to 376/1,024), but we have added substantially to the chances of rejecting a true hypothesis (α increased from $\frac{1}{32}$ to $\frac{6}{32}$)—thus verifying again the old adage that you can't have your cake and eat it too.

We can make the β error zero by always rejecting the hypothesis, regardless of the result. As a decision rule, this leaves a good bit to be desired, because in this case a true hypothesis will always be rejected. That is, $\beta = 0$, but $\alpha = 1$. The fallacy in such a decision rule is obvious.

The set of all possible results can be divided into two categories by the decision rule. The results which cause rejection of the hypothesis are said to be in the *region of rejection.* All others are in the *region of acceptance.* For example, if we adopt the second decision rule above and let y = number of successes, the region of rejection will consist of $y = 4$ and $y = 5$. The region of acceptance will contain $y = 0$, $y = 1$, $y = 2$, and $y = 3$.

The *power of a test,* against a particular alternative, is the probability that the results will lie in the region of rejection of the hypothesis *if the alternative is true.* That is, it is the probability that we shall accept Mr. McQueedy's hypothesis when he is actually correct. Clearly, then, the power of a test equals $1 - \beta$. One test is said to be more powerful than another if, for the same α, the power of the first $(1 - \beta)$ is greater than the power of the second.

3-4. RELATIONSHIP OF DECISION TO COST

One may inquire, what is a good decision rule? The question is not always easy to answer. Sometimes, however, a cost can be assessed against each type of error, in which case one determines the decision rule in such a manner as to minimize this cost (or to optimize gain). Consider the following illustration.

Mr. Bone raises chickens for the market. He has an opportunity to buy 1,000 baby chicks. The hatchery operator tells Mr. Bone that the 1,000 chicks represent either 250 males and 750 females or 750 males and 250 females. His hatches of chicks have become mixed, and he knows it is one or the other, but he doesn't know which combination it actually is. Mr. Bone believes that at the quoted price he can make 15 cents on each male chick, but he expects to lose 10 cents on females. He asks whether he can select 5 chicks at random and test them for sex. He is granted this privilege. What should be his decision rule?

Suppose he sets up the hypothesis that $\frac{3}{4}$ are males. The alternative hypothesis is that $\frac{1}{4}$ are males. If y = number of males in the sample of 5, he may decide to let $y = 0$ be the region of rejection. That is, he will reject the hypothesis that $\frac{3}{4}$ are males if he finds no males among the 5 examined; otherwise he will accept it. This is his decision rule. In this case

$$\alpha = {}_5C_0\left(\tfrac{3}{4}\right)^0\left(\tfrac{1}{4}\right)^5 = \frac{1}{1{,}024}$$

$$\beta = 1 - {}_5C_0\left(\tfrac{1}{4}\right)^0\left(\tfrac{3}{4}\right)^5 = 1 - \frac{243}{1{,}024} = \frac{781}{1{,}024}$$

Note that N (size of the population) is large enough for us to assume that the separate trials are independent. If H_0 ($p = \frac{3}{4}$) is true, Mr. Bone stands to gain \$87.50. This figure is computed as 750 males times 15 cents profit less 250 females times 10 cents loss. (We shall ignore the 5 chicks tested in order to simplify the problem.) If H_1 ($p = \frac{1}{4}$) is true, Mr. Bone will lose \$37.50, that is, 250 males times 15 cents profit less 750 females times 10 cents loss. Since Mr. Bone has no other information about the population, he can logically assume that, without the test, H_0 is just as likely as H_1.

According to Mr. Bone's decision rule he will reject the hypothesis that $p = \frac{3}{4}$ if no males are found in the sample of 5. In other words, he will not buy the lot of chicks. In this case he will neither gain nor lose any money (except for an "opportunity cost" due to failure to buy a good lot of chicks—a consideration that we shall ignore). We show all the possible gains or losses for Mr. Bone in Table 3-1. The table shows that,

Table 3-1. Possible Gains and Losses for Mr. Bone in the Chick-buying Example

Actual proportion of males	Mr. Bone's decision	
	Buy	Don't buy
$\frac{3}{4}$	\$87.50	0
$\frac{1}{4}$	− 37.50	0

if $\frac{3}{4}$ are males and Mr. Bone buys, he will gain \$87.50. If he does not buy, he gains nothing. If $\frac{1}{4}$ are males and he buys, he loses \$37.50. If he does not buy, he gains or loses nothing.

We must introduce the α and β probabilities computed above to evaluate Mr. Bone's decision rule. If $\frac{3}{4}$ of the chicks are males, the probability that he will accept that hypothesis is $1 - \alpha = 1{,}023/1{,}024$. The probability that he will reject that hypothesis is $1/1{,}024$ (which is α). Therefore, if the hypothesis that $\frac{3}{4}$ are males is true, he will, on the average, gain \$87.50 about $1{,}023/1{,}024$ of the time. Hence his *expected gain* is

$$\frac{1{,}023}{1{,}024}(\$87.50) + \frac{1}{1{,}024}(0) = \$87.41$$

This is a conditional expected gain, based upon the condition *that* $\frac{3}{4}$ *are males.* It represents simply the average gain if this condition is true. That is, $1{,}023/1{,}024$ of the time he will gain \$87.50 and $1/1{,}024$ of the time he will gain nothing.

Now suppose that only $\frac{1}{4}$ of the chicks are males. By the decision rule Mr. Bone will accept the false hypothesis that $\frac{3}{4}$ are males with probability $781/1{,}024$ (which is β). He will accept the right hypothesis (and not buy) with probability $1 - \beta$, or $243/1{,}024$. Therefore, his expected loss *if* $\frac{1}{4}$ *are males* is

$$\frac{781}{1{,}024}(-\$37.50) + \frac{243}{1{,}024}(0) = -\$28.60$$

Since both situations are equally likely (by assumption), we can average the two results above to find his expected gain under the given decision rule. His expected gain is

$$\frac{\$87.41 - \$28.60}{2} = \$29.40$$

We wonder whether he can do better. We try the decision rule: Reject H_0 if $y = 0$ or 1; otherwise accept.

$$\alpha = {}_5C_0\left(\tfrac{3}{4}\right)^0\left(\tfrac{1}{4}\right)^5 + {}_5C_1\left(\tfrac{3}{4}\right)^1\left(\tfrac{1}{4}\right)^4 = \frac{16}{1{,}024}$$

$$\beta = 1 - {}_5C_0\left(\tfrac{1}{4}\right)^0\left(\tfrac{3}{4}\right)^5 - {}_5C_1\left(\tfrac{1}{4}\right)^1\left(\tfrac{3}{4}\right)^4 = \frac{376}{1{,}024}$$

Now his expected gain if H_0 is true is

$$(1 - \alpha)(\$87.50) + \alpha(0) = \frac{1{,}008}{1{,}024}(\$87.50) + \frac{16}{1{,}024}(0) = \$86.13$$

His expected loss if H_1 ($\frac{1}{4}$ males) is true is

$$\beta(-\$37.50) + (1 - \beta)(0) = \frac{376}{1{,}024}(-\$37.50) + \frac{648}{1{,}024}(0) = -\$13.77$$

The average of these expected results is

$$\frac{\$86.13 - \$13.77}{2} = \$36.18$$

This is a better decision rule since the average gain is greater.

EXERCISE 3-1. Try other decision rules and find the one which maximizes the expected gain.

EXERCISE 3-2. In a particular taste-testing experiment, the subject is presented with 3 bits of food or drink of which 2 are alike and 1 is different. Whether the 2 alike will be A or B is chosen at random (say, by the toss of a coin). Suppose that a taste panel is composed of 5 members and that the decision rule is to reject the hypothesis that panelists cannot distinguish the odd food if 4 or 5 make the correct choice. (*a*) If a panelist really cannot make the distinction, what is the probability that he will make a correct choice by chance? (*b*) What is α? (*c*) Suppose our alternative hypothesis is that panelists can make the distinction $\frac{2}{3}$ of the time. Compute β. (*d*) What is the power of the test against this alternative? What is the region (*e*) of rejection? (*f*) of acceptance?

3-5. TESTS OF HYPOTHESIS WITH MULTIPLE ALTERNATIVES

In Sec. 3-3 a method was presented for testing a hypothesis $H_0 = p_0$ against a single alternative, $H_1 = p_1$. It must be apparent that the assumption of a single alternative is unduly restrictive. For example, why must we choose $p = \frac{3}{4}$ as Mr. McQueedy's only possible alternative? Why not permit 0.8, 0.6, 0.9, or any other possible value for p? We must, indeed, provide for testing a hypothesis against such general classes of alternatives.

Suppose one is grading a product as grade A or grade B. His hypothesis is that 60 per cent of the items produced are grade A. If he draws a sample of 10 items and grades them, what region of rejection will he choose? The answer obviously will depend upon α, the risk he is willing to run of rejecting a true hypothesis, but it is also dependent upon the alternatives he wishes to guard against.

If he wishes to guard against all possible alternatives to the hypothesis H_0 ($p_0 = 0.6$), then clearly he must reject the stated hypothesis if the number of grade A units in his sample is either very small or very large. However, if he is concerned only with guarding against the alternatives in which p is greater than p_0, he will reject p_0 only if the observed number of grade A units is very high. Conversely, to guard against the possibility

that the true number of grade A units is small, he will reject p_0 if the observed number of grade A units is quite small.

These ideas are pursued further in a later chapter, where the use of continuous probability functions makes the manipulations easier.

EXERCISE 3-3. Assume that you are buying eggs and that you decide to test 100 to see whether they are good or bad. You adopt this decision rule: If there are 0, 1, or 2 bad eggs, accept; if more, reject. What is the probability that you will buy a lot of eggs that is exactly 5 per cent bad? *Hint:* Use the Poisson distribution.

EXERCISE 3-4. Using the data of Exercise 3-3, find a decision rule which will assure you of running a risk of less than 0.05 of accepting a lot of eggs more than 4 per cent bad (*a*) by sampling 100 eggs, (*b*) by sampling 200 eggs.

EXERCISE 3-5. In the above exercises, what costs would you consider in choosing a "best" decision rule?

EXERCISE 3-6. Mr. A manufactures electrical fuses. He decides whether they are good or bad by this rule: Draw a random number from 1 to 100; if it is 95 or under, call the fuse good; if it is 96 or over, call the fuse bad. (*a*) What is the probability that a good fuse will be rejected (i.e., what is α)? (*b*) What is wrong with this decision rule? *Hint:* Compute β.

EXERCISE 3-7. Mr. A discovers that 50 per cent of the fuses with a particular observable characteristic are bad but that only 10 per cent of the others are bad. He decides to call all fuses with the observable characteristic bad. Suppose 20 per cent have the observable characteristic and 80 per cent do not. (*a*) What is α? (*b*) What is β? Remember that α is the probability that a *good* fuse will be rejected.

4. STATISTICAL DESCRIPTION

4-1. MEASUREMENTS

We can never determine the size of anything which must be measured. This fact is shocking to some people. They don't believe it. But suppose you pick up a pencil from your desk, and it appears to be about 6 in. long. This is a first approximation to its length. Then suppose you lay it along a ruler and find that it is about $6\frac{1}{4}$ in., which is your next approximation to its length. Next, you obtain an inexpensive set of calipers and determine that a more exact measurement is 6.251 in. Then, you take the pencil to the machine shop and by careful measurement are able to determine that a more exact length is 6.2508 in., although here there may be some real doubt about the significance of the last decimal place. The reason is that the ends of the pencil are uneven and microscopic examination reveals that they are not smooth, as you might have imagined, but are composed of hills and depressions. Therefore, the measurement you get depends upon the exact position of the measurement. You decide to go no further with the measuring process but to let 6.2508 in. represent an approximation to the length. And that is precisely what every measurement is—an approximation. Even if one obtains a bar of the finest tool steel which is carefully machined, he cannot determine its length. If he is able to carry out its measurement to 10 decimal places (in inches), there is surely an eleventh place which is unknown.

Another thing which is often misunderstood is the significance of the measurement. An illustration may help to clarify the point. A truck loaded with gravel weighs 15 tons. The driver weighs 185 lb. He wears $4\frac{1}{2}$ lb of clothing and carries a mechanical pencil weighing 1 oz. How much does the whole load weigh? The only significant answer, obviously, is 15 tons.

A rough rule to follow is that the sum of a set of figures should be carried

only to the number of decimal places in the addend of least accuracy. Thus the sum of 37, 329.4, 1.256, and 0.0002 should be given as 368, not 367.6562.

Before a rule is given on multiplication, it is desirable to examine what is meant by "significant digits." Each of the following numbers has three significant digits: 2.45, 31.4, 0.00269, 981.0. Now, let us examine a number such as 320,000. Here there is some doubt about the number of significant digits. If one's measuring device is accurate only to the nearest 10,000, then there are two significant digits. But the zeros following 32 may represent true numbers. To avoid this ambiguity, the number may be written as $32(10)^4$ if there are actually two significant digits or as $320(10)^3$ if there are three, and so forth. This specification of precision, called *scientific notation*, is more common in engineering than in management, but is worth mentioning here.

Now for a rule on multiplication: If two approximate numbers are multiplied together, their product is significant to the number of significant digits in the factor having the fewest significant digits. This sounds confusing, but an illustration may help. A hallway is measured and found to be 6.7 by 82.2 ft. How many square feet are in the area? We find 82.2 ft $\times$ 6.7 ft = 550.74 sq ft. By the rule, 6.7 has two significant digits and 82.2 has three. Therefore, the product will have two significant digits and should be called 550 sq ft. Let's see why this is true. The measurement 82.2 could conceivably have been rounded downward from 82.25 or upward from 82.15. Similarly, 6.7 could have been rounded downward from 6.75 or upward from 6.65. Taking the two extremes, we could have

$$82.25 \text{ ft} \times 6.75 \text{ ft} = 555.1875 \text{ sq ft}$$

or

$$82.15 \text{ ft} \times 6.65 \text{ ft} = 546.2975 \text{ sq ft}$$

Therefore, we may as well say that the product is 550 sq ft. There is perhaps some merit in the argument that one is unlikely actually to have the extreme rounding error in the same direction on both measurements. For this reason a third significant digit may have some meaning, but certainly no more digits. Common practice, then, would list the product as 551 sq ft.

Note that the above rule applies if there is rounding (or measurement) error in both factors. If there is error in only one, then the number of significant places in that measurement governs the significant places in the product. Similarly, if there is no error in either, there is no error in the product. For example, suppose a class is composed of 54 persons, none of whom is an amputee. There are certainly *exactly* 54 $\times$ 4 = 216 arms and legs in the class.

EXERCISE 4-1. A machine costs $392,469.18 and is estimated to last 9 years. (*a*) Assuming a straight-line depreciation, how much does the machine depreciate during the first year? (*b*) Why is the figure reported as $43,607.69 on the statement of profit and loss?

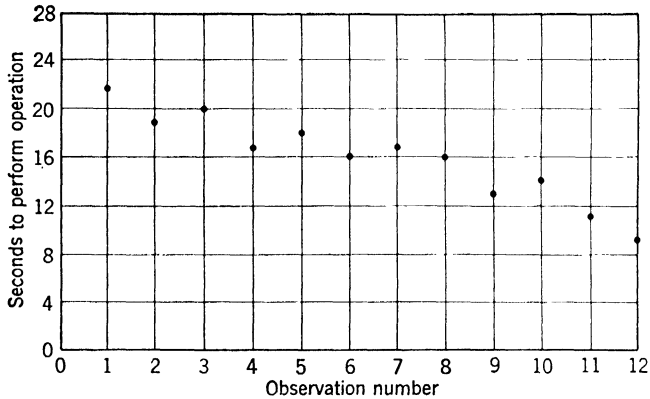

FIG. 4-1. Successive observations on a constantly changing quantity

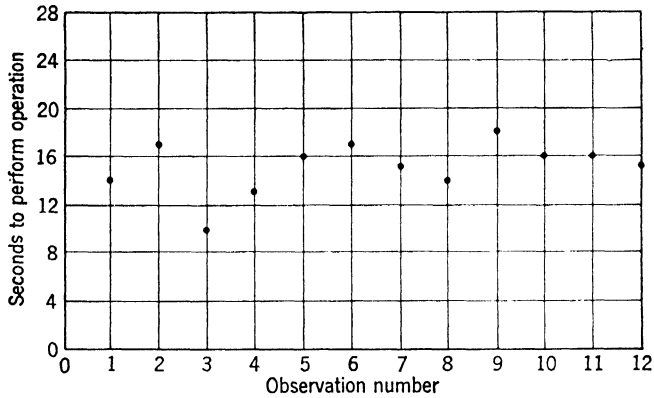

FIG. 4-2. Successive observations on a fixed quantity

Let us consider another aspect of measurement which sometimes is overlooked. We shall suppose that in a time study the following observations, in seconds, are recorded on the length of time required to perform an operation: 22, 19, 20, 17, 18, 16, 17, 16, 13, 14, 11, 9. The observations are listed in the order in which they are taken. When plotted, they appear as in Fig. 4-1. What meaning can be attached to an average of these figures? The answer is none. A look at Fig. 4-1 shows us that the thing we are trying to measure changes from one observation to the next. The average of the 12 observations does not give us much information about the value to be expected on the thirteenth observation. Perhaps the operator is just learning the operation. Perhaps he has

become aware that he is being timed. A number of factors could account for the trend in observed time. In any case, we do not have readings on a constant process, but on a constantly changing quantity.

A series of 12 readings on an unchanging quantity is shown in Fig. 4-2. Here it is clear that there is no growth or decline of the quantity being measured during the 12 observations. The average of these 12 observations has some real meaning in predicting the value to be expected on the thirteenth and succeeding observations. Note that the significant distinction between Figs. 4-1 and 4-2 is the lack of pattern, or "drift," in Fig. 4-2. We direct our attention now to a discussion of the average by which we can characterize the 12 observations.

4-2. THE MEAN

The point of the previous section is that, if a population remains unchanged from observation to observation, there exists a value (a parameter) around which the observations vary in a random manner. This central value we call the *mean of the population* and denote it by the Greek letter μ (mu). Another way of expressing the same thing is to say that each observation, Y_i, in the population can be expressed as the sum of two components, the first being the mean of the population and the second a random deviation from this mean. Expressed algebraically, this statement becomes

$$Y_i = \mu + e_i \tag{4-1}$$

The random elements e_i have the property that their average value *over the whole population* is 0. Another way of saying this is to say that their *expected value* is 0. This does not mean, of course, that the average value of a sample of them will be 0, but only that the average of the whole population will be 0.

Another assumption which we make in the beginning in order to simplify the manipulation is that the e_i are independent from one observation to the next. From our definition of independence in Chap. 2, we know this means that the value of e drawn on a particular observation does not affect the probability that any particular value of e will be drawn on the next trial. The illustrations in Chap. 2 pointed out that, if the population is finite (e.g., 10 items, of which 4 are defective), the probability of drawing a particular class of item (say, defective) changes from trial to trial, depending upon what items were drawn on previous trials. This is a characteristic of sampling from finite populations. Therefore, we shall assume for the present that all sample items are drawn from an infinite population, in order to retain the independence assumption. This assumption will be relaxed in a later chapter.

The mean of the population, represented by μ in Eq. (4-1), is almost always unknown. It is, however, a parameter which we wish to estimate. In the illustration in the previous section it was assumed that some observations were taken on the length of time required for a worker to perform an operation. In this case we are interested in the *average* length of time, or expected length of time, to perform the operation rather than the length of time it takes in this trial or the next, or in the following 20 trials. In other words we want to know the *real* average time, which is the parameter μ. We can never actually determine the true average, however, so we must settle for an estimate based upon the sample observations.

When we consider the problem of estimating the population average, our intuition may tell us to use as an estimate the average of the sample observations. This is, in fact, an excellent estimate, and we shall discuss some of its properties.

When we were in grade school, we learned to compute the "average" by adding up the figures to be averaged and dividing by the number of figures added. Symbolically, we denote the average by $\bar{Y}$ and call it the *arithmetic mean*. Its formula is

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{4-2}$$

where n is the sample size (number of figures added), the Y_i are the observations, and the sigma stands for "sum of." In words, the summation symbolism tells us to sum up the Y_i from i equals 1 to i equals n. This simply means that if we have n Y values we add up all n of them. Since in most statistical work we add over all the n values, it is customary to omit the subscripts and superscripts on Σ and to write

$$\bar{Y} = \frac{\Sigma Y_i}{n} \tag{4-3}$$

Summations appear in so many statistical formulas that it is worthwhile devoting some time to an understanding of them. Suppose we have the following Y values: 10, 7, 8, 9, 10. We make the following identification: $Y_1 = 10$, $Y_2 = 7$, $Y_3 = 8$, $Y_4 = 9$, $Y_5 = 10$. Then

$$\Sigma Y_i = 10 + 7 + 8 + 9 + 10 = 44$$

The order in which computations are performed is important. For example,

$$\Sigma Y_i^2 = 10^2 + 7^2 + 8^2 + 9^2 + 10^2 = 394$$

but $(\Sigma Y_i)^2 = 44^2 = 1{,}936$. Also, the formula $\Sigma(Y_i - 10)^2$ tells us first to subtract 10 from each Y, then to square the results, and finally to add them up. That is,

$$\begin{aligned}\Sigma(Y_i - 10)^2 &= (10 - 10)^2 + (7 - 10)^2 + (8 - 10)^2 \\ &\qquad + (9 - 10)^2 + (10 - 10)^2 \\ &= (0)^2 + (-3)^2 + (-2)^2 + (-1)^2 + (0)^2 \\ &= 0 + 9 + 4 + 1 + 0 = 14\end{aligned}$$

EXERCISE 4-2. (*a*) Compute $\bar{Y}$ for the above data. (*b*) Compute $\Sigma(Y_i + 2)^2$, $\Sigma(Y_i^2 - Y_i + 2)$, and $n\Sigma Y_i^2 - (\Sigma Y_i)^2$.

Suppose we are asked to add up a constant. For example, to find $\sum_{i=1}^{5} 10$, we see that we have simply $10 + 10 + 10 + 10 + 10 = 50$. In general, if we denote a constant by C, then $\sum_{i=1}^{n} C = nC$. But suppose the constant is multiplied by a variable, as follows:

$$\begin{aligned}\Sigma CY_i &= CY_1 + CY_2 + \cdots + CY_n \\ &= C(Y_1 + Y_2 + \cdots + Y_n) = C\Sigma Y_i\end{aligned}$$

Thus we see that the constant can be taken outside the summation sign and only the variable part added.

EXERCISE 4-3. Compute $\Sigma 10(Y_i - 2)^2$ for the data of Exercise 4-2.

We return now to the arithmetic mean and observe what happens if we take the mean of the Y_i of Eq. (4-1). We have

$$\bar{Y} = \frac{\Sigma Y_i}{n} = \frac{n\mu}{n} + \frac{\Sigma e_i}{n} = \mu + \bar{e} \tag{4-4}$$

where $\bar{e}$ represents an average of the errors, or random deviations. The idea behind this expression is quite simple; yet it is extremely important to statistical methodology and theory. The idea is that a sample mean is equal to the population mean plus an average of the errors associated with the sample observations. Only by pure chance would $\bar{e}$ ever be 0. In fact, the probability that $\bar{e}$ is 0 is ordinarily so small that it can be ignored.

The expression $Y_i = \mu + e_i$ is called a *model* of the population from which the sample values are drawn. A numerical example may assist in understanding the nature of the arithmetic mean of a sample drawn from a population which is described by this model. We must specify some value for μ, so suppose we say that $\mu = 17$. Since the e_i are random variables, they must follow some probability distribution (see Sec. 2-5).

Let us assume that the e_i follow the simple probability distribution given in Table 4-1.

Table 4-1. Hypothetical Probability Distribution for the Random Errors of Model 4-1

e_i	$P(e_i)$
−5	0.1
−1	0.5
0	0.2
5	0.2

We have specified that the e_i have an average value of 0. We can verify that the above distribution has an average value (expected value) of 0 by computing

$$E(e_i) = \Sigma e_i P(e_i) = -5(0.1) - 1(0.5) + 0(0.2) + 5(0.2) = 0$$

It may be worth noting here that this is a general form for an expected value, that is, the sum of the variables times their associated probabilities.

We wish to obtain some random drawings from our hypothetical population. We may obtain them by drawing numbers from the table of random numbers, Table 1-1, by identifying digit 1 with $e = -5$, digits 2, 3, 4, 5, and 6 with $e = -1$, digits 7 and 8 with $e = 0$, and digits 9 and 0 with $e = 5$. This is equivalent to saying that we are drawing the e_i with the probabilities listed as $P(e_i)$ in Table 4-1. This is true because −5 is given 1 chance out of 10 of being included, −1 is given 5 chances out of 10, and so on.

Suppose we start with the first digit of row 6, column 5, of Table 1-1, and decide to proceed to the right, taking each digit as it comes. Then we can simulate the random drawing of observations listed in Table 4-2.

Table 4-2. Simulated Random Drawings from Model (4-1): $Y_i = \mu + e_i$

Random digit	e_i	μ	$\mu + e_i = Y_i$
7	0	17	17
1	−5	17	12
7	0	17	17
0	5	17	22
5	−1	17	16
8	0	17	17
3	−1	17	16
1	−5	17	12
4	−1	17	16
6	−1	17	16
Sum	−9	170	161

In the first row and first column of Table 4-2 is listed the random digit 7, which we have decided will denote the drawing of $e_i = 0$, and so forth. We note that an average of the Y's yields a mean of 16.1, that the true population mean is 17, and that the difference, -0.9, is the average of the random deviations. As the sample size increases, we should expect the average of the errors, $\bar{e}$, to come close to 0.

EXERCISE 4-4. Continue the simulation of random drawings for 40 more trials and compute $\bar{Y}$ and $\bar{e}$ for the 50 observations (including the 10 above).

EXERCISE 4-5. Draw a vertical-bar diagram, similar to Fig. 2-3, of the probability distribution of the random errors.

One of the important properties of the arithmetic mean is that the sum of the deviations from it is 0. For example, consider the last column of Table 4-2 and compute the sum of deviations from the mean for this column. The computations are shown in Table 4-3. In fact, this can be used as a definition of the arithmetic mean. It is that figure such that the sum of the deviations from it is equal to zero.

Table 4-3. Computation of the Sum of Deviations and Sum of Squares of Deviations from the Mean

Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
17	0.9	0.81
12	−4.1	16.81
17	0.9	0.81
22	5.9	34.81
16	−0.1	0.01
17	0.9	0.81
16	−0.1	0.01
12	−4.1	16.81
16	−0.1	0.01
16	−0.1	0.01
Sum	0.0	70.90

It should be noted that saying the sum of deviations equals 0 is not the same thing as saying that the arithmetic mean is the central value in an array of numbers. For example, in the array 2, 5, 9, 13, 20, the central value is 9, but the arithmetic mean is 9.8. The central value is called the *median* in statistics.

Another property of the arithmetic mean is that the sum of squares of deviations from it is less than the sum of squares of deviations from any

other point. Refer again to Table 4-3. We cannot find any point other than 16.1 such that the sum of squares of deviations from it will be less than 70.90. This is called the *property of least squares.* Its importance will be apparent in a later section.

EXERCISE 4-6. Consider some point other than 16.1 (say, 16), compute the sum of squares of deviations from it, and compare with 70.9.

Another property of the arithmetic mean is that it is an unbiased estimate of the population parameter μ if μ exists. By an unbiased estimate of μ, we mean one which, on the average, will be equal to the population parameter μ. In any given case, $\bar{Y}$ may be above or below μ, but if one were to take many samples from a fixed population, computing $\bar{Y}$ for each sample, the average of the $\bar{Y}$'s would approach the value of μ.

4-3. THE VARIANCE AND STANDARD DEVIATION

As indicated earlier, variation is the heart of statistics. Without it we should be dealing only with constants, and no statistical computations would be necessary. A measure of the amount of variation is called a *measure of dispersion.*

A measure of dispersion not only helps to describe the frequency distribution but also tells us something about how accurately the mean of a sample estimates the mean of a population. We can demonstrate this principle by the following illustration. Consider two populations, Z and W, each consisting only of the values given below.

Population Z		*Population W*	
5	5	1	400
5	5	10	800
5	5	50	1,000
5	5	100	2,000
5	5	200	4,000

A sample of 2 taken from population Z will yield precisely the population mean (5), and a sample of 2 taken from population W may yield a mean far different from the population mean of 856.1. It may, in fact, vary from $\bar{Y} = (1 + 10)/2 = 5.5$ to $\bar{Y} = (2{,}000 + 4{,}000)/2 = 3{,}000$. We say, then, that a sample mean drawn from population Z is *more reliable* than a sample mean from population W.

To consider another case, suppose we take 10 random samples of concerns from each of two industries with the object of estimating the

average rate of absenteeism (in days missed per 100 working days) in each of the industries. We record the following observations:*

Industry A		*Industry B*	
3.75	5.04	3.47	3.90
5.16	3.98	0.22	4.11
4.65	4.55	8.48	5.99
4.27	4.74	4.05	6.88
4.47	5.50	3.58	5.43

The average absenteeism rate in industry A is 4.611, with little variation in the sample observations (3.75 to 5.50). The average rate in industry B is also 4.611, with wide variation (0.22 to 8.48). In the case of A we should expect the mean of the sample to be more representative of the mean of the population from which it was drawn than in the case of B, because additional sample units drawn from industry B, with their indicated extreme variations, would be expected to exert greater influence on the sample mean than would additional sample units from industry A, with their apparent small variation. Here we are assuming, of course, that the variation in the sample is indicative of the variation in the population. Devices are given later for testing the "significance" of observed differences in variation.

There are a number of devices through which the statistician can place a numerical value on the amount of dispersion. By far the most common measures are the variance and its square root, the standard deviation. For a finite population we can define *variance* as the average of the squares of deviations from the arithmetic mean. We may write the formula for the variance as

$$\sigma^2 = \frac{\Sigma(Y - \mu)^2}{N} \tag{4-5}$$

where N is the number of individuals in the population and μ is the population mean. The *standard deviation* of a finite population then becomes

$$\sigma = \sqrt{\frac{\Sigma(Y - \mu)^2}{N}} \tag{4-6}$$

where the positive root is always taken. These measures apply to populations and are, therefore, *parameters*. Suppose we apply formula

* It is assumed that the concerns all have approximately the same number of employees. Otherwise the rates would have to be weighted by number of employees in the averaging process.

(4-5) to populations Z and W given above. We find that the variance of population Z is 0 and that the variance of population W is 1,452,353.

EXERCISE 4-7. Verify these computations.

For infinite populations the definitions of σ^2 and σ rely upon the integral calculus and will not be given here.* Perhaps it is sufficient to state that in practically all the infinite populations with which the statistician works, the variance σ^2 exists (although it may be unknown) and may be estimated from sample data.

The idea of an infinite population may need some clarification. The farms in Iowa, the retail merchants in Philadelphia, and the grains of sand on Coney Island all are finite populations. On the other hand, the number of times a coin can be tossed and the number of babies born or to be born are infinite populations, because they are boundless. For statistical purposes, very large finite populations can be handled by the same statistical procedures as infinite populations. Since infinite populations are somewhat simpler to handle, the discussion in the earlier chapters of this book is largely limited to infinite populations.

It must be remembered that σ^2, whether referring to a finite or an infinite population, is a parameter. If we have an infinite population† and wish to estimate σ^2 from a sample drawn from that population, we may do so by computing

$$s^2 = \frac{\Sigma(Y - \bar{Y})^2}{n - 1} \tag{4-7}$$

where $\bar{Y}$ is the sample mean and n is the number of observations in the sample. The estimated standard deviation is then $s = \sqrt{s^2}$.

We say that s^2 is an *unbiased* estimate of σ^2. We have discussed bias briefly before. If s^2 is an unbiased estimate of σ^2, then, *on the average*, s^2 will equal σ^2. In any particular case, however, s^2 may be above or below σ^2. If, on the average, an estimator is greater than the parameter it is designed to estimate, we say it is biased upward. If, on the average, it is too small, we say it is biased downward. It should be noted that, even though an estimator is biased downward, occasional values of the esti-

* For the mathematically inclined reader, the variance of a continuous probability distribution is given by

$$\sigma^2 = \int (y - \mu)^2 f(y)\, dy$$

where $f(y)$ is the functional form of the probability distribution, and the integral is over the range of y.

† For a finite population of size N, we may obtain an unbiased estimate of σ^2 by the formula

$$s^2 = \frac{\Sigma(Y_i - \bar{Y})^2}{n - 1} \frac{N - 1}{N}$$

mator may be too large. That is, if one shoots at a target with a rifle whose sights are not set correctly, the rifle may, on the average, shoot to the left of the target, but an occasional shot may go to the right of the target because of inaccuracies in aiming and imperfections in the rifle and its ammunition.

In spite of the fact that s^2 is an unbiased estimate of σ^2, s is not an unbiased estimate of σ. A correction for bias applicable to normal populations and depending upon the sample size can be applied.* In any case, we are interested primarily in an estimate of the variance and only incidentally in its square root, the standard deviation.

We note that the formula for the estimate (4-7) differs from the formula for the parameter (4-5) by the replacement of μ by $\bar{Y}$ and N by $n - 1$. Division by $n - 1$ is called division by degrees of freedom. The concept of degrees of freedom is a mathematical one. The number of degrees of freedom is a property of a sum of squares, being determined by the number of independent linear comparisons which can be made among the n observations. In order to understand this definition, one would need to define *linear comparison* and *independence*. Rather than devote space to these definitions, we shall merely observe the mathematical fact that only $n - 1$ independent linear comparisons can be made among n things. Hence the sum of squares $\Sigma(Y_i - \bar{Y})^2$ has $n - 1$ degrees of freedom associated with it. On the other hand, the sum of squares $\Sigma(Y_i - \mu)^2$ has n degrees of freedom, because it contains one more possible comparison, that of $\bar{Y}$ with μ. For our purposes, we need only remember that the number of degrees of freedom associated with the variance is the number of observations, n, minus the number of statistics estimated from them. Thus $\Sigma(Y_i - \mu)^2$ has n degrees of freedom, and $\Sigma(Y_i - \bar{Y})^2$ has

* For a mathematical discussion, see A. Hald, *Statistical Theory with Engineering Applications*, pp. 299, 300, John Wiley & Sons, Inc., New York, 1952. Correction factors for selected values of n are as follows:

n	c
2	1.253
3	1.128
4	1.085
6	1.051
8	1.036
10	1.028
20	1.013
50	1 005
100	1.002

The correction is made by multiplying s by c. Note that for moderate sample sizes the correction is quite small.

$n - 1$ degrees of freedom, because, in the latter case, μ has been estimated from the data by $\bar{Y}$.

The proof that $\Sigma(Y - \bar{Y})^2/(n - 1)$ is an unbiased estimate of σ^2 is mathematically beyond the scope of the discussion at this point. However, we can verify arithmetically, for a particular case, that this is true. Incidentally, we shall verify, at the same time, that $\Sigma Y/n$ is an unbiased estimate of μ.

Consider a finite population consisting entirely of the 3 values 1, 8, and 9. Its mean, variance, and standard deviation are computed in Table 4-4. Now suppose we draw samples of 2 from this population.

Table 4-4. Computation of Mean, Variance, and Standard Deviation of Finite Population 1, 8, 9

Y	$Y - \mu$	$(Y - \mu)^2$
1	−5	25
8	2	4
9	3	9
18	0	38

$$\mu = \frac{\Sigma Y}{N} = \frac{18}{3} = 6$$

$$\sigma^2 = \frac{\Sigma(Y - \mu)^2}{N} = \frac{38}{3} = 12\tfrac{2}{3}$$

$$\sigma = \sqrt{12\tfrac{2}{3}} = 3.559$$

One might imagine that we write the numbers 1, 8, and 9 on 3 disks and put them in a hat. A sample of 2 is constructed by drawing a disk, recording its value, replacing it in the hat, drawing again, and recording the value of the second disk (which might, in fact, be the same value as the first).* There are 9 possible different samples which might be obtained, and they are all equally likely to occur. They are listed in Table 4-5 with their means and variances. The variances are computed by the usual formula:

$$s^2 = \frac{\Sigma(Y - \bar{Y})^2}{(n - 1)}$$

For example, the variance of sample 2 is

$$\frac{(1 - 4.5)^2 + (8 - 4.5)^2}{1} = 24.50$$

* Disks are replaced after each drawing in order to make the population inexhaustible, or infinite.

The student can easily verify the remaining computations. Now if we compute the mean of the $\bar{Y}$ column (the average of the sample means), we obtain $\Sigma\bar{Y}/m = \frac{54}{9} = 6$, where m is the total number of different samples. This average is equal to the population parameter μ. If we compute the average of the sample variances, we have $\Sigma s^2/m = \frac{114}{9} = 12\frac{2}{3}$, which is precisely the population parameter σ^2. Note, however, that if we compute the average of the sample standard deviations s, we have $\Sigma s/m = 22.628/9 = 2.514$, rather than the population parameter $\sigma = 3.559$. This is an arithmetic demonstration of the fact that s is not an unbiased estimate of σ.

Table 4-5. Computation of Sample Means and Variances of All Possible Samples from the Finite Population 1, 8, 9 (with Replacement to Simulate an Infinite Population)

Sample	Sample values	$\bar{Y}$	s^2	s
1	1, 1	1.0	0	0
2	1, 8	4.5	24.5	4.950
3	1, 9	5.0	32.0	5.657
4	8, 1	4.5	24 5	4 950
5	8, 8	8.0	0	0
6	8, 9	8.5	0.5	0.707
7	9, 1	5.0	32.0	5 657
8	9, 8	8.5	0 5	0.707
9	9, 9	9.0	0	0
Total		54.0	114.0	22 628

Leaving theoretical considerations for the moment, let us consider an estimate of the population variance based upon the sample values taken from the illustration of rates of absenteeism in industry A. The computations are given in Table 4-6. Note that the mean was rounded from 4.611 to 4.61 before subtraction from the Y value. This rounding error accounts for the fact that the $Y - \bar{Y}$ column does not sum to 0.

Similar computations for industry B yield $\bar{Y} = 4.61$; $s^2 = 5.0452$; and $s = 2.25$. Comparison of the two standard deviations (0.53 and 2.25) shows us immediately that the rates in industry B are more variable than the rates in industry A. Keep in mind, however, that these figures are both sample estimates of population dispersion and hence are subject to sampling error. Methods are presented later for judging whether the two variances differ significantly. As a matter of fact, they do differ significantly, which means that such a big difference as the one observed (0.2859 vs. 5.0452) would not be likely to occur by chance if the variances of the two populations were equal. Read this over again. The idea is important.

Table 4-6. Computations of Variance of Absenteeism Rates in Industry A

Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
3.75	−0.86	0.7396
5.16	0.55	0 3025
4.65	0.04	0.0016
4.27	−0.34	0.1156
4.47	−0.14	0.0196
5.04	0.43	0.1849
3.98	−0.63	0.3969
4.55	−0.06	0.0036
4 74	0 13	0.0169
5.50	0.89	0.7921
46.11	0.01	2.5733

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{46.11}{10} = 4.61$$

$$s^2 = \frac{\Sigma(Y - \bar{Y})^2}{n - 1} = \frac{2.5733}{9} = 0.2859$$

$$s = \sqrt{0.2859} = 0.53$$

4-4. SHORT CUTS IN COMPUTATION

The computation of the variance according to the formula given above is tedious, because of the necessity for taking deviations from the arithmetic mean. Since the mean is seldom integral, squares of decimal figures are involved. This tedium can be avoided by devising formulas which permit computation from the original Y values rather than deviations of the Y's from the mean.

Let us consider n ungrouped sample values, $Y_1, Y_2, \ldots, Y_n$. The following formula, though it appears more complex than (4-7), is actually simpler to use:

$$s^2 = \frac{n\Sigma Y^2 - (\Sigma Y)^2}{n(n - 1)} \tag{4-8}$$

Exercise 4-8. Derive formula (4-8), starting from (4-7).

The use of Eq. (4-8) is demonstrated in Table 4-7. It may be noted that both the numerator and the denominator of Eq. (4-8) are exact numbers; hence the only approximation is introduced by the final division operation. From the standpoint of accuracy, then, Eq. (4-8) is to be preferred to Eq. (4-7), which introduces inaccuracy by virtue of taking deviations from $\bar{Y}$, which is an approximate calculation.

One obvious way of reducing the amount of computation is to group the observations into classes by size of observation before any computa-

Table 4-7. Computation of Variance by Alternative Formulas

Y	Y^2	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
10	100	−2.6	6.76
12	144	−0.6	0.36
13	169	0.4	0.16
13	169	0.4	0.16
15	225	2.4	5.76
63	807		13.20

$$s^2 = \frac{n\Sigma Y^2 - (\Sigma Y)^2}{n(n-1)} = \frac{5(807) - (63)^2}{5(4)} = 3.30$$

$$s^2 = \frac{\Sigma(Y - \bar{Y})^2}{n-1} = \frac{13.20}{4} = 3.30$$

Table 4-8. Sizes of Families, United States, 1950

Number of persons in family	*Number of families, in thousands*
2	13,084
3	9,984
4	8,228
5	4,434
6	2,136
7 or more	1,956
Total	39,822

SOURCE: *Statistical Abstract of the United States*, p. 46, 1952.

tions are performed. Such a grouping is called a *frequency distribution.* An illustration which shows the number of families of various sizes in the United States according to the 1950 census is presented in Table 4-8. A *family* is defined as a group of two or more persons related by blood, marriage, or adoption and residing together. The data appear in graphic form in Fig. 4.3.

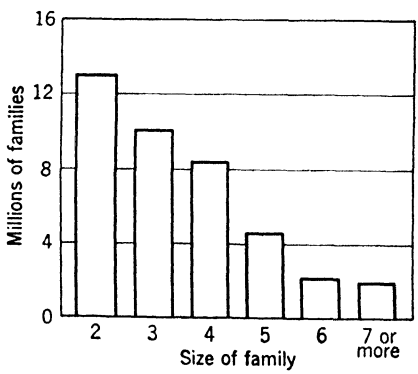

FIG. 4-3. Size of families, United States, 1950

In this illustration, size of family is an exact number (i.e., 2, 3, 4, etc.). The term *discrete* is sometimes used for this sort of variable. In contrast, a *continuous* variable can take on infinitely small graduations in size (e.g., lengths of human hairs).

Table 4-9. Families and Unrelated Individuals by Income Level, United States, 1949

Income level, dollars	*Number of families, in thousands*
Under 500	6,336
500– 999	4,272
1,000–1,499	3,882
1,500–1,999	3,615
2,000–2,499	4,274
2,500–2,999	3,931
3,000–3,499	4,447
3,500–3,999	3,428
4,000–4,499	2,783
4,500–4,999	1,883
5,000–5,999	2,990
6,000–6,999	1,598
7,000–9,999	1,760
10,000 and over	1,111
Total	46,310

SOURCE: *Statistical Abstract of the United States*, 1952.

NOTE: Income not reported by 3,270,000 families and unrelated individuals.

When the number of different size groups is large (or if the variable is continuous), the construction of a frequency distribution becomes tedious unless the number of observations is small. Statisticians, being lazy and intelligent human beings, use the *class interval* in such cases. That is, they group similar-sized observations together and record only the numbers of such observations in each group. Table 4-9 shows an example. The data appear in graphic form in Fig. 4-4.

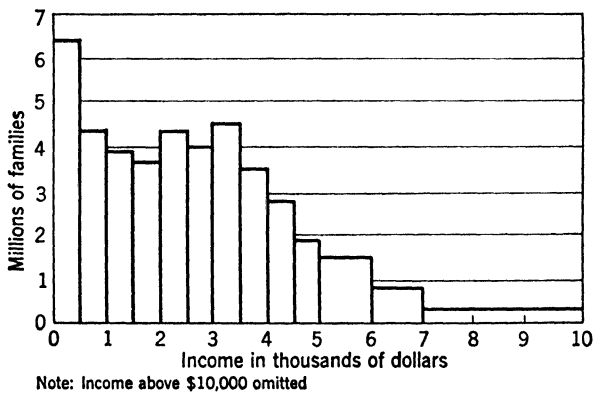

FIG. 4-4. Families and unrelated individuals, United States, by income level, 1949

Grouping into class intervals has two important advantages: (1) it reduces the data to workable size (imagine working with 46,310,000 separate incomes), and (2) it makes it possible for the researcher to get a picture of the distribution of income data. Note that the change in width of class interval necessitates a change in vertical scale in Fig. 4-4. The first 10 class intervals have a width of $500, the next 2 have a width of $1,000, and the thirteenth has a width of $3,000.

To explain the nature of the required adjustment, consider the interval $5,000 to $5,999, a width of $1,000. This interval contains 2,990,000 observations (families). If it were divided into two intervals of width $500, each interval would have approximately half of the 2,990,000 observations, or approximately 1,495,000 observations. Therefore, in order to graph the data on a comparable scale, the frequencies for the intervals of width $1,000 must be divided by 2. The frequency of the last interval, of width $3,000, must be divided by 6. Why?

Note also that these data have been plotted in vertical columns which are shoved up against each other, in contrast to Fig. 4-3, which shows vertical bars with a space in between. Figure 4-3 is called a *vertical-bar diagram*. Generally vertical bars are used to plot discrete data, and a *vertical-column diagram* (a histogram) is used to plot class-interval data.

Grouping observations into class intervals introduces error. In fact the error is similar to that introduced by rounding of observations. It is wise to devote some attention to the method of grouping so that undue error or bias is not introduced.

Consider a set of data reported in integers (whole numbers), or rounded to integers, and suppose that class intervals are to be set up with a width of 5 units. Two methods, which we shall call method A and method B, are in vogue for showing such class intervals. They are as follows:

Method A	*Method B*
5– 9.9	4.5 a.u. 9.5
10–14.9	9.5 a.u. 14.5
15–19.9	14.5 a.u. 19.5
20–24.9	19.5 a.u. 24.5

NOTE: The abbreviation "a.u." means "and under."

The notation of method A implies that the first interval includes 5 and all values greater than 5 but less than 10. This would include, in particular, a number such as 9.997. We are assuming, however, that the data are rounded to integers, or are integers at the start. If the data are integers, the values which can lie in the first class interval of method A are 5, 6, 7, 8, 9. Their average (as well as their mid-point) is 7; hence it is logical to

let the value 7 represent every frequency in the interval. The value which is assigned to all cases in the interval is called the *class mark.*

Now, suppose the data have been rounded to integers. Any value between 4.5 and 9.5 would be rounded to an integer in the interval. The mid-point between 4.5 and 9.5 is 7, so again 7 is the logical number to represent the first interval, 12 the second interval, 17 the third interval, and so on. Note that methods A and B have precisely the same meaning.

One may feel that 7, 12, 17, and so forth, are awkward class marks. If so, he may select 5, 10, 15, and so on, by setting up the intervals as follows: 2.5 a.u. 7.5, 7.5 a.u. 12.5, 12.5 a.u. 17.5, and so forth.

Table 4-10. Gross Income of 50 Retail Drugstores, with Computation of Mean and Variance

Class intervals, thousands of dollars	Mid-point Y	Frequency f	fY	fY^2
10 a.u. 15	12.5	3	37.5	468 75
15 a.u. 20	17 5	5	87 5	1,531 25
20 a.u. 25	22.5	12	270 0	6,075.00
25 a.u. 30	27.5	9	247.5	6,806.25
30 a.u. 40	35 0	7	245.0	8,575.00
40 a.u. 50	45 0	5	225 0	10,125.00
50 a.u. 60	55.0	4	220.0	12,100.00
60 a.u. 80	70 0	3	210.0	14,700.00
80 a.u. 100	90 0	0	0	0
100 a.u. 120	110.0	1	110.0	12,100.00
120 a.u. 140	130 0	0	0	0
140 a.u. 180	160.0	1	160.0	25,600.00
Total		50	1,812.5	98,081.25

$$\bar{Y} = \frac{\Sigma fY}{n} = \frac{1,812.5}{50} = 36.25$$

$$s^2 = \frac{n\Sigma fY^2 - (\Sigma fY)^2}{n(n-1)} = \frac{50(98,081.25) - 3,285,156.25}{50(49)}$$

$$= \frac{4,904,062.5 - 3,285,156.25}{2,450} = \frac{1,618,906.25}{2,450}$$

$$= 660.78$$

$$s = 25.7$$

Sometimes, as a matter of convenience, one accepts the point halfway between two consecutive lower limits as the class mark. For example, in Table 4-9, the class mark for the second class interval would be recorded as \$750 rather than the more exact figure \$749.50, which would arise from grouping data rounded to the nearest dollar. The error is, of course, trivial in this case.

One of the reasons for grouping data is to facilitate computation. With the widespread use of modern calculating equipment the necessity

for grouping to reduce computation is becoming less acute. The student should beware of grouping sample data where the number of observations is small, say, less than 100. The most important reason for grouping sample data is to get an idea of the shape of the frequency distribution through graphing (see Figs. 4-3 and 4-4). The grouping of comparatively few sample observations can be effective for this purpose.

We consider next some computing formulas for grouped data. If we let Y equal the exact value of the discrete variable, or the class mark of the class interval, then the formulas for means and variances become

$$\bar{Y} = \frac{\Sigma fY}{n} \qquad \text{where } n = \Sigma f \tag{4-9}$$

$$s^2 = \frac{n\Sigma fY^2 - (\Sigma fY)^2}{n(n-1)} \tag{4-10}$$

An example of the computations is given in Table 4-10.

A useful short cut is to subtract a constant from each value (or class mark) before performing the computations. The variance is not affected by this procedure, and the correct arithmetic mean can be obtained easily by adding the subtracted constant to the mean of the new variables.

EXERCISE 4-9. Given the frequency distribution

Y	f
85	2
95	8
105	24
115	31
125	16
135	5

compute the mean and variance (*a*) by using formulas (4-9) and (4-10), (*b*) by subtracting 100 from each Y value and recomputing.

4-5. THE STANDARD DEVIATION AND THE NORMAL DISTRIBUTION

It seems desirable at this point to introduce some geometric concept of the standard deviation. We recall that the arithmetic mean is a point on the scale of values which in some sense is representative of the distribution being averaged. The standard deviation represents a *distance* on the scale of values which is indicative of the amount of dispersion. The rate of absenteeism in industry B (Fig. 4-5) can be used to illustrate the concept. We see that all the rates are included within a band about the mean which extends from 2 standard deviations on the negative side to 2 standard deviations above the mean. We cannot jump to the con-

clusion that in every distribution all observations will be included within $\pm 2s$ from the mean. We can say, however, that for any population having a finite mean and standard deviation the probability of drawing an observation at random a distance as much as $k\sigma$ away from the mean is equal to or less than $1/k^2$.† For example, the probability of drawing a sample value more than 3 standard deviations away from the mean is

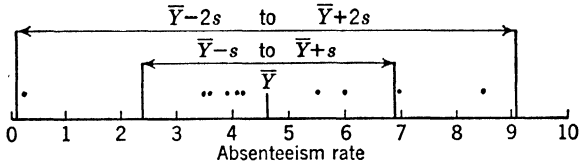

FIG. 4-5. Absenteeism rates in industry B, showing relationship of standard deviations to dispersion

equal to or less than $\frac{1}{9}$, regardless of the shape of the population distribution. (Note that the rule applies to population parameters σ rather than to sample estimates s.)

If we know the exact shape of the population from which the sample observations were drawn, we can often make a much more precise statement about the probability of a randomly drawn observation falling within any specified interval. This is particularly true when the population is "normal." In statistics the term *normal distribution* refers to a particular probability distribution described by the equation

$$f(y) = (1/\sigma\sqrt{2\pi})\, e^{-(y-\mu)^2/2\sigma^2} \qquad (4\text{-}11)$$

The quantities π and e are the well-known mathematical constants. The parameters μ and σ are the mean and standard deviation, respectively. A picture of the distribution is shown in Fig. 4-6.

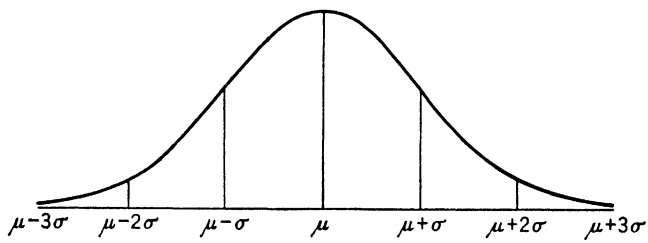

FIG. 4-6. The normal distribution

A population is said to be normally distributed if it can be described by Eq. (4-11). Actually, the populations that we meet in practice are not normal, but many of them are approximately so, and this is what makes the normal distribution so valuable to us.

The functional form of the normal curve was derived mathematically

† This is called *Tchebycheff's inequality.*

early in the eighteenth century as the limiting form of the binomial expansion. Toward the end of the eighteenth century much work was done which showed that the normal distribution describes the behavior of errors in measurement, that is, that repeated measurements of the same quantity tend to be normally distributed. This information was of particular interest to the astronomers of the time. These applications of the normal distribution led to the name *normal curve of error*, a term which still appears in some statistical writings. The wide applicability of the distribution led some to believe that it represented some universal law of error or chance. We realize now that it is only one of many useful probability distributions, but it still retains its position of first importance.

The significance of the standard deviation to a population which is normally distributed is shown in Fig. 4-6. The figure shows that about $68\frac{1}{4}$ per cent of the area of the normal curve lies in the interval from 1 standard deviation below the mean to 1 standard deviation above the mean. About $95\frac{1}{2}$ per cent of the area lies between 2 standard deviations below the mean to 2 standard deviations above the mean. About $99\frac{3}{4}$ per cent of the area lies within the range of plus and minus 3 standard deviations. Since the normal curve extends infinitely far in both directions, no finite number of standard deviations would include *all* of the area under the curve, but so little lies outside of, say, 4 standard deviations that ordinarily we ignore it.

Using area as our measure from which to compute probability (according to the definitions of Chap. 2), we can translate the above percentages directly into probabilities. That is, if we draw an observation (sample value) from a normal population, the probability that it will lie between $\mu - \sigma$ and $\mu + \sigma$ is 0.6826. (We shall see shortly how we get this figure.) Likewise, the probability of drawing a sample observation whose value is more than 3 standard deviations away from the mean is 0.0027 (approximately $\frac{1}{4}$ of 1 per cent). Clearly this is an unusual event.

Now let us see how the above probabilities are read from a table. Suppose we think of a normal distribution whose mean is 0, whose standard deviation is 1 (unity), and whose total area under the curve is 1. This we call a standardized normal distribution. Then, since the area below (to the left of) 0 is half the total area, the probability of obtaining a sample value less than 0 is certainly $\frac{1}{2}$. Similarly, the area under any other portion of the curve is exactly equal to the probability of obtaining a value in the corresponding interval of the horizontal axis.

Any normally distributed variable Y can be transformed into a standardized normal variable by the following transformation:

$$Z = \frac{Y - \mu}{\sigma} \tag{4-12}$$

This, in effect, shifts the mean from μ to 0 and changes the horizontal scale so that $\sigma = 1$. Now, we have this important relationship: *The probability that Y lies between a and b is equal to the probability that Z lies between* $(a - \mu)/\sigma$ *and* $(b - \mu)/\sigma$. We express this in symbols as follows:

$$P_r(a < Y < b) = P_r\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \tag{4-13}$$

Refer to Table I, Appendix, which shows the areas (probabilities) under the standardized normal curve. This table makes it possible for us to find approximate probabilities of obtaining various values of Z. The table is organized to show the area (probability) between the center of the standardized normal curve and a point a given number of standard units (Z's) away from the mean on either side. You will remember that the normal distribution is symmetric; hence only one-half of it needs to be tabled.

EXAMPLE *a*. Assume that we have a normal distribution, represented by the variable Y, whose mean is 100 and whose standard deviation is 10. What is the probability that, in drawing an observation Y_i at random from this population, its value will be between 90 and 100?

From Eq. (4-13) we have

$$\begin{aligned} P_r(90 < Y < 100) &= P_r\left(\frac{90 - 100}{10} < Z < \frac{100 - 100}{10}\right) \\ &= P_r(-1 < Z < 0) \end{aligned}$$

That is, we seek the probability that Z lies between -1 and 0, which is the same as the probability that Z lies between 0 and 1, because of the symmetry of the distribution. Reference to Table I shows that 0.3413 of the area lies between 0 and 1; hence this is the required answer. [Read the value from the table corresponding to $(Y - \mu)/\sigma = 1$.]

EXAMPLE *b*. Now, for a slightly more complex problem, suppose Y is normally distributed with mean 50 and standard deviation 5. We wish to know the probability of obtaining a value, by random sampling, between 43 and 54. Since 43 lies on one side of the mean of 50 and 54 on the other, it is convenient to break up the area into two parts as follows:

$$\begin{aligned} P_r(43 < Y < 54) &= P_r(43 < Y < 50) + P_r(50 \leq Y < 54) \\ &= P_r\left(\frac{43 - 50}{5} < Z < \frac{50 - 50}{5}\right) \\ &\qquad + P_r\left(\frac{50 - 50}{5} \leq Z < \frac{54 - 50}{5}\right) \\ &= P_r(-1.4 < Z < 0) + P_r(0 \leq Z < 0.8) \\ &= P_r(0 < Z < 1.4) + P_r(0 \leq Z < 0.8) \\ &= 0.4192 + 0.2881 = 0.7073 \end{aligned}$$

The last figures, of course, come from Table I. The areas are shown in Fig. 4-7.

EXAMPLE *c*. Suppose we assume the same data as in *b* except that now we require the probability that Y is between 43 and 45. Since all probabilities in Table I are cumulative from the mean out to some

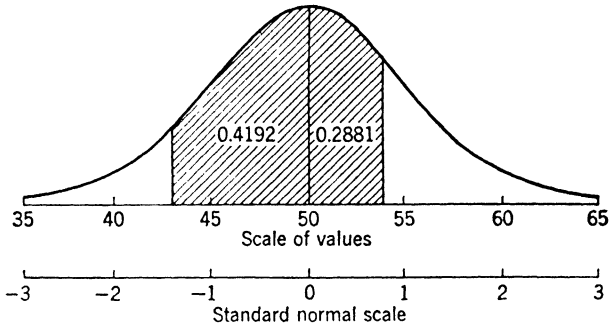

FIG. 4-7. Probability that Y lies between 43 and 54 when Y is normally distributed with mean 50 and variance 25

specified value of Z, we must find the probability that Y lies between 43 and 50 and subtract from it the probability that Y lies between 45 and 50.

$$P_r(43 < Y < 45) = P_r(43 < Y < 50) - P_r(45 < Y < 50)$$
$$= P_r\left(\frac{43 - 50}{5} < Z < \frac{50 - 50}{5}\right) - P_r\left(\frac{45 - 50}{5} < Y < \frac{50 - 50}{5}\right)$$
$$= P_r(0 < Z < 1.4) - P_r(0 < Z < 1)$$
$$= 0.4192 - 0.3413 = 0.0779$$

4-6. STANDARD ERROR

As indicated earlier, $\bar{Y}$ is an estimate of a population parameter μ, but repeated samples of a given size from the same population will yield varying values of $\bar{Y}$. We see, then, that a distribution of means of samples from a common population contains variation, just as does a distribution of individual observations. A further word is in order concerning the expression "a distribution of means." Let us suppose we have available a whole population of values, say, all the yearly earnings of the 10,000 employees of the MOM Beverage Bottling Company. (Don't ask for "pop"—ask for "MOM.") Now, suppose we draw a sample of 10 yearly earnings and compute the mean of the 10. We replace the 10 (put them back in the population, so that the population will not ultimately be exhausted), draw another 10, and compute the mean. We

repeat this process ad infinitum. We then make a frequency distribution of the resulting means. This is a sample distribution of means. If we have drawn enough samples, it will approach closely the distribution of the sample mean for samples of size 10. More will be said about this concept later.

It is important for us in statistics to establish some measure of the expected variance of the distribution of means. This can then be used to tell us a great deal about the reliability of the mean of the sample as an estimate of the parameter μ. One possible approach might be to take many random samples of the same size from a population, compute their means, and then compute the variance of these means as though they were any other observations. The difficulty of this approach is that taking a great many samples is too time-consuming, and often impossible. Through statistical theory a formula has been developed which gives us the variance of the mean of a given-sized sample. This formula is

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n} \tag{4-14}$$

where n is the size of the sample and σ^2 is the variance of an infinite population from which the sample was drawn. This formula will not be derived here, but we can see intuitively that, the larger the sample, the more stable will be the mean and that, the greater the variance in the population, the greater will be the expected variance in the mean of samples drawn from the population. We see, then, that the variance of the mean is directly proportional to the variance in the population and inversely proportional to the size of the sample.

As an arithmetic verification of Eq. (4-14), suppose we refer again to the 9 samples of size 2 from the finite population 1, 8, 9 in Sec. 4-3. The variance of the distribution of 9 sample means is computed in Table 4-11.

We recall that the population variance was 12.667 and the sample size $n = 2$; hence from Eq. (4-14) we have

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n} = \frac{12.667}{2} = 6.33$$

This result checks exactly with the variance of the sample means computed above.

This is a good place for you to examine your reading habits critically. Did you actually follow through the above illustration? If the answer is no, then you had better go back and try it again, because the idea presented is a fundamental one. Remember that we are demonstrating the validity of a formula for the variance of a *distribution of means*, that is, the variance of the mean. The first column in Table 4-11 represents the theoretical distribution of the mean, and its variance is 6.33. If one were

Table 4-11. Computation of Variance of All Possible Sample Means, from Finite Population 1, 8, 9

$\bar{Y}$	$\bar{Y} - \mu$	$(\bar{Y} - \mu)^2$
1.0	−5.0	25.00
4.5	−1.5	2.25
5.0	−1.0	1.00
4.5	−1.5	2.25
8.0	2.0	4.00
8.5	2.5	6.25
5.0	−1.0	1.00
8.5	2.5	6.25
9.0	3.0	9.00
54.0		57.00

$$\mu = \frac{\Sigma \bar{Y}}{N} = \frac{54}{9} = 6$$

$$\sigma_{\bar{Y}}^2 = \frac{\Sigma(\bar{Y} - \mu)^2}{N} = \frac{57}{9} = 6.33$$

to draw 9 samples of 2 each, the 9 means would not be likely to have the 9 values shown; but if one were to draw 90,000, approximately $\frac{1}{9}$ (or 10,000) would have the value 1.0, and so forth. The distribution of the sample mean would then look something like Fig. 4-8. The vertical bars

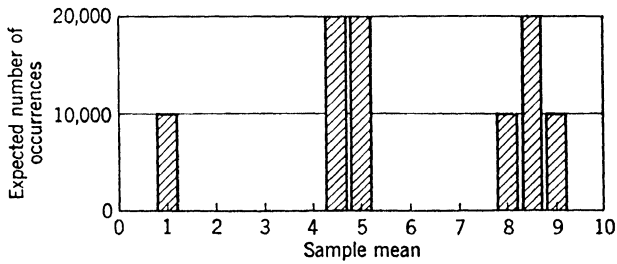

FIG. 4-8. Expected frequency of occurrence of sample means in 90,000 drawings (with replacement) of samples of size 2 from the population 1, 8, 9

represent the expected frequency of occurrence. There naturally would be some variation from these expected values due to sampling error.

The obvious difficulty with Eq. (4-14) is that seldom, if ever, is the variance of the population (σ^2) known. We therefore adapt the formula as follows to employ an estimate of the population variance:

$$s_{\bar{Y}}^2 = \frac{s^2}{n} \tag{4-15}$$

The square root of this expression might be called the *standard deviation of the mean,* but it is more commonly called the *standard error of the mean.* It is written

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} \tag{4-16}$$

EXERCISE 4-10. One hundred steel pins are measured for length and the observations recorded in centimeters. It is found that $\Sigma Y = 560$ and $\Sigma Y^2 = 3{,}167$. (*a*) What is your best estimate of the average length? (*b*) Assuming that the lot of pins from which the sample was taken is very large, what is your estimate of the standard error of the mean for samples of size 100? (*c*) What is your estimate of the standard error of the mean for samples of size 36?

Equation (4-16) assumes that random samples are drawn from an infinite population. When the population from which the sample was drawn is known to be finite, a finite-population correction should be applied to the standard error of the mean, as follows:

$$s_{\bar{Y}} = \sqrt{\frac{s^2}{n}\frac{N-n}{N}} \tag{4-17}$$

where N is the number of values in the population and n is the number of sample values. It is obvious that, when N is very large in relation to n, the finite-population correction has little effect and need not be applied. An arbitrary rule sometimes used is that, if the sample size is less than 5 per cent of the population size, the finite-population correction can be ignored.

Suppose we wish to compute the standard error of the mean of the sample of gross incomes of retail drugstores (Table 4-10). Let us assume that there are 400 in the population. We have

$$s_{\bar{Y}} = \sqrt{\frac{s^2}{n}\frac{N-n}{N}} = \sqrt{\frac{660.78}{50}\left(\frac{350}{400}\right)} = 3.40$$

The units are, of course, thousands of dollars, so that the standard error of the mean is $3,400.

EXERCISE 4-11. Refer to Exercise 4-10. Answer the same questions assuming that the total *lot* size is 500. Note that the sample size remains the same but that we are now sampling from a finite population.

Now for an important idea. Assume that we draw many samples of a given size n from some continuous population of data. (This implies, of course, that the population is infinite.) Each time a sample is drawn, the sample mean $\bar{Y}$ is computed. We then put these sample means into a class-interval distribution and plot the distribution as a column diagram.

Suppose we take many such samples and keep reducing the width of the class intervals as the total number of samples increases. The column diagram will gradually approach the shape of a smooth curve. We call this limiting curve the *distribution of the sample mean.* A constant sample size is implied, of course.

If a sample of size n is drawn from a normal population with mean μ and variance σ^2, the distribution of the mean is normal with mean μ and variance σ^2/n. That is, if the population from which we draw our samples is normal, the limiting distribution made up of the sample means themselves is normal. Furthermore, it has the same mean as the population itself and has variance σ^2/n.

Also, regardless of the shape of the original population (just so long as it has a finite mean and variance), the distribution of the sample mean approaches normality as n becomes large. This is a consequence of the *central limit theorem*, which is often called the most important theorem in statistics. Even for relatively small samples, the tendency toward normality is remarkable, as is demonstrated by one of the problems at the end of the chapter.

In any case the mean of the distribution of means is always μ. We say, then, that the expected value of $\bar{Y}$ is μ, which is another way of saying that $\bar{Y}$ is an unbiased estimate of μ. Also, the variance of the distribution of means is always σ^2/n. This is true whether or not the distribution is normal. However, it is more difficult to make use of the variance of the mean when the variable $\bar{Y}$ is not normally distributed.

Suppose we have a normal population with mean 100 and variance 25. What is the probability that the mean of a random sample of 64 items drawn from this population will have a value between 99 and 101? We know, from the above, that the variance of the mean is $\sigma^2/n = \frac{25}{64}$. The standard error of the mean is the square root of this quantity, or $\frac{5}{8} = 0.625$. We can approach the problem as follows:

$$
\begin{aligned}
P_r(99 < \bar{Y} < 101) &= P_r(99 < \bar{Y} < 100) + P_r(100 \leq \bar{Y} < 101) \\
&= P_r\left(\frac{99 - 100}{0.625} < Z < \frac{100 - 100}{0.625}\right) \\
&\quad + P_r\left(\frac{100 - 100}{0.625} \leq Z < \frac{101 - 100}{0.625}\right) \\
&= P_r(-1.6 < Z < 0) + P_r(0 \leq Z < 1.6) \\
&= 0.4452 + 0.4452 = 0.8904
\end{aligned}
$$

The probabilities are read from Table I in the Appendix.

Let us see how we can use these ideas to attack a hypothetical statistical problem. A processor of frozen foods plans to put 10 oz of frozen peas in each package, on the average. His statistician has reason to believe that

weights of the contents are normally distributed with a standard deviation of 0.5 oz. To find out whether the weights are up to standard, he weighs the contents of a random sample of 25 packages. He finds that $\bar{Y} = 9.5$. Is he justified in believing that his standard has dropped, or can this much variation be accounted for by random factors?

In this case the assumption of normality is not a critical one. The means of samples of size 25 are almost certain to be approximately normally distributed. We know, then, that if his packages meet the standard, the sample mean is normally distributed with mean $\mu = 10$ and variance $\sigma^2/n = (0.5)^2/25 = 0.01$. The standard error of the mean, which you will remember is the standard deviation of the distribution of means, is $\sigma/\sqrt{n} = \sqrt{0.01} = 0.1$.

We want to know the probability of drawing a sample value of 9.5 *or less* from a normal distribution whose mean is 10 and whose standard deviation is 0.1; hence we write

$$P_r(\bar{Y} \leq 9.5) = P_r\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{9.5 - 10}{0.1}\right) = P_r(Z \leq -5)$$

Referring to the normal-area table, we see that this probability is so small that it is not even tabled. We must conclude then that, since the probability of this event's happening by chance is so small, it is likely due to some other cause, namely, a real drop in average contents. In other words, the processor should have a look at his packaging process if he wishes to maintain his advertised standard.

EXERCISE 4-12. A machine produces steel pins whose length is approximately normally distributed with mean of 1 in. and standard deviation of 0.05 in. What is the probability that a pin chosen at random will have a length (*a*) greater than 1.05 in.? (*b*) less than 0.95 in.? (*c*) between 0.95 and 1.05 in.? (*d*) between 1.00 and 1.03 in.? (*e*) between 1.01 and 1.04 in.?

EXERCISE 4-13. If 100 steel pins are drawn at random from the population described in Exercise 4-12, how many of them would you expect to have values (*a*) less than 1.0 in.? (*b*) between 0.98 and 1.02 in.? (*c*) greater than 1.10 in.?

EXERCISE 4-14. If the *average* of the 100 steel pins is taken, what is the probability that the average will be (*a*) less than 0.99 in.? (*b*) between 0.98 and 1.02 in.? (*c*) between 1.01 and 1.02 in.?

EXERCISE 4-15. This exercise is designed to demonstrate the central limit theorem. First, draw 200 digits from the table of random numbers, Table 1-1. Prepare a frequency distribution of them, recording the number of 1s, 2s, 3s, and so on, and plot this distribution by a vertical-bar diagram. Next, group the digits by twos, finding the average of the first 2 digits, the average of the second 2, and so forth. Prepare a frequency distribution of these 100 averages and plot. Follow the same procedure again, grouping the first 4 observations, the second 4, and so forth, and finding new means. (One can achieve the same results by averaging the first 2 *means*, the second 2 means, and so on.) Repeat the process

again, using averages of 8 observations. Note how the distributions approach the symmetric bell-shaped form as the sample size increases. It will be helpful to use the same set of class intervals for each distribution. The following are suggested: −0.25 a.u. 0.25, 0.25 a.u. 0.75, 0.75 a.u. 1.25, 1.25 a.u. 1.75, 1.75 a.u. 2.25, and so on.

EXERCISE 4-16. Dressed chickens are being purchased for a chain restaurant. A random sample of 12 from a lot of 1,000 had the following individual weights, in ounces: 24, 26, 22, 25, 22, 23, 26, 24, 28, 25, 21, 24. (*a*) Estimate the average weight of the 1,000. (*b*) Estimate the *total* weight of the 1,000. (*c*) Compute the sample variance and standard deviation, using Eq. (4-7) and then Eq. (4-8). (*d*) Estimate the standard error of the mean. (*e*) Would you have been surprised to obtain the average weight you found in the sample of 12 if you had known the population of weights to be normally distributed with mean 23 and standard deviation 2?

EXERCISE 4-17. Answer the questions of Exercise 4-16 assuming that there were 100 chickens in the lot instead of 1,000. *Note:* Consider use of the finite-population correction.

EXERCISE 4-18. Make a frequency distribution of the digits in the first four lines of Table 1-1. Compute the mean and variance.

EXERCISE 4-19. In Exercise 4-18, what mean and variance would you expect, knowing that the digits were randomly determined? *Hint:* All of the digits from 0 to 9 are equally likely, so you can find the theoretical mean and variance by letting $f = 1$ for each digit in the formulas for mean and variance.

EXERCISE 4-20. (*a*) What is your estimate of the standard error of the mean in Exercise 4-18? (*b*) Do you think that the mean is nearly normally distributed? Why or why not?

EXERCISE 4-21. What is the probability distribution of Y, where Y is a random digit?

5. STATISTICAL INFERENCE II

5-1. INTRODUCTION

Some examples of statistical inference were given in Chap. 3. In that chapter the illustrations were limited to situations in which one of the discrete distributions of Chap. 2 could be used to assign probabilities. In this chapter we consider statistical inference under a wider set of conditions which includes many continuous probability distributions as well as the discrete distributions of Chap. 2.

Statistical inference may assume several forms. It may be that the objective is to estimate certain parameters of a population, such as total inventory, average age of accounts receivable, and proportion of customers who are women. We refer to this sort of estimation as *point estimation*, because we are interested in making a single estimate (a number) which can be represented by a point on the scale of all real numbers.

Generally speaking, point estimation is an incomplete solution to a problem. If we state, for example, that the average life of the electric bulbs we manufacture is estimated at 2,000 hr, the question arises, how accurate is this estimate? We found in Chap. 4 that the standard error of the mean could be helpful in gauging the precision of an estimate. That is, *if* the variable under consideration is normally distributed and *if* we know the standard error of the mean, then the probability that the difference between the sample mean and the population mean is less than 1 standard error is approximately equal to 0.68. This, we will remember, is the probability that an observation drawn at random from a normal population will lie within 1 standard deviation of the population mean. Therefore, in the above illustration, if our estimated mean is 2,000 hr and if the true standard error of the mean is 50 hr, we should be surprised if the real population average, μ, were less than 1,850 hr or more than 2,150 hr. These are the figures which are 3 standard errors away from

the sample mean. If the standard error of the mean has also been estimated, then some adjustment in the probability statements is necessary. This is discussed later.

It is important to note that what we have said above is *not* equivalent to saying that the probability is 0.68 that μ lies between $\bar{Y} - \sigma_{\bar{Y}}$ and $\bar{Y} + \sigma_{\bar{Y}}$. This last statement is meaningless, because μ is a parameter. Either it lies in the interval or it does not, so the probability is either 0 or 1, and not 0.68 as stated.

How, then, can we make a probability statement indicating the precision of an estimate? Confidence intervals give us an answer to this problem. We select a value for α, where α is the probability of Type I error, discussed in Chap. 3. Then $1 - \alpha$ is called the *confidence coefficient.* Next, let us assume that the sample mean (for the given sample size in which we are interested) is normally distributed with mean μ and variance σ^2/n (or standard error $\sigma/\sqrt{n}$). Then for any given μ and $\sigma/\sqrt{n}$ we can determine two values, A and B, such that $\alpha/2$ of the area of the normal curve lies above (to the right of) A and $\alpha/2$ of the area lies below (to the left of) B. Thus $1 - \alpha$ of the area lies between A and B.

For example, suppose $\mu = 50$, $\sigma = 5$, $n = 16$, and $\alpha = 0.10$. Then $\sigma/\sqrt{n} = 1.25$, and the two-tailed normal deviate for $\alpha = 0.10$ is 1.645. Hence

$$A = 50 + 1.645(1.25) = 52.056$$
$$B = 50 - 1.645(1.25) = 47.944$$

and 90 per cent of the area of the normal curve lies between A and B. That is, the probability is 0.90 that a mean of a sample of 16 observations drawn from this population will lie between 47.944 and 52.056.

Now, suppose that $\mu = 40$ rather than 50, as assumed above. In this case $A = 42.056$ and $B = 37.944$. Likewise, if $\mu = 60$, we have $A = 62.056$ and $B = 57.944$. We see then that, given any value for μ (and the constant assumptions about σ, n, and α), we can find values for A and B. Further, the probability is 0.90 that the *sample mean* will lie between A and B for any given assumption about μ. These observations permit us to construct the lines A and B in Fig. 5-1. The horizontal axis represents the possible values of μ from 10 to 60. The vertical axis represents the values of $\bar{Y}$ that can be obtained for a given value of μ. For a given value of μ, then, 90 per cent of the distribution of means will lie between the values represented by the lines A and B.

For the case we are considering, however, we do not know the value of μ. We know only the value $\bar{Y}$ which has been computed from a sample of n values. Suppose that the sample mean has a value of 32. This mean is plotted as a horizontal dotted line in Fig. 5-1. It crosses the A line at about 30 and the B line at about 34 on the scale of μ. That

is, the sample mean of 32 is at the upper 5 per cent point if $\mu = 30$ and at the lower 5 per cent point if $\mu = 34$. Thus we say that the 90 per cent confidence limits are 30 and 34.

This is *not* the same thing as saying that the probability is 0.90 that μ lies between 30 and 34. What it does mean is that, if we place confidence limits on means by this procedure, in a long sequence of trials the intervals, so computed, will contain the parameter μ about 90 per cent of the time, on the average.

We have investigated a graphical solution for the confidence interval when $\bar{Y} = 32$. Let us consider an algebraic solution. The lower

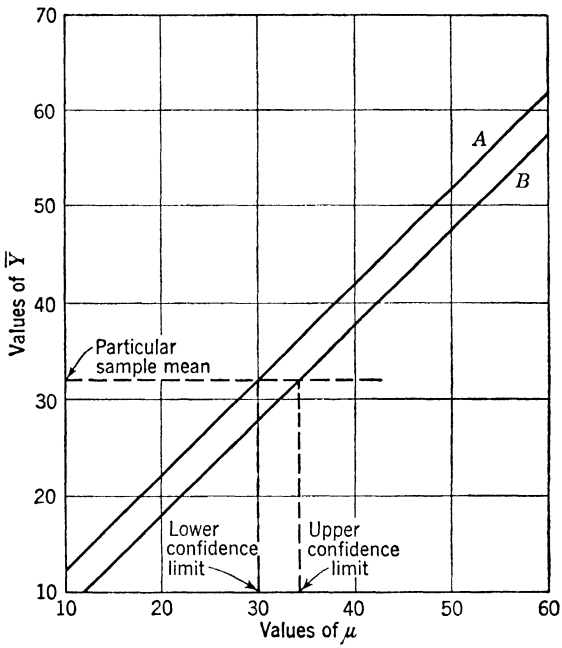

FIG. 5-1. Confidence intervals for the mean

confidence limit, C_1, is the value for μ such that $\mu + 1.645\,\sigma/\sqrt{n} = 32$. Solving this for μ, we have $\mu = 32 - 1.645(1.25) = 29.94$, which is more precise than our graphical solution of 30. The upper confidence limit, C_2, is the value for μ such that $\mu - 1.645\,\sigma/\sqrt{n} = 32$. Solving for μ yields $C_2 = 34.06$, which is again more precise than the corresponding graphical solution, 34.

In order to have a convenient expression for what we have computed, we can write

$$P_c(\bar{Y} - Z_\alpha \sigma_{\bar{Y}} < \mu < \bar{Y} + Z_\alpha \sigma_{\bar{Y}}) = 1 - \alpha \qquad (5\text{-}1)$$

where Z_α is the normal deviate which excludes $\alpha/2$ of the area on the

positive side of the curve and $\sigma_{\bar{Y}}$ is the true standard error of the mean. It must be remembered that this is not a probability statement in the usual sense. Probability can be associated with it only by considering many repetitions of the confidence-interval procedure. With this concept, about $1 - \alpha$ of such confidence intervals will contain the population parameter.

EXERCISE 5-1. Place 95 per cent confidence limits on the mean of the light-bulb illustration, where $\bar{Y} = 2{,}000$ and $\sigma_{\bar{Y}} = 50$.

Since we shall apply the confidence-interval technique to various parameters, it is desirable to have a general expression for a confidence interval. We may write

$$P_c(C_1 < \theta < C_2) = 1 - \alpha \tag{5-2}$$

where C_1 and C_2 are the lower and upper limits, respectively, and θ is the parameter being considered.

The confidence interval, as exemplified by Eq. (5-2), is an example of the risks inherent in a management decision. The manager does not know whether the confidence interval actually includes the parameter or not, but he governs his actions as though it does. In the long run his risk of wrong decision is α. He can control α arbitrarily, but if he makes α very small, his confidence interval may become so wide as to be meaningless. In the light-bulb illustration, for example, he may be sure that the average life of his bulbs is somewhere between 1,000 and 3,000 hr, but this information is not of much use to him if he is trying to make a comparison with a competitor's published average life of 2,200 hr.

We have discussed the point estimate and the interval estimate as examples of statistical inference. A third form, which has already received some attention in Chap. 3, is the test of hypothesis. Sometimes it is more convenient to consider a decision problem within the framework of hypothesis testing than within the framework of confidence intervals. We shall formalize the procedure for testing a hypothesis by considering seven steps.

1. State the hypothesis and the alternative. Sometimes the hypothesis can be expressed in terms of the parameters of the population. For example, $\mu = 50$, $\sigma_1^2 = \sigma_2^2$ are hypotheses about parameters in the population. At other times the hypothesis is simply a statement, in words, about the population. For example, "X is normally distributed" and "X and Y are independent" are hypotheses of this sort. The alternative hypothesis indicates the hypothesis which must be accepted if the stated hypothesis is rejected. Sometimes there is a single alternative to be considered. Sometimes there are infinitely many alternatives ($\mu > 50$, $\sigma_1^2 \neq \sigma_2^2$, etc.).

2. Choose a level of significance. This step involves choosing a level for α, that is, the probability that a true hypothesis will be rejected. Some common levels chosen are 0.10, 0.05, 0.01, 0.001. Remember that if one makes α very small, say, 0.001, he runs a great risk (β) of accepting a false hypothesis. This risk can only be computed for specific alternatives and is not known in general. The term *level of significance* arises because the test of hypothesis judges the "significance" of the difference between the hypothesis and the sample result. If $\alpha = 0.05$ and the hypothesis is rejected, we may say that the results are significantly different at the 0.05 level. This implies, by the usual methods of selecting the region of rejection, that *if the hypothesis were true so unusual a sample could be expected less than 5 per cent of the time.* This last statement is quite important and is worth digesting.

In view of the relationship of α and β, it seems wise to devote some careful thought to the selection of α (which then determines β for particular alternatives). We shall use the normal distribution and a hypothesis about the mean to illustrate the point.

Suppose we are producing small cardboard containers whose quality is judged by the number of pounds of pressure it takes to break them. Our present process produces containers with an average crushing strength of 27 lb and a standard deviation of 4 lb. We are considering a new forming process which is slightly more expensive. We have reason to believe that the new process will produce a product which is just as variable as the old (that is, we still expect a standard deviation of 4 lb), but we do not know what mean to expect. Our situation calls for testing some of the new product for crushing strength and examining these results in the light of the hypothesis that $\mu = 27$ lb.

It must be obvious that we shall not change our process if the new process produces a product which is inferior to the old. (Remember that it costs more to produce the new product.) We want, then, to reject the hypothesis that $\mu = 27$ if our sample mean is somewhat greater than 27 lb. But how much greater? Suppose we decide that it isn't worthwhile changing processes unless the *true* average of the new process is at least 28 lb. This figure then represents an alternative hypothesis, in the language of Chap. 3.

Suppose we examine a sample of 25 containers. There is reason (by virtue of the central limit theorem) to believe that the mean is normally distributed. Under the hypothesis that $\mu = 27$, we expect the sample mean to be normally distributed with mean 27 and standard error equal to $4/\sqrt{25} = 0.8$.

Now, suppose we decide to reject the hypothesis that $\mu = 27$ if the sample mean is so large that it could have occurred by chance less than 5 per cent of the time. That is, we shall run a risk of 0.05 of rejecting

a true hypothesis. This is α. Under this decision rule, what is the probability (β) that we shall accept the hypothesis $\mu = 27$ when the alternative hypothesis $\mu = 28$ is actually true?

Reference to Appendix Table I shows that the standardized normal deviate which has 5 per cent of the area to the right of it is 1.64. Put in terms of our illustration, this means that we shall reject the hypothesis that $\mu = 27$ when $\bar{Y}$ is equal to or greater than 27 plus 1.64 standard errors, or $27 + 1.64(0.8) = 28.31$. Conversely, we shall accept the hypothesis that $\mu = 27$ whenever $\bar{Y}$ is less than 28.31. Now, if our alternative hypothesis is true, what is the probability that we shall get a

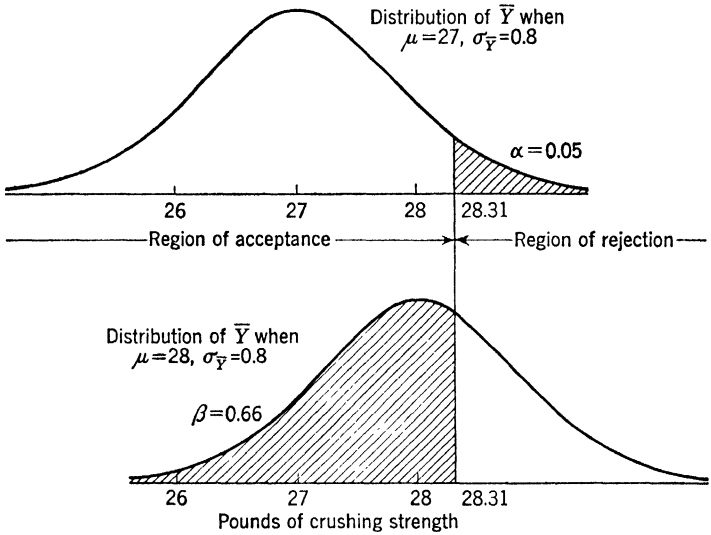

FIG. 5-2. α and β errors for a particular decision rule

sample mean less than 28.31 (i.e., what is the probability that we shall accept the wrong hypothesis)? Using the notation of Chap. 4, we set up the problem as follows:

$$P(\bar{Y} < 28.31) = P\left(\frac{\bar{Y} - 28}{0.8} < \frac{28.31 - 28}{0.8}\right)$$
$$= P(Z < 0.4) = P(Z < 0) + P(0 \leq Z < 0.4)$$
$$= 0.500 + 0.16 = 0.66 \qquad \text{(approximately)}$$

Therefore $\beta = 0.66$ and we are running a risk of 0.66 of not recognizing a better process. We may consider this risk excessive and decide to reduce it in order to make risks α and β more nearly equal. Figure 5-2 shows the quantities α and β for the decision rule to reject when $\bar{Y}$ is equal to or greater than 28.31.

EXERCISE 5-2. (*a*) Let $\alpha = 0.10$ and compute β. (*b*) Do you think this is a better decision rule? Why? (*c*) Draw a diagram similar to Fig. 5-2.

3. Select a probability distribution. This step will become more meaningful as this chapter progresses. Since results must be interpreted in terms of probabilities, it is necessary to find a probability distribution which fits our needs. In the container illustration above, we used the normal distribution. In Chap. 3 we used the binomial distribution. Other distributions commonly used are the t distribution, the chi-square (χ^2) distribution, and the F distribution. These are all described in this chapter.

4. Choose a region of rejection. We divide the entire space of possible answers for our test results into two regions: a region of rejection and a region of acceptance. It might be more descriptive to call the latter region the "region of nonrejection," since sometimes we may simply withhold judgment when the sample results lie in this region.

Mathematical criteria have been developed for determining a "best region of rejection," but for the cases we meet in this beginning book the region can be determined quite easily. We simply include in the region of rejection those portions of the probability distribution which are most likely in view of the alternative hypothesis being tested. The illustrations later in this chapter will clarify this point.

5. Do the computations.

6. Make the statistical decision. This decision is either to reject or not to reject the hypothesis, depending upon whether the computed value of the test criterion in step 5 lies in the region of rejection or the region of nonrejection.

7. Make the managerial decision. This decision lies outside the realm of statistics, but it is included here to emphasize that the purpose of the statistical decision is to contribute information for the benefit of management.

5-2. INFERENCES ABOUT MEANS

We consider first the situation in which we wish to test a hypothesis about a population mean. That is, as a result of examining sample data we shall make a decision whether to accept or to reject a given hypothesis about the population mean.

We shall consider again the frozen-peas illustration of Chap. 4. Remember that we wished to put 10 oz of peas in each package. The problem is considered in the light of the seven steps suggested above.

1. *Hypothesis:* $\mu = 10$ oz. *Alternatives:* $\mu < 10$ or $\mu > 10$. (Note that the alternatives permit any departure at all from 10 oz. Either a

very high value or a very low value of $\bar{Y}$ will tend to discredit the hypothesis that $\mu = 10$.)

2. $\alpha = 0.05$. This is an arbitrary decision on our part. Whether it is good or bad can be tested only by knowing the cost of making wrong decisions.
3. The probability distribution is normal with $Z = (\bar{Y} - \mu)/\sigma_{\bar{Y}}$, where $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$. This distribution is used because the sample mean is likely to be nearly normally distributed and its variance is known to be σ^2/n.
4. As noted above, the alternative hypothesis is stated in such a manner that either a very high value or a very low value for $\bar{Y}$ would tend to discredit the stated hypothesis and lead to acceptance of the alternative. That is, we can reject the hypothesis if we are putting too much in the package or too little. If $\bar{Y}$ is very high, Z will be positive.

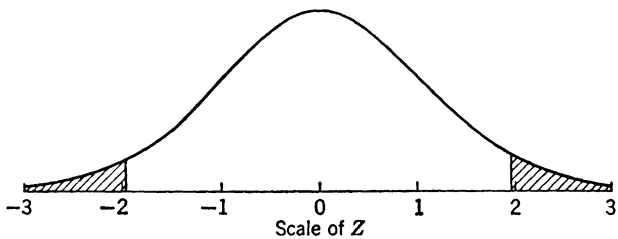

FIG. 5-3. Normal curve with regions of rejection for the two-tailed test for $\alpha = 0.05$

If $\bar{Y}$ is very low, Z will be negative. Therefore, our region of rejection will consist of the two ends of the normal distribution. Since $\alpha = 0.05$, we must have half of this, or 0.025 of the area, in each half of the region of rejection. The area from the mean out to the region of rejection will be $0.500 - 0.025$, or 0.475. We look in the *body* of the normal-curve-area table (Table I, Appendix) for 0.475 and find that the Z value corresponding to this area is 1.96. Therefore, the region of rejection can be specified by $Z \leq -1.96$ or $Z \geq 1.96$. A diagram of the normal distribution with the region of rejection is given in Fig. 5-3. Since both ends of the curve are included in the region of rejection, we call this test of sigificance a *two-tailed test.*

5. The computations are

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{9.5 - 10}{0.5/\sqrt{25}} = -5$$

6. Since Z is clearly less than -1.96, it lies in the region of rejection. Therefore we reject the hypothesis that our standard has not changed. We accept the alternative hypothesis that $\mu < 10$, that is, that we

are actually putting less than 10 oz of product in each package, on the average.

7. The managerial decision is to examine the process, to see whether the average can be brought back to standard.

EXERCISE 5-3. What is the probability that, if our average has actually dropped to 9.8 oz, we shall fail to recognize the change by the decision rule of step 4 above?

Suppose we are not interested in how much the package of frozen peas exceeds 10 oz but are vitally concerned with any weight deficiencies. Modifications are required in steps 1 and 4, and possibly in 5 and 6.

1. *Hypothesis:* $\mu = 10$ oz. *Alternative:* $\mu < 10$.
2. $\alpha = 0.05$ (as before).
3. $Z = (\bar{Y} - \mu)/\sigma_{\bar{Y}}$ (as before).
4. Only small values of $\bar{Y}$ will tend to make us reject the hypothesis and to accept the alternative; hence the entire region of rejection must be at the small end (negative end) of the normal distribution. The point which has 5 per cent of the area to the left of it is -1.645. (Verify this from Table I, Appendix.) The one-tailed region of rejection is shown in Fig. 5-4.

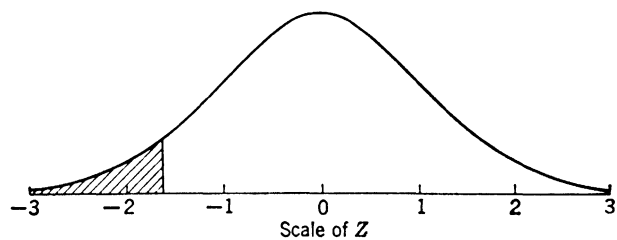

FIG. 5-4. Normal curve with region of rejection for the one-tailed test for $\alpha = 0.05$

5. $Z = -5$ (as before).
6. Reject the hypothesis, since $Z = -5$ is in the region of rejection; that is, it is less than -1.645.
7. Same as before.

The alternative hypothesis calls for a *one-tailed test.* In other words, the entire region of rejection is at one end of the probability distribution (in this case, the normal curve). Generally speaking, when the alternatives permit departure from the hypothesis in only one direction, a one-tailed test is called for. If departure is permitted in both directions, a two-tailed test is required.

EXERCISE 5-4. Recompute Exercise 5-3 using the one-tailed test and the region of rejection specified by step 4, above.

In the above illustration it appears that our population average has changed and is no longer 10 oz. Our best estimate of the new average is 9.5 oz. We may place a confidence interval around this estimate as an indication of its reliability. Using formula (5-1) with $\alpha = 0.10$, we have

$$9.5 - 1.64(0.1) < \mu < 9.5 + 1.64(0.1)$$
$$9.34 < \mu < 9.66$$

In this case the selection of α is purely arbitrary. We can assert with some confidence that the true mean is somewhere between 9.34 and 9.66 oz. In a long sequence of such statements, we should be correct 90 per cent of the time, or 9 times out of 10.

The obvious weakness in the above examples is that the variance (or standard deviation) of the population must be known in order to compute the standard error of the mean, $\sigma_{\bar{Y}}$. In the usual case we shall not have any more information about σ than we have about μ, and we must estimate both. In this case we use the t distribution as our probability distribution rather than the normal.

The t probability distribution takes the following form for testing hypotheses about the mean:

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \tag{5-3}$$

The only difference between this form and that of the normal is that σ is replaced by s. Note that the numerator of the t ratio is normally distributed (under the assumption that Y itself is normally distributed) with mean 0 and that the square of the denominator is an unbiased estimate of the variance of the numerator. The ratio is distributed as t *if the variable* Y *is normally distributed.* Even if Y is not normally distributed, the ratio tends to be distributed as t with increasing sample size, because of the central limit theorem mentioned earlier. A precise definition of the t distribution is given in Sec. 5-4.

The distribution of t is dependent upon the number of degrees of freedom in the computation of the variance s^2. There is a different t distribution for every number of degrees of freedom. When the degrees of freedom become large, the t distribution approaches the normal distribution, but for small degrees of freedom there is substantial difference between the two, as illustrated by Fig. 5-5.

Table II, Appendix, gives various values of t for selected probability levels and various degrees of freedom. The probability levels are the probabilities that the given values of t will be exceeded *in absolute amount.* That is, for 16 degrees of freedom the probability that t is less than -2.12 or is greater than 2.12 is 0.05. This can be written more

compactly as

$$P_r(|t| > 2.12) = 0.05$$

In words, the probability that the absolute value of t is greater than 2.12 is 0.05. We see, then, that the t table is set up for two-tailed tests.

Note also, with respect to Table II, that with infinite degrees of freedom the probability that t is less than -1.96 or greater than 1.96 is 0.05, which is exactly the value we found earlier for the normal distribution. This is a consequence of the fact that the t distribution with infinite degrees of freedom is the normal distribution.

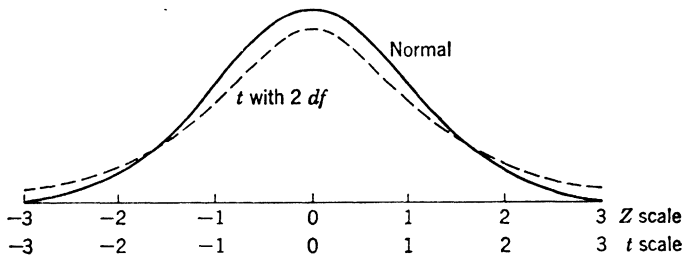

FIG. 5-5. Comparison of normal and t distributions

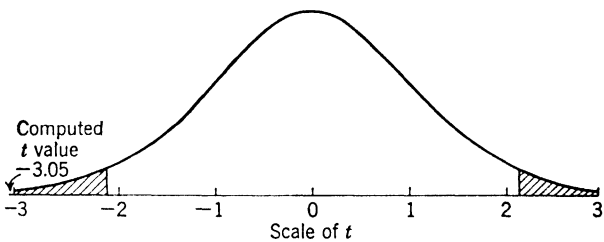

FIG. 5-6. Distribution of t with 15 degrees of freedom showing region of rejection for two-tailed test and $\alpha = 0.05$

Suppose our company is manufacturing steel wire with an average tensile strength of 100 lb. Our laboratory examines 16 pieces and finds that $\bar{Y} = 95.8$, $s^2 = 30.25$. Are these results in accordance with the hypothesis that the population average is 100 lb?

1. *Hypothesis* (H_0): $\mu = 100$. *Alternatives:* $\mu > 100$ or $\mu < 100$.
2. $\alpha = 0.05$.
3. *Test criterion:* $t = (\bar{Y} - \mu)/(s/\sqrt{n})$ with $(n - 1) = 15$ *df*.
4. *Region of rejection:* $t \leq -2.13$ or $t \geq 2.13$.
5. $t = (95.8 - 100)/(5.5/4) = -4.2(4)/5.5 = -3.05$.
6. Reject H_0 since the computed t lies in the region of rejection. The region of rejection and the computed t value are shown in Fig. 5-6.
7. Examine the production process to find out why.

Note that the above hypothesis calls for a two-tailed test. It is quite likely that we should not be concerned if our tensile strength exceeded

the standard, so we might wish to state the alternative hypothesis as $\mu < 100$. This statement of the hypothesis calls for a one-tailed test and requires a revision of the region of rejection. Since the t table is set up for a two-tailed test, *we must double the level of α to find the t value for a one-tailed test.* Since only small values of $\bar{Y}$ will tend to discredit the hypothesis, only the negative end of the t distribution will lie in the region of rejection. The region of rejection for the one-tailed test is $t \leq -1.75$.

It may be desirable to examine this argument somewhat more closely. We are seeking the value of t that has 0.05 of the area below it. Since the t table is set up for two-tailed tests, we must find the value of t such that its absolute value will be exceeded 10 per cent of the time on the average. That is, 5 per cent of the values will be too small and 5 per cent too large. This justifies the rule of doubling α in order to make the one-tailed test.

Again, we can make a confidence-interval estimate of the mean. Since we do not know the true value of σ, we must use the t distribution instead of the normal, modifying Eq. (5-1) accordingly. The revised formula with its evaluation for a 95 per cent confidence interval is as follows:

$$P_c[\bar{Y} - t_{(\alpha,n-1)}s_{\bar{Y}} < \mu < \bar{Y} + t_{(\alpha,n-1)}s_{\bar{Y}}] = 1 - \alpha \qquad (5\text{-}4)$$

$$95.8 - 2.13(1.375) < \mu < 95.8 + 2.13(1.375)$$

$$92.87 < \mu < 98.73$$

Note that placing a $1 - \alpha$ confidence interval around the mean is equivalent to making a two-tailed test with probability α. If μ lies within the confidence interval, we cannot reject the hypothesis. If μ lies outside the confidence interval, we do reject the hypothesis.

Exercise 5-5. Over a long period of time a certain food-processing plant has prepared a product with an average water content of 37.5 per cent and a standard deviation of 3.4 per cent. A sample of 10 units, taken after a change in cooking procedure, shows an average water content of 39.2 per cent. (*a*) Is this significantly different? (*b*) How would your procedure differ if you had not known the standard deviation of 3.4 per cent but had estimated it as 3.4 from the sample?

Exercise 5-6. Place 90 per cent confidence limits on the mean in Exercise 5-5, (*a*) under the assumption that the standard deviation is known and (*b*) under the assumption that it was estimated from the sample.

An interesting problem arises in the evaluation of the probability of Type II error (β) when the standard deviation of the population is unknown. The schema of Fig. 5-7 illustrates the situation.

For the evaluation of α the quantity $(\bar{Y} - \mu_0)/(s/\sqrt{n})$ is distributed as t with $n - 1$ degrees of freedom. But for the evaluation of β the quantity $[(\bar{Y} - \mu_1) + (\mu_1 - \mu_0)]/(s/\sqrt{n})$ is distributed as *noncentral t*

with $n - 1$ degrees of freedom. It is difficult to work with because it contains an extra parameter, $\sqrt{n}\,(\mu_1 - \mu_0)/\sigma$. This is called the *noncentrality parameter.* Some exact tables* have been prepared for the evaluation of β, but they will not be discussed here. It may be pointed out that if the sample size is reasonably large, say, 30 or more, then one can obtain a reasonably close approximation to β by using the normal distribution (Z), as was done in the previous section. One simply uses the estimate s in place of σ in the distribution of Z. Even the exact tables referred to above require one to express the difference between the hypothesis and its alternative in terms of the true standard deviation,

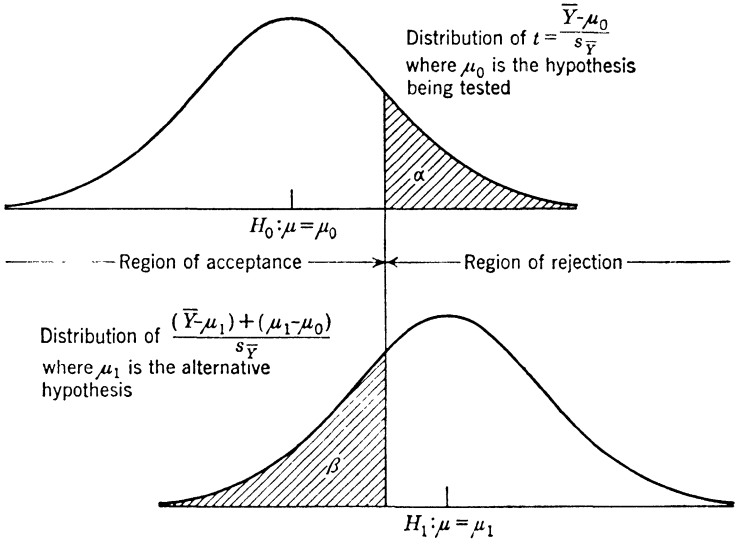

FIG. 5-7. α and β errors with the t distribution

which usually is not known. Even an approximate answer can be quite helpful to management, however.

EXERCISE 5-7. A standard production process produces gadgets with an average waste of 14 per cent of the weight of the new material in the finished product. A new process is proposed which is supposed to cut waste. Assume that the raw material is worth 10 cents per pound and that there are 5 lb of it in the finished product. Assume also that the new process costs 5 cents more per unit than the old. (*a*) What is a logical hypothesis to test? Suppose a sample of 25 shows an average waste of 3 per cent, with a standard deviation of 2 per cent. (*b*) Would

* Jerzy Neyman, with the cooperation of K. Iwaszkiewicz and St. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation," *Suppl. J. Roy. Statist. Soc.*, vol. 2, no. 2, pp. 107–180, 1935. The exact methods require that we know the standard deviation precisely. But if we know the exact value of σ, we use the normal distribution rather than the t distribution. Hence, some approximation must be introduced in any method of estimating β.

you keep the old process or introduce the new on the basis of this evidence? (*c*) What is a logical alternative? (*d*) What is your decision rule? (*e*) What is α? (*f*) What is β (approximately) for the alternative you have chosen?

When one is designing a test of a hypothesis, the appropriate alternative may not suggest itself until after the trials have been completed. Suppose in a chain grocery store one experiments with the display of barbecued chickens to see whether average sales per customer can be increased. He may establish the hypothesis that there will be no change, without having in mind a particular alternative. After the experiment has been conducted, he may reject the stated hypothesis with a risk of Type I error of α. Then, his best estimate of the correct hypothesis (the parameter he is estimating) is the average he has found as a result of experimentation. It is this revised estimate of the population mean which requires a statement concerning its reliability, either in the form of a listing of its standard error or in terms of its confidence interval.

5-3. INFERENCES ABOUT THE VARIANCE

Sometimes we are as concerned about variation in the population as we are about the average. Take the wire illustration. If our wire is to be used in a tying process, its diameter and tensile strength must be controlled carefully. If the wire thins out, it may break under the tension necessary in tying. If it thickens, it may not pass through the tying mechanism.

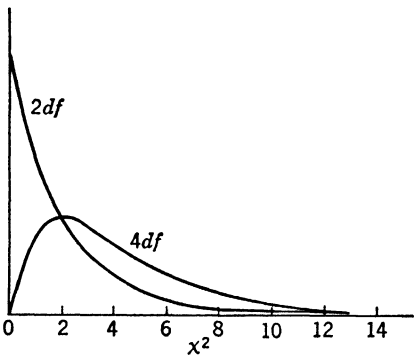

FIG. 5-8. Chi-square distributions with 2 and 4 degrees of freedom

The distribution we use in testing the variance is the chi-square (χ^2) distribution. If Y is normally distributed, then the ratio

$$\frac{(n-1)s^2}{\sigma^2} \tag{5-5}$$

is distributed as χ^2 with $n-1$ degrees of freedom. Like the t distribution, there is a different χ^2 distribution for each number of degrees of freedom. Two χ^2 distributions are shown in Fig. 5-8. The distribution of χ^2 is highly skewed to the right for small degrees of freedom and becomes approximately normal for large degrees of freedom, say, more than 30. Interesting properties of the χ^2 distribution are that its mean is always equal to the number of degrees of freedom and that its variance is always equal to twice the degrees of freedom. These properties make

it easy to find approximate χ^2 values for large numbers of degrees of freedom by using the normal-area table.

Let us refer again to the wire illustration in which, with $n = 16$, we found $\bar{Y} = 95.8$ and $s^2 = 30.25$. Suppose we wish to test the hypothesis that the population variance is 22, which we assume is a suitable standard of uniformity, against the alternative that $\sigma^2 > 22$. Suppose $\alpha = 0.10$ is satisfactory for our risk of rejecting a true hypothesis.

1. $H_0: \sigma^2 = 22$. $H_1: \sigma^2 > 22$.
2. $\alpha = 0.10$.
3. $\chi^2 = (n - 1)s^2/\sigma^2$ with $n - 1$ *df*.
4. The region of rejection is found by reference to Table III, Appendix, with $n - 1 = 15$ degrees of freedom. Note that the alternative hypothesis calls for a one-tailed test. Only large values of s^2 will tend to discredit the hypothesis. Large values of s^2 will cause large values of χ^2. Hence we are required to put the upper tail of the χ^2 distribution in the region of rejection. The χ^2 table is set up to show the probability that a particular value of χ^2 will be *exceeded*. For $\alpha = 0.10$ we want the χ^2 value which has 10 per cent of the area above it. This value, with 15 degrees of freedom, is 22.31. Therefore, our region of rejection is $\chi^2 \geq 22.31$.
5. $\chi^2 = 15(30.25)/22 = 20.63$.
6. We do not reject the hypothesis that our standard has not changed, since our computed χ^2 does not fall in the region of rejection.
7. It is interesting to note the effect of choice of α on the managerial decision. With $\alpha = 0.10$ we should expect, on the average, to examine the production process needlessly about 10 per cent of the time. If such examination is quite costly we should prefer to make α smaller, say, 0.02 or 0.01. Also the size of sample is directly related to the sensitivity with which one can detect a change in standard. Suppose that a sample of 31, rather than 16, had resulted in $s^2 = 30.25$. Then the region of rejection would have been $\chi^2 \geq 40.26$, and the computed χ^2 would have been 41.25. Hence the same difference between hypothetical and actual variance would have caused us to examine the production process. Here the cost of testing must be balanced against the risk of producing an inferior product.

For a two-tailed test of the hypothesis $\sigma^2 = 22$ with 15 degrees of freedom and $\alpha = 0.10$, we divide the region of rejection into two equal parts. We find the χ^2 value which has 95 per cent of the area to the right of it (7.26) and the value that has 5 per cent of the area to the right of it (25.00). Note that 90 per cent of the area then lies between 7.26 and 25.00. Then the region of rejection is $\chi^2 \leq 7.26$ or $\chi^2 \geq 25.00$. The one-tailed and two-tailed regions of rejection are plotted in Fig. 5-9.

Occasionally one may have more than one sample from the same

population and may wish to combine the separate variance estimates into a single estimate. We may construct a pooled estimate of the variance by computing a weighted average of the separate variances, using degrees of freedom as weights. That is, suppose we have three

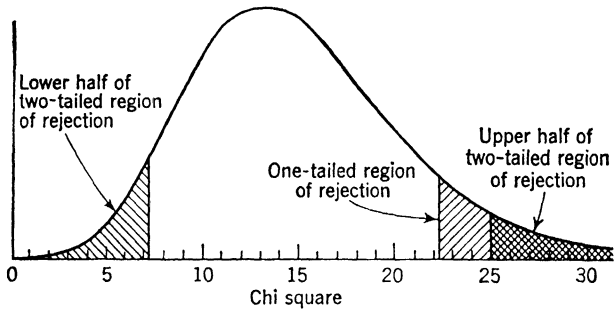

FIG. 5-9. Chi-square distribution with 15 degrees of freedom showing one- and two-tailed regions of rejection for $\alpha = 0.10$

estimates of the variance denoted by s_1^2, s_2^2, and s_3^2. A pooled estimate of the variance is formed from

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n_1 + n_2 + n_3 - 3} \tag{5-6}$$

If the population from which the three samples were drawn is normally distributed, then the ratio

$$\frac{(n_1 + n_2 + n_3 - 3)s_p^2}{\sigma^2}$$

is distributed as χ^2 with $n_1 + n_2 + n_3 - 3$ degrees of freedom. Hypotheses can be tested using the combined variance in much the same manner as explained above for a single estimate.

Ordinarily, a combined estimate of the mean is more useful than a combined estimate of the variance. If one has two or more random samples from the same population and wishes to combine them into one estimate of the mean, he may do so by a weighted average, using sample sizes as weights. That is, for three samples one would compute

$$\bar{Y} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2 + n_3\bar{Y}_3}{n_1 + n_2 + n_3} \tag{5-7}$$

For averaging means, one uses sample sizes as weights, and for averaging variances, one uses degrees of freedom as weights.

A different problem arises when one has two or more estimates of the population mean, each having a different variance. In this case one should average the estimates using reciprocals of the variances as weights.

The problem sometimes arises in experimental work and in survey sampling.

Sometimes experiments or sample surveys are conducted with the specific purpose of estimating the variance in the population. More often, the estimated variance is obtained as a result of an estimate of the mean or some other parameter. In either case, a confidence-interval estimate of the variance may be required. Here, again, we seek an interval such that the "confidence probability" is $1 - \alpha$ that the interval C_1 to C_2 will contain the population variance σ^2. In the case of the variance, however, the limits C_1 and C_2 are not symmetrically distributed about the estimator s^2. We recall that the ratio $(n - 1)s^2/\sigma^2$ is distributed as χ^2 with $n - 1$ degrees of freedom. We make use of this information to estimate σ^2. We solve

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

for σ^2 to obtain

$$\sigma^2 = \frac{(n - 1)s^2}{\chi^2} \tag{5-8}$$

By inserting the values of $(n - 1)$, s^2, and χ^2, we can estimate two values of σ^2, which we shall denote by C_1 and C_2. The values of χ^2 chosen depend upon the level of confidence and the given degrees of freedom. If we wish a 0.90 confidence interval we choose the χ^2 values for the 0.05 and 0.95 levels of probability, remembering that the χ^2 table gives the probability that a given value of χ^2 will be exceeded. The confidence interval may be written as follows:

$$P_c\left[\frac{(n - 1)s^2}{\chi^2_{(\alpha/2,n-1)}} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}\right] = 1 - \alpha \tag{5-9}$$

The subscripts on χ^2 indicate the probability levels in the χ^2 table and the degrees of freedom.

Suppose that for a sample of 6 we find $s^2 = 20$ and we wish a 0.98 confidence-interval estimate. We refer to a χ^2 table with 5 degrees of freedom and observe that a χ^2 of 0.554 is exceeded 99 per cent of the time and a χ^2 of 15.086 is exceeded 1 per cent of the time. We compute

$$C_1 = \frac{5(20)}{15.806} = 6.63$$

$$C_2 = \frac{5(20)}{0.554} = 180.5$$

Therefore, the 98 per cent confidence interval is

$$6.63 < \sigma^2 < 180.5$$

The reader may be impressed by the width of this interval. However, the sample size is small, and these are limits on the variance rather than the standard deviation. If square roots are taken, the limits on the standard deviation are approximately 2.57 and 13.4. In any case, the student should note the asymmetry of these limits about the sample variance (or standard deviation).

An element of the method which may be confusing is that the upper value from the χ^2 distribution is used in the computation of the lower limit on the variance. Remember that the relationship between χ^2 and σ^2 is inverse; that is, a low value of σ^2 results in a high χ^2 and vice versa.

EXERCISE 5-8. Compute 90 per cent confidence limits on the variance in the steel-wire illustration.

Sometimes we are not concerned with placing a lower confidence limit on the variance but are concerned only with an upper confidence limit. In such cases we say that the variance is equal to or less than some specific figure. We make this statement with "confidence probability" $1 - \alpha$. We construct such an upper limit by finding

$$C_2 = \frac{(n-1)s^2}{\chi^2_{(1-\alpha,n-1)}} \tag{5-10}$$

where the χ^2 is that value which has $1 - \alpha$ of the area *to the right of it.* The interval $0 < \sigma^2 < C_2$ is called a "one-sided confidence interval." There is, of course, a corresponding confidence interval, which may be written $C_1 < \sigma^2 < \infty$, which is the other one-sided confidence interval.

EXERCISE 5-9. Find the lower one-sided confidence-interval estimate for the variance in the steel-wire illustration, letting $\alpha = 0.01$.

Sometimes we have data from two samples (say, an experimental group and a control group) and wish to test hypotheses about differences between the two populations. We shall consider the test for differences in the variances.

If Y_1 and Y_2 are normally distributed variables drawn independently from the same normal population and if s_1^2 and s_2^2 are their respective variances, then the ratio s_1^2/s_2^2 is distributed as F. Here we have a new distribution with which to work. It is clear that if s_1^2 is very small and s_2^2 is very large, F will approach 0. Conversely, F may approach ∞. If s_1^2 and s_2^2 come from the same population (which is our hypothesis), then the average value of F is near 1. We see then that F is highly skewed to the right, ranging from 0 to ∞, with average value near 1 (unity).

The F distribution has two sets of degrees of freedom associated with it rather than one as in the t distribution or the χ^2 distribution. One value for degrees of freedom is associated with the numerator of the F

ratio and the other with the denominator. The shape of the F distribution for 9 and 4 degrees of freedom is shown in Fig. 5-10.

Since both values for degrees of freedom must be considered, tabling of the F distribution is quite complex, and only the 1, 5, and 10 percentage points are shown in Table IV, Appendix. Furthermore, only the right-hand (upper) end of the distribution is shown. If we adopt the convention of placing the larger variance in the numerator, we can make satisfactory use of the upper-end tables.

Let us examine how this affects the probability level. Under the hypothesis that s_1^2 and s_2^2 both come from the same population, the ratio s_1^2/s_2^2 will be less than 1 for part of the time and greater than 1 for part of the time. Its average value will be near 1, however. Thus, if we had the entire F table, the hypothesis $\sigma_1^2 = \sigma_2^2$ would call for the usual two-tailed test with $\alpha/2$ of the area at the left and $\alpha/2$ of the area at the right.

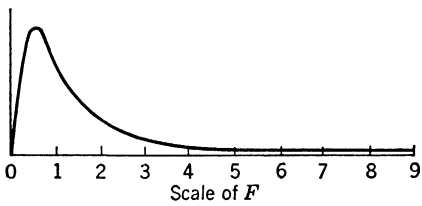

FIG. 5-10. F distribution with 9 and 4 degrees of freedom

For $\alpha = 0.10$ we should find the 5 and 95 percentage points, and the region outside these points would be the region of rejection. Since we decide to put the larger variance in the numerator, the ratio s^2(larger)/s^2(smaller) will always be greater than 1, and the entire region of rejection must be placed at the upper end of the F distribution. Therefore, placing the entire 5 per cent region of rejection at the upper end is equivalent to a two-tailed test of the hypothesis $\sigma_1^2 = \sigma_2^2$, with the region of rejection split 0.05 at the lower end and 0.05 at the upper end. It is, in fact, a test with $\alpha = 0.10$ for the two-tailed alternative.

If one needs to know the values of F at the lower end of the distribution, he may find them by the following relationship:

$$F_{(1-\alpha,\text{ with } a \text{ and } b\ df)} = \frac{1}{F_{(\alpha,\text{ with } b \text{ and } a\ df)}} \tag{5-11}$$

Suppose that two sources of raw materials are under consideration. Both sources seem to have about the same average characteristics, but we wonder about their variability. A sample of 13 lots from source A has a variance of 146 and a sample of 21 lots from source B has a variance of 200. Is it likely that their true variances are equal?

1. H_0: $\sigma_1^2 = \sigma_2^2$. *Alternatives:* $\sigma_1^2 < \sigma_2^2$ or $\sigma_2^2 < \sigma_1^2$.
2. $\alpha = 0.10$ (an arbitrary decision on our part).
3. $F = s_l^2/s_s^2$ with $n - 1$ degrees of freedom, where the subscript l indicates larger variance and s indicates smaller variance (not sample sizes).

4. The region of rejection is found from Table IV to be $F \geq 2.54$ (with 20 and 12 degrees of freedom). Remember that we must use the 5 per cent table.
5. $F = \frac{200}{148} = 1.37$.
6. Do not reject the hypothesis, since the computed F is not greater than 2.54.
7. Since we have not been able to reject the hypothesis of equality (statistically), management appears to be free to choose either source. However, if the sources are equal in all other respects, one would choose the source A, since it has the smaller sample variance. If the decision about source is critical, one might increase the size of the sample before reaching a final decision.

The discussion under step 7, above, raises a point which might well be discussed here: Why not determine the sample size based upon requirements for the managerial decision? Suppose we decide that, if the standard deviation of one source is greater than the standard deviation of the other by 20 per cent, preference will be given to the source with the smaller variation. Since our judgment is in terms of standard deviations and our probability distribution F is in terms of variances, we must convert standard deviations to variances. It is clear that 1.2 standard deviations is equivalent to 1.44 variances, so we want to know how large a sample to take from each population in order to have an F ratio of 1.44 significant at, say, the 0.10 level. Unless there is a difference in cost of taking samples, we should probably decide to take the same number from each population (source). Reference to the F table shows that about 80 samples from each source would be adequate to show the specified difference as significant at the 0.10 level. This approach to the problem makes considerably more sense than blundering ahead with an arbitrarily chosen sample size. The process can be improved still further by a device called *sequential sampling*, in which samples are drawn and examined consecutively until a decision is reached.

5-4. A REVIEW OF PROBABILITY DISTRIBUTIONS

At this point we have introduced all the common statistical probability distributions. They are the binomial, hypergeometric, Poisson, normal, t, χ^2, and F. There are, of course, many others, but they are used less frequently then these. It seems advisable to review again the concepts which give rise to these common distributions.

The binomial, hypergeometric, and Poisson distributions are discrete. That is, there is a set of specific values which the variable y can attain. This is easily remembered if we recall that with these distributions we are concerned with the number of "successes," or the number of occurrences

of an event in a given class of events. The y variable is a "counted" variable and is discrete.

The normal, t, χ^2, and F distributions are continuous distributions. They are all based on the assumption that the sample values are drawn from a normal population. The t, χ^2, and F distributions are called "derived" distributions because they are derived by assuming normality of the population from which the sample (or the pair of samples, as the case may be) was drawn. It may be helpful to see how one could derive such distributions empirically, that is, by experimentation only, rather then by mathematical proof. The procedure is, of course, extremely tedious, and it is fortunate indeed that we do not need to employ this technique in practice.

First, let us assume that we have a normal population from which to draw. This is hard to grasp because a normal population must contain infinitely many values. Let's compromise and assume that we have a population which is approximately normally distributed. Such a population is provided for us by a table of random normal deviates.* This is essentially a table of Z values which are random in the same sense that a table of random digits is random. They are Z values because their average is 0 and their standard deviation is 1.

Let us assume that we shall draw samples of 5 numbers from such a table. Any number other than 5 (except 1) would serve just as well since we are attempting to illustrate a general principle. Suppose our first set of 5 is -1.016, -0.838, 0.637, 0.267, -0.142. The sample mean of these 5 observations is -0.218, and the variance is 0.498307. Remember that the true mean of the population is 0 and the true variance is 1.000000. The differences between these figures and the sample estimates represent sampling variation. The sample standard deviation is found to be 0.706. Now, suppose we draw 100,000 such samples of 5 observations. We then have 100,000 sample means, 100,000 sample variances, and 100,000 sample standard deviations.

For *each* of the 100,000 samples we compute

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

remembering that in our particular population μ is 0 and may be omitted from the computation. We now have 100,000 t values. If we put them into class intervals, using a very narrow interval width, we get a frequency distribution of t values. Then, if we divide the frequency of each class interval by 100,000, we shall obtain the fraction of the total that falls

* See, for example, The Rand Corporation, *A Million Random Digits with* 100,000 *Normal Deviates*, Glencoe, Ill., Free Press, 1955.

within the interval. This number is an approximation to the probability that t with 4 degrees of freedom will lie within the specified interval. We shall have constructed, then, an approximation to the distribution of t with 4 degrees of freedom.

In an actual problem we have only one set of sample values, but that is all we need because we know the distribution of t and need not reconstruct it every time a problem arises. The basis for a test of the hypothesis is to compute t for the particular set of values obtained from the sample and then to ask whether this is an unusual t value when judged against the average of all possible values of t (the t distribution). If it is unusual, that is, if it lies far out toward the tail of the distribution, we conclude that the hypothesis *upon the basis of which the t computation was made* is not true.

Now, let us consider how we might construct a χ^2 distribution from the 100,000 samples drawn (hypothetically) above. Here, for each sample, we should compute

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

giving us 100,000 χ^2 values. When plotted and reduced to percentages, they would give a close approximation to the χ^2 distribution with 4 degrees of freedom.

Similarly, if we paired the samples by using the first and second, third and fourth, and so on, we could get an approximation to the F distribution. Letting the first sample variance of the pair be denoted by s_1^2 and the second by s_2^2, we could compute

$$F = \frac{s_1^2}{s_2^2}$$

There would be 50,000 such F values, and they would form a close approximation to the F distribution with 4 and 4 degrees of freedom. Again, an F value is judged unusual if it lies far away from the center of the distribution. It is indicative, therefore, that the hypothesis upon which it is based is false.

The generation of t, χ^2, and F by the drawing of random samples as suggested above yields special cases of those distributions. They are, however, the most common cases encountered in statistical work. We now introduce more general definitions which apply to a broader range of cases.

There are two distributions which are basic to our definitions. They are the normal distribution and the χ^2 distribution. The normal distribution has been previously defined by the probability density

$$f(Y) = (1/\sigma\sqrt{2\pi})e^{-(Y-\mu)^2/2\sigma^2}$$

The χ^2 density is a special form of the so-called "gamma distribution" and may be defined as follows:

$$f(\chi^2) = \frac{(\chi^2)^{(\nu/2)-1}e^{-\chi^2/2}}{2^{\nu/2}[(\nu/2) - 1]!} \tag{5-12}$$

where ν is degrees of freedom (the only parameter). This functional form of the χ^2 distribution is useful in theoretical work and in the construction of tables, but is used little by the applied statistician.

One special application of the χ^2 distribution which we have encountered arises when a sample of size n has been drawn at random from a normal population. Then the ratio

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\Sigma(Y_i - \bar{Y})^2}{\sigma^2} \tag{5-13}$$

is distributed as χ^2 with $n - 1$ degrees of freedom. If it should happen that we know the population mean μ, then the ratio

$$\frac{\Sigma(Y_i - \mu)^2}{\sigma^2} \tag{5-14}$$

is also distributed as χ^2, but now with n degrees of freedom. We shall encounter still another application in Chap. 6.

It is possible for us to define the t ratio in terms of the normal distribution and the χ^2 distribution. The t distribution has a probability density, of course, which will not be given here. The t ratio which we encounter in practice may be said to be the ratio of a variable which is normally distributed with mean 0 and variance 1 and the square root of a variable which is independent of the numerator and which is distributed as χ^2/df. We have used the symbol Z to stand for a normal variable with variance 1 and mean 0, so we could write t as

$$t = \frac{Z}{\sqrt{\chi^2/df}} \tag{5-15}$$

Now, let us see whether the ratio

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

meets these requirements. We may write

$$t = \frac{\bar{Y} - \mu}{\sqrt{s^2/n}} = \frac{(\bar{Y} - \mu)/\sigma}{\sqrt{s^2/n\sigma^2}} = \frac{(\bar{Y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}} = \frac{Z}{\sqrt{\chi^2/df}} \tag{5-16}$$

We shall find uses for this more general definition of t in later chapters.

The F ratio can be defined entirely in terms of χ^2 distributions. Suppose we have two independent χ^2's, say, χ_1^2 and χ_2^2 with ν_1 and ν_2 degrees of freedom. Then the ratio

$$\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \tag{5-17}$$

is distributed as F with ν_1 and ν_2 degrees of freedom. Let us see whether the familiar form of the F distribution satisfies the definition

$$F = \frac{s_1^2}{s_2^2} = \frac{s_1^2/\sigma^2}{s_2^2/\sigma^2} = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \tag{5-18}$$

Note that division of both numerator and denominator by σ^2 to form the χ^2 distributions automatically specifies the hypothesis $\sigma_1^2 = \sigma_2^2$.

The relationship between the t distribution and the F distribution is sometimes useful. That is, F with 1 and ν_2 degrees of freedom is equal to t^2 where t has ν_2 degrees of freedom. Note that the relationship holds only when the numerator of the F ratio has only 1 degree of freedom. The student is urged to examine the tables to observe this relationship.

5-5. INFERENCES ABOUT TWO MEANS

Generally, we are more interested in testing hypotheses about means than about variances. When we compare a control group with an experimental group, or when we compare two experimental groups, we are concerned primarily with differences in their means. From the point of view of methodology, we must distinguish two cases, one in which the two samples are independent and the other in which observations are paired.

Independent samples. We need not concern ourselves with a rigorous definition of *independence* here. It is sufficient to say that we may consider two samples independent when the individuals in one cannot be related to, or associated with, the individuals of the other in a meaningful manner. The concept will be clearer after we have discussed the case of paired samples later in this section.

Suppose that two worktable designs are under study. A time and motion study shows that 10 employees working on table A have an average assembly time of 250 sec with variance of 100 and that 20 employees on table B have an average assembly time of 225 sec with variance of 80. We want to know whether the average assembly times for the two worktables are significantly different.

The following rule governs our method: If Y_1 and Y_2 are normally distributed variables with means μ_1 and μ_2 *and the same variance* σ^2, then

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{(1/n_1) + (1/n_2)}} \tag{5-19}$$

is distributed as t with $n_1 + n_2 - 2$ degrees of freedom, where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5-20}$$

The assumption that the variances are homogeneous, that is, that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, is capable of being tested as a hypothesis by the methods in the previous section. We now proceed with this test, letting $\alpha = 0.10$.

1. H_0: $\sigma_1^2 = \sigma_2^2$. *Alternatives:* $\sigma_1^2 < \sigma_2^2$ or $\sigma_2^2 < \sigma_1^2$.
2. $\alpha = 0.10$.
3. $F = s_l^2/s_s^2$ with $n_l - 1$ and $n_s - 1$ degrees of freedom.
4. *Region of rejection:* $F \geq 2.43$.
5. $F = \frac{100}{80} = 1.25$.
6. Do not reject.

We may now proceed with the test of the means.

1. H_0: $\mu_1 = \mu_2$. *Alternatives:* $\mu_1 < \mu_2$ or $\mu_2 < \mu_1$.
2. $\alpha = 0.01$ (let's say).
3. $t = (\bar{Y}_1 - \bar{Y}_2)/s_p \sqrt{(1/n_1) + (1/n_2)}$.
4. *Region of rejection:* $t \leq -2.76$ or $t \geq 2.76$.
5. $$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{9(100) + 19(80)}{28}} \sqrt{\frac{1}{10} + \frac{1}{20}} = \sqrt{\frac{72{,}600}{5{,}600}} = 3.6.$$
   $$t = \frac{250 - 225}{3.6} = 6.9.$$
6. Reject hypothesis of equality and accept alternative that $\mu_2 < \mu_1$.

We could have tested the hypothesis that $\mu_1 = \mu_2$ against the one-sided alternative $\mu_2 < \mu_1$. Had we chosen to do so with $\alpha = 0.01$, our region of rejection would have been $t \geq 2.47$. Only sample results in which $\bar{Y}_1$ exceeds $\bar{Y}_2$ furnish evidence to reject the hypothesis $\mu_1 < \mu_2$; hence only the right-hand end of the t distribution is in the region of rejection.

When the hypothesis $\sigma_1^2 = \sigma_2^2$ must be rejected, unfortunately there is no very good test for the means which can be substantiated in statistical theory. One fairly common approximate test is the following:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{s_{\bar{Y}_1}^2 + s_{\bar{Y}_2}^2}} \tag{5-21}$$

with

$$\frac{(s_{\bar{Y}_1}^2 + s_{\bar{Y}_2}^2)^2}{[(s_{\bar{Y}_1}^2)^2/(n_1 + 1)] + [(s_{\bar{Y}_2}^2)^2/(n_2 + 1)]} - 2\,df$$

where $s_{\bar{Y}_1}^2$ is the variance of the first sample mean and $s_{\bar{Y}_2}^2$ the variance of the second sample mean. The computation for degrees of freedom usually does not yield an integer, but one may obtain a value for the region of rejection by interpolating in the t table.

Dependent samples. When samples are paired, we can relate each observation in one sample with some particular observation in the

second sample. Pairing may be accomplished if the same individual has a score in each sample or if individuals on whom observations are taken are paired or matched by sex, age, and other characteristics.

Suppose that efficiency in an operation is measured by a score which takes into account quantity of production as well as quality. Twenty-five employees are scored before and after taking a refresher course. The results are shown in the first three columns of Table 5-1. (The last two columns will be explained presently.)

To effect the test, we subtract the second observation from the first for each employee, arriving at a new variable D. If there is no difference

Table 5-1. Scores of 25 Employees before and after Refresher Training Course, with Computations for the Test of the Hypothesis $\mu_1 = \mu_2$

Employee number	First test	Second test	D	D^2
1	125	179	−54	2,916
2	135	168	−33	1,089
3	78	120	−42	1,764
4	116	171	−55	3,025
5	128	171	−43	1,849
6	185	238	−53	2,809
7	138	152	−14	196
8	177	225	−48	2,304
9	159	182	−23	529
10	160	176	−16	256
11	151	142	9	81
12	110	165	−55	3,025
13	140	170	−30	900
14	153	175	−22	484
15	194	180	14	196
16	106	158	−52	2,704
17	142	194	−52	2,704
18	170	218	−48	2,304
19	195	244	−49	2,401
20	172	193	−21	441
21	185	202	−17	289
22	128	151	−23	529
23	151	221	−70	4,900
24	153	202	−49	2,401
25	138	196	−58	3,364
Total	3,689	4,593	−904	43,460

between the first and second samples, D will average 0, so our test procedure is based upon testing the hypothesis that the average of D is 0. We require the standard error of the average D and employ the t test.

1. H_0: $\mu_1 = \mu_2$. *Alternative:* $\mu_1 > \mu_2$.
2. $\alpha = 0.01$.
3. $t = \bar{D}/s_{\bar{D}}$ with $n - 1$ degrees of freedom, where n is the *number of pairs* and $s_{\bar{D}}$ is $s_D/\sqrt{n}$, the standard error of the average D.
4. *Region of rejection:* $t \leq -2.49$.
5. *Computations:*

$$\bar{D} = \Sigma D/n = -904/25 = -36.16.$$
$$s_D^2 = [n\Sigma D^2 - (\Sigma D)^2]/n(n-1) = [25(43{,}460) - (-904)^2]/25(24) = 448.8.$$
$$s_{\bar{D}} = \sqrt{(448.8/25)} = 4.24.$$
$$t = \bar{D}/s_{\bar{D}} = -36.16/4.24 = -8.5.$$

6. Reject the hypothesis that scores before training were equal to or better than scores after training. The training program apparently was of some benefit.

EXERCISE 5-10. An experiment was conducted to determine the percentage of free fatty acids produced by frying sardines by two different methods. The results of the experiment are as follows:

Sample number	Free fatty acids, per cent of total weight	
	Heated in air	Heated away from air
1	0.7	0.3
2	0.4	0.1
3	1.6	0.9
4	1.2	1.0
5	0.7	0.5
6	0.8	0.5
7	1.2	0.6

SOURCE: U.S. Bureau of Fisheries, Doc. No. 1020, p. 169.

(*a*) Can you conclude that the two methods yield different results? Note that the samples are not independent. (*b*) What would your answer be if the samples were independent?

In using the $\bar{D}$ test one can easily obtain a confidence-interval estimate of the difference between the population means by the following device:

$$P_c[\bar{D} - t_{(\alpha,n-1)}s_{\bar{D}} < \Delta < \bar{D} + t_{(\alpha,n-1)}s_{\bar{D}}] = 1 - \alpha \qquad (5\text{-}22)$$

where Δ is the true difference between the population means.

EXERCISE 5-11. Place 90 per cent confidence limits on the mean difference in the sardines illustration (Exercise 5-10).

Where two samples are independent, a confidence-interval estimate of the true difference between the population means is provided by

$$P_c[(\bar{Y}_1 - \bar{Y}_2) - t_{(\alpha, n_1+n_2-2)}s_{(\bar{Y}_1-\bar{Y}_2)} < \mu_1 - \mu_2 < (\bar{Y}_1 - \bar{Y}_2) + t_{(\alpha, n_1+n_2-2)}s_{(\bar{Y}_1-\bar{Y}_2)}] = 1 - \alpha \quad (5\text{-}23)$$

where $s_{(\bar{Y}_1-\bar{Y}_2)}$ is the standard error of the difference between two means.

EXERCISE 5-12. Place 95 per cent confidence limits on the difference between the means in the time-and-motion-study illustration at the beginning of this section. Notice that, if one limit is positive and the other negative, we must accept the hypothesis that $\mu_1 = \mu_2$. This again illustrates the correspondence between confidence intervals and tests of hypotheses.

EXERCISE 5-13. An experiment is conducted to see whether sales of a new line of neckties can be increased by "pushing" the line. The experiment is conducted in the following manner. There are 11 clerks in the men's clothing department. Six of them are chosen at random and told to suggest the line of ties to the customer, regardless of his purchase. The remaining 5 clerks are told nothing about the experiment. The number of neckties sold during the day are taken from the sales tickets and recorded for each clerk. The results follow:

Control group		Experimental group	
Employee	Sales	Employee	Sales
A	8	F	14
B	10	G	7
C	5	H	12
D	6	I	17
E	8	J	11
		K	12

(*a*) Test the hypothesis that the variances are equal for the two groups. (*b*) Test the hypothesis that the means are equal. (*c*) What are the populations from which these figures are samples? (*d*) Place 90 per cent confidence limits on the difference in means.

EXERCISE 5-14. Assume that in the above experiment the clerks had been assigned to the two experimental groups as follows. The clerks were ranked according to their previous month's sales, with the clerk having the highest sales listed first: B, I, H, A, E, F, D, K, J, C, G. The clerks then were paired as follows: (B,I), (H,A), (E,F), (D,K), (J,C). G was omitted from the experiment. From each pair one was chosen by the toss of a coin to receive the special instruction. (*a*) If this had been the experimental situation, answer parts (*b*) and (*d*) of Exercise 5-13. (*b*) Is this a better experimental design? Why?

EXERCISE 5-15. (*a*) What are the differences in the assumptions between the two exercises above? (*b*) How do you feel about their validity?

5-6. DETERMINATION OF SAMPLE SIZE

In the preceding pages it has been assumed that a sample of given size n has been taken and that on the basis of the sample observations one is to make judgments about certain population parameters. Frequently, however, the researcher is required not only to analyze data but to design an experiment in such a way that a certain prescribed precision is obtained. In the simplest case this includes the determination of sample size.

Suppose that the researcher has some information, obtained through sampling surveys, or by any other means, to indicate that the population variance is 100. He wishes to determine a sample size such that a difference of 3 from the hypothesis μ will be significant at the 0.01 level of significance. Since the direction of the difference is not specified, a two-tailed test may be assumed. We write the normal deviate in the form

$$Z_\alpha = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

where Z_α is the deviate required for significance at the 0.01 level and n is the required sample size. Reference to a table of areas of the normal curve (or to the infinity row of the t table) shows that $Z_{0.01} = 2.58$. Also, we know from the conditions of the problem that $\bar{Y} - \mu = 3$ and $\sigma = 10$. Solving for n and substituting the known data, we have

$$n = \frac{Z_\alpha^2 \sigma^2}{(\bar{Y} - \mu)^2} = \frac{(2.58)^2(100)}{3^2} = 74$$

which is the required sample size.

In the above illustration it was assumed that the researcher had considerable information about the variance of the population, so that the estimate $\sigma^2 = 100$ was subject to little error. In the usual situation, however, the variance is estimated from a sample. If one's estimate of the variance is too low and he uses the above formula to approximate sample size, inserting s^2 for σ^2, his estimate of n will be too low to give him the precision required. If, on the other hand, his estimate of σ^2 is too high, an unnecessarily large sample will be taken, with resulting loss in economy.

If one wishes to be ultraconservative, he may set an upper confidence limit on s^2 by the usual methods and use this as an approximation for σ^2. The method is wasteful if additional sample values are costly to procure.

A method in which the sample may be taken in two stages is suggested by Stein.* A preliminary sample is drawn, using any information whatever to guess at sample size. As a matter of practice, the estimate is usually purposely planned to be too small. Additional sample units are drawn according to the following rules:

Let n_1 = size of preliminary sample
s_1 = standard deviation of preliminary sample
t_1 = corresponding t value for $n_1 - 1$ df and probability level α
d = desired half width of the precision interval, that is, the limits of desired variation on $\bar{Y} - \mu$

If
$$\frac{t_1 s_1}{\sqrt{n_1}} < d$$

then n_1 is large enough, and no additional sample values need be drawn.

If
$$\frac{t_1 s_1}{\sqrt{n_1}} > d$$

additional sample values are drawn until $n = t_1^2 s_1^2/d^2$. This method assures us that the probability of the sample mean's being more than d away from the true mean is less than or equal to α.

EXERCISE 5-16. It is believed that the standard deviation, in hours, for the life of an electric mechanism is 25. How large a sample is necessary to establish the average life within 2 hr with probability 0.90?

EXERCISE 5-17. If approximately 20 per cent of produced items are defective, how large a sample is required to determine the true percentage within 1 percentage point with probability 0.95?

5-7. CONFIDENCE INTERVALS VERSUS TESTS OF HYPOTHESES

Throughout this chapter we have discussed statistical inference from the viewpoint of tests of hypotheses as well as from the standpoint of confidence intervals. In practice, the business manager may be more interested in confidence-interval estimation than in tests of hypotheses. Testing the null hypothesis is in many cases vacuous. To assert, by hypothesis, that two machine speeds will produce the same number of good items per hour seems stupid. Of course they are different! Once the difference is estimated, it seems to lead naturally to the placement of a confidence-interval estimate on the difference to serve as an index of the reliability of the estimate.

* A Two-sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance. C. Stein, *Ann. Math. Statistics*, vol. 16, pp. 243–258.

We have seen previously how tests of hypotheses and confidence intervals are related. That is, if a $1 - \alpha$ confidence interval on the mean yields the figures $30 < \mu < 40$, we cannot reject the hypothesis that $\mu = 32$ (or any other value between 30 and 40) at the α level of significance. Since, in addition to providing the necessary information for a test of a hypothesis, the confidence interval provides a range whose width is indicative of one's faith in the estimation of the parameter, it is sometimes preferred to the test of a particular hypothesis.

There is considerable merit in the point of view that two populations are always different, since they would otherwise be the same population. Therefore, if one chooses a large enough sample from each, the difference will be revealed to be "significantly" different. If any real problem exists, it can be stated in these terms: How large a sample must be taken to reveal the difference as statistically significant a given proportion of the time? This approach might lead to a better understanding of the economic importance of a significant difference.

There is, however, a large body of statistical methodology which is based, and properly so, on the null hypothesis. This is that part of statistics known as *statistical quality control*, and its principles are outlined in a later chapter. Studying the testing of hypotheses leads to an understanding of the two types of error in decision making and their related probabilities, α and β. This background is necessary for an understanding of the risks involved in operating a quality-control sampling plan.

Another matter, about which there is frequent misunderstanding, is the role played by "other information" in the management decision. For example, if a coin is tossed three times in your presence and falls heads each time, will you still assume that the probability is $\frac{1}{2}$ of its coming up tails on the next toss? If your answer is yes, you are basing your judgment on information not contained in the experiment. Presumably, your reasoning was something like this: "I have never seen a two-headed coin, nor one which is terribly biased one way or the other. Therefore, the hypothesis that heads will appear half of the time is a logical hypothesis. Three heads in a row is insufficient evidence to reject this hypothesis; hence I will still assume that the probability of tails on the next toss is $\frac{1}{2}$."

Notice the role played by prior information about the world of coins in this illustration. It seems fairly safe to assume that a primitive man with some intelligence, but no knowledge of coins, would bet on heads coming up again. There is some evidence that rats exhibit this kind of "reasoning" in learning to turn right or left in a T maze in search of food. There are good grounds for believing that the primitive man (or the rat) is using better judgment than the statistics student in this case.

Another example of the same principle is the case in which a company

is deciding whether to purchase machine A or machine B on the basis of a test of the two machines. *In the absence of other information* the company should choose the machine which appears to be "best" on the trial, *whether the difference between the two is significant or not.* This assumes, of course, that the test takes into account any differences in cost as well as in output. Usually, however, other information is available, such as knowledge of what competitors are doing, experience with the maintenance policies of the two companies producing the machines, and knowledge of the engineering design of the machines. Were it not for this other information, there would be no role, other than a mathematical one, for the business manager.

EXERCISE 5-18. The following are some observations on the length of time, in minutes, that it takes a barber to cut hair: 18, 22, 20, 20, 21, 19, 23, 20, 28, 21, 20. Estimate the mean and place 95 per cent confidence limits on it. The barber plans to spend 20 min on each haircut. Are the figures in accord with this standard? Note that an answer is provided by the confidence interval itself, without the necessity for testing a separate hypothesis.

EXERCISE 5-19. Using the data of Exercise 5-18, place 95 per cent confidence limits on the number of haircuts the barber can expect to give (assuming an ever-present waiting line) in 8 hr.

EXERCISE 5-20. The barber of Exercise 5-18 took 3 other samples with the following results:

n	$\bar{Y}$	s^2
5	22.1	8.5
9	21 4	5.7
13	24 3	12.2

With these additional data, answer the questions of Exercise 5-18.

EXERCISE 5-21. Two methods of producing cartons are being tested. A sample of 10 produced by method A has a sample variance of 100. A sample of 15 from method B has a variance of 50. Place two-tailed 90 per cent confidence limits on the true ratio of the variance of A to the variance of B. Note that you must find lower F values by Eq. (5-11).

EXERCISE 5-22. An office manager wishes to estimate, within 1 invoice, the average output of invoices per hour of a clerk whose sole duty is preparation of invoices. He has observed that the clerk has never produced fewer than 2 invoices per hour or more than 17. How large a sample of hours should he take? *Hint:* Estimate the standard deviation from the observed range, assuming normality.

EXERCISE 5-23. An auto-parts company is interested in increasing the sales of a particular oil-filter unit. It holds a sales meeting at which the factory representative describes the product. There are 8 salesmen, and their sales for the week before the meeting and the week after are shown below:

Salesman	Sales	
	Week before	Week after
1	0	1
2	5	8
3	2	2
4	7	11
5	4	3
6	5	8
7	1	6
8	6	15

Place 90 per cent confidence limits on the amount sales were increased by the sales meeting.

EXERCISE 5-24. A preliminary sample of 10 yields $\bar{Y} = 47.8$, $s^2 = 81$. How many additional units must be sampled if one is to be at least 95 per cent confident that the true mean is estimated within 1 unit?

6. ANALYSIS OF FREQUENCIES

Sometimes the raw data to be analyzed consist simply of numbers of occurrences rather than measures on individual elements drawn into the sample. For example, we may observe the number of defective parts in a production process, number of days of employee absences, number of pedestrians, by sex, passing a store location, and so on. This chapter is devoted to the analysis of such "counted data."

6-1. THE BINOMIAL DISTRIBUTION

The binomial distribution was introduced in Chap. 2. Review of that presentation will show that the binomial distribution applies whenever:

1. The possible results can be classified into one of two mutually exclusive categories.
2. The probability (p) of classification into one of the categories remains constant from trial to trial.

We now investigate this distribution within the general framework for testing hypotheses.

A complex and expensive mechanism is being produced for the first time. Policies have been established in the belief that only 10 per cent of the mechanisms will require extensive adjustment during the 6-month guarantee period. Of the first 8 units produced, 4 required extensive adjustment before 6 months expired. Must we reject the hypothesis that, on the average, only 10 per cent will require adjustment? In binomial language, $n = 8$, $p = 0.10$, $q = 0.90$, and the probability of Y defective units in 8 trials is

$$_8C_Y(0.1)^Y(0.9)^{8-Y}$$

We can formalize the testing procedure in the usual manner.

1. H_0: $p = 0.1$. H_1: $p > 0.1$. (We have only one alternative hypothesis, since we are not concerned if our product is *better* than anticipated.)

2. $\alpha \leq 0.05$. (The reason for the inequality is that we are working with a discrete variable, and we may not be able to find a number of defective items such that the probability of obtaining that many, or more, is exactly equal to α.)
3. The probability distribution is the binomial distribution.
4. The region of rejection is the possible number of defective items at the upper end of the scale such that the sum of their probabilities will be less than or equal to 0.05, but such that the addition of another value would make the sum greater than 0.05. The principle is much easier to explain by illustration, as shown in Table 6-1.

Table 6-1. Probabilities of Various Numbers of Defective Items with $n = 8$, $p = 0.1$

Number of defectives, Y	Probability of Y defectives	Cumulative probability
8	${}_8C_8(0.1)^8(0.9)^0 = 0.00000$	0.00000
7	${}_8C_7(0.1)^7(0.9)^1 = 0.00000$	0 00000
6	${}_8C_6(0.1)^6(0.9)^2 = 0.00002$	0.00002
5	${}_8C_5(0\ 1)^5(0.9)^3 = 0\ 00041$	0.00043
4	${}_8C_4(0.1)^4(0.9)^4 = 0.00459$	0.00502
3	${}_8C_3(0.1)^3(0.9)^5 = 0.03307$	0.03809
2	${}_8C_2(0.1)^2(0.9)^6 = 0.14881$	0.18690

The probability of obtaining a value of Y equal to or greater than 3 is 0.038. The probability of obtaining a value of Y equal to or greater than 2 is 0.187. Hence, 3 is the smallest value of Y such that the probability of getting that number or more of defective items is less than or equal to 0.05. Therefore the region of rejection is $Y \geq 3$.

5. Our observed value of Y is 4, which lies in the region of rejection.
6. Reject the hypothesis that $p = 0.1$.

Computation of binomial probabilities can be avoided by a good set of binomial probability tables.*

6-2. THE NORMAL APPROXIMATION TO THE BINOMIAL

When n is larger, it becomes tedious to compute the exact binomial probabilities or even, in some cases, to read them from a table. In these circumstances we can employ the normal distribution as an approximation to the binomial. The following principle governs our procedure: Let y be a binomial variable with fixed probability p in a single trial. As n becomes larger without limit, the distribution of the ratio $(Y - np)/\sqrt{npq}$ approaches the standardized normal distribution, that is, the

* Such as *Tables of the Cumulative Binomial Probability Distribution*, by the Staff of the Harvard University Computation Laboratory, Harvard University Press, Cambridge, Mass., 1955.

normal distribution with mean 0 and variance 1. For any value of n, the mean of the binomial is np and its variance is npq, so we see that the above ratio is, in effect, the familiar $(Y - \mu)/\sigma$.

We cannot use the normal approximation indiscriminately for any values of n and p. Cochran* suggests the sample sizes necessary before application of normal-curve procedure to binomial populations. The recommended sample sizes for various values of p are given in Table 6-2.

Table 6-2. Recommended Minimum Sample Sizes for Use of Normal Approximation

Observed proportion of success, p	*Sample size for normal approximation to apply*
0.5	30
0.4 or 0.6	50
0.3 or 0.7	80
0.2 or 0.8	200
0.1 or 0.9	600
0.05 or 0.95	1,400

Suppose we consider the problem of Sec. 6-1 again, assuming now that we have produced 800 units which have been in service 6 months or more and that 100 of these have required major maintenance. Can we reject the hypothesis that $p = 0.1$?

1. H_0: $p = 0.1$. H_1: $p > 0.1$.
2. $\alpha = 0.05$.
3. *Probability distribution:* normal, with $Z = (Y - np \pm \frac{1}{2})/\sqrt{npq}$. The $\pm\frac{1}{2}$ is called a *continuity correction.* It has the effect of adjusting for the use of a continuous distribution (the normal) to estimate the probability in a discrete distribution (the binomial). The plus sign is used if $Y - np$ is negative and the minus sign if $Y - np$ is positive. The continuity correction has the effect of replacing the integer Y by a point halfway between Y and the next higher (or lower) integer. Remember that Y can only take on integral values, such as 97, 98, 99, 100. Suppose we use a continuous distribution to find the probability that $Y = 100$ or more and add to it the probability that $Y = 99$ or less. The two probabilities will not add to 1 because we have omitted the interval from 99 to 100 on the continuous scale. To avoid this difficulty, we let $Y = 99.5$ in finding the probability that $Y = 100$ or more. If n is large (such as in this case), there is little error in omitting the continuity correction.
4. *Region of rejection:* $Z \geq 1.645$. (Note that a one-tailed test is called for.)

* By permission from William G. Cochran, *Sampling Techniques*, p. 41, John Wiley & Sons, Inc., New York, 1953.

5. $Z = [100 - 800(0.1) - \frac{1}{2}]/\sqrt{800(0.1)(0.9)} = 19.5/\sqrt{72} = 2.30.$
6. Reject the hypothesis.

Since we reject the hypothesis that $p = 0.1$ we should substitute a revised estimate of p. Our best estimate at this point is $p = \frac{100}{800} = 0.125$.

Instead of making the test on frequencies we can make the test on proportions. Let p' be the observed proportion of defective units. The standard error of a proportion is $\sqrt{pq/n}$ where p is the hypothesis being tested.

1. H_0: $p = 0.1$. H_1: $p > 0.1$.
2. $\alpha = 0.05$.
3. *Probability distribution:* normal, with $Z = (p' - p \pm 1/2n)/(\sqrt{pq/n})$.
4. *Region of rejection:* $Z \geq 1.645$.
5. $Z = (0.125 - 0.1 - 1/1{,}600)/\sqrt{0.1(0.9)/800} = 2.30.$
6. Reject the hypothesis.

Since the stated hypothesis must be rejected, it may be of interest to place confidence limits on the true proportion of units requiring extensive maintenance. A difficulty arises because the standard error of the proportion defective is $\sqrt{pq/n}$, but we do not know p or q. However, we may replace p by p' and q by $1 - p'$ in order to arrive at an estimate of the standard error. It can be shown that division by $n - 1$ rather than n results in an improved estimate, but since n is large in any situation where the standard error is used, this refinement will be ignored. The form for the confidence-interval estimate is

$$P_c\left[p' - t_{(\alpha,\infty)}\sqrt{\frac{p'q'}{n}} < p < p' + t_{(\alpha,\infty)}\sqrt{\frac{p'q'}{n}}\right] = 1 - \alpha \qquad (6\text{-}1)$$

Note that the t value with infinite degrees of freedom has been used. As we have seen earlier, this is, in fact, the normal deviate. It is more convenient to obtain this constant from the infinity row of the t table than from the table of areas under the normal curve.

EXERCISE 6-1. Out of 5,000 bolts examined, 200 had defective threads. Is this in accord with the advertised standard that there are not more than 3 per cent defective?

EXERCISE 6-2. Place 99 per cent confidence limits on the above proportion defective.

EXERCISE 6-3. Suppose we estimate that $p = 0.5$ when actually it equals 0.4. How much is the standard error affected by our error in estimating p?

EXERCISE 6-4. Plot a curve showing the values of the standard error $\sqrt{pq/n}$ for all values of p from 0 to 1; compute for values of p equal to 0, 0.1, 0.2, 0.3, and so on; and draw the curve by approximation.

EXERCISE 6-5. Plot a curve showing the values of the coefficient of variation $(\sqrt{pq/n})/p$ for all values of p from 0 to 1.

6-3. THE CHI-SQUARE TEST ON FREQUENCIES

In the previous chapter we have seen how the χ^2 distribution can be used to test hypotheses about the variance. In this section we discuss another application, that of testing hypotheses about frequencies. The sum

$$\chi^2 = \sum \frac{(|f_o - f_c| - \frac{1}{2})^2}{f_c} \tag{6-2}$$

where f_o is the observed frequency, f_c is the computed (or theoretical) frequency, the vertical lines mean "the absolute value of," and the summation is over all classes being compared, is distributed approximately as χ^2 with degrees of freedom equal to the number of classes compared less the number of restrictions imposed on the comparison. A few examples in this and later sections will help to clarify the computation of degrees of freedom. The subtraction of the constant $\frac{1}{2}$ is a continuity correction, similar to that employed in the previous section.

Suppose that in our organization we have 300 women employees and 700 men employees. In a given year we observe 3,500 days of absence on the part of the men and 1,800 days of absence on the part of the women. We wish to know whether the difference in rate of absenteeism between men and women is significant at the 0.05 level.

1. H_0: Days of absenteeism are distributed in the ratio 3:7. H_1: The ratio is not 3:7.
2. $\alpha = 0.05$.
3. *Distribution:* $\sum \frac{(|f_o - f_c| - \frac{1}{2})^2}{f_c} = \chi^2$ (approximately) with 1 degree of freedom. Since there are two categories and we employ one linear restriction, namely, that the observed and theoretical sums must be equal, we have 1 degree of freedom remaining.
4. *Region of rejection:* $\chi^2 \geq 3.84$ (from Table III). Note that any departure from the 3:7 ratio will cause an inflation of χ^2. Therefore, the entire region of rejection must be at the upper end of the χ^2 distribution.
5. *Computations:*

Sex	f_o	f_c	$\lvert f_o - f_c\rvert - \frac{1}{2}$	$(\lvert f_o - f_c\rvert - \frac{1}{2})^2/f_c$
Men	3,500	3,710	209.5	11.83
Women	1,800	1,590	209.5	27.60
Total	5,300	5,300		39.43

$\chi^2 = 39.43$ with 1 degree of freedom.

6. Reject the hypothesis that absenteeism is equally distributed between the sexes. In our case (purely theoretical, of course) women tend to take more time off than men.

A word of caution is in order concerning the use of the test. It should not be used when the theoretical frequency in any class is very small, say, less than 5, since division by the small denominator results in an inflated χ^2 value. For example, suppose $f_o = 2$, $f_e = 0.1$; then the contribution of this one class to χ^2 would be $(1.4)^2/0.1 = 19.6$; yet the difference between observed and theoretical frequency is only 1.9.

When we have one or more classes with small frequencies, sometimes we can group them in some meaningful manner before employing the test. At other times we may use the binomial or multinomial* distribution to obtain exact probabilities.

EXERCISE 6-6. Redo Exercise 6-1 using the χ^2 criterion.

EXERCISE 6-7. A die is cast 120 times with the following results:

Face	*Number of times*
1	15
2	20
3	27
4	28
5	14
6	16

Do you feel that the die is fair? Use the χ^2 test, noting that, since there are 6 categories being compared, there are 5 degrees of freedom.

EXERCISE 6-8. In an accounting department 50 accounts are examined for errors. The following information is gathered:

Number of errors per account	*Number of accounts*
0	25
1	15
2	8
3	0
4	1
5	1

* The multinomial distribution is a simple extension of the binomial. Let there be k classes, and let p_i $(i = 1, 2, \ldots, k)$ be the probability of obtaining an element of the ith class in a single trial. We have $\Sigma p_i = 1$. Let n be the number of trials and y_i $(i = 1, 2, \ldots, k)$ be the number of sample elements in the ith class. The $\Sigma y_i = n$. The distribution of the y_i is as follows:

$$P(y_1, y_2, \ldots, y_k) = \frac{n!}{y_1! y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$$

Note that the p_i are assumed to remain constant from trial to trial; otherwise the hypergeometric distribution is used.

Test the hypothesis that the errors are distributed by the Poisson law. (*a*) Find the average errors per account and call them m. (*b*) Fit the Poisson distribution using this value of m. (*c*) Compare the frequencies using χ^2, grouping the frequencies so that there are 5 or more computed frequencies in each class. The degrees of freedom will be 2 less than the number of classes being compared. One degree of freedom is lost by estimating m from the data, and the other degree of freedom is lost because the frequencies are made to equal 50 in each case.

6-4. CROSS CLASSIFICATION

In an earlier chapter we saw that "classification" of data simply means the grouping together of similar observations. Sometimes the criterion of classification is a numerical one. An example is the formation of class intervals in a frequency distribution. Sometimes the criterion for classification is nonnumerical, such as sex, color of eyes, success or failure, and so forth. Also, it may be that observations can be classified by two or more criteria simultaneously. When this is done, we refer to the resulting classification as a cross classification.

A complete cross classification (for example, Table 6-3) permits the reader to determine all the factors of classification about a particular figure by inspection of the stub (the left-hand column) and the column captions. It is an objective to be desired in tabulation.

Table 6-3. Number of Grocery Stores in Cinder City by Ownership and Sales Volume for 1955 (Complete Classification)

Sales volume, 1955	Type of ownership		
	Total	Home-owned	Chain
Total	41	32	9
Under $100,000	31	29	2
$100,000 and over	10	3	7

A three-way classification necessitates repetition of one of the classifications either in the stub or in the column headings. The student will soon discover that repeating the classification in the stub is much more economical in terms of space used than repeating the column headings. Table 6-4 is an example of a complete three-way classification.

Now, let's see whether you can read Table 6-4. Try to answer the following questions:

1. How many people were interviewed? Watch this one! The correct answer is "I don't know," or perhaps "At least 560." The title says

Table 6-4. Number of Smokers Carrying Each Brand of Cigarettes at Time of Interview, by Sex and Age

Sex and age	Total	Brand X	Brand Y	Brand Z	Other
Total all ages	560	292	208	42	18
Total men	365	189	138	27	11
Total women	195	103	70	15	7
Under 30 years	243	88	122	24	9
Men	157	67	72	14	4
Women	86	21	50	10	5
30 to 49 years	212	136	57	15	4
Men	146	84	46	12	4
Women	66	52	11	3	0
50 years and over	105	68	29	3	5
Men	62	38	20	1	3
Women	43	30	9	2	2

that the table reports smokers. You don't know how many non-smokers were interviewed.

2. How many smokers interviewed carried brand X? How many men smokers carried brand X?
3. How many interviewed smokers were under 30 years of age? How many of these were women? How many women under 30 carried brand Z?
4. Does age appear to have an effect on preference? If so, what? In answering this question it may be helpful to refer to Table 6-5, in which the numbers carrying each brand have been reduced to percentages of the total for the age group. The first figure (percentage of those under 30 carrying brand X) is computed as $\frac{88}{243}$ (100) = 36 per cent. Other figures are found in similar manner. The table shows a clear preference for brand Y among young people and a preference for brand X among middle-aged and older smokers.

Table 6-5. Percentage of Age Group Carrying Each Brand of Cigarettes

Age	Total	Brand X	Brand Y	Brand Z	Other
Under 30	100	36%	50%	10%	4%
30 and under 50	100	64	27	7	2
50 and over	100	65	28	3	5

5. Does sex apear to affect preference? It will help to complete Table 6-6.

Table 6-6. Percentage Carrying Each Brand of Cigarettes, by Sex

Sex	Total	Brand X	Brand Y	Brand Z	Other
Male	100				
Female	100				

Although Tables 6-5 and 6-6 show differences in preference, it must be remembered that our conclusions are drawn from sample data. Therefore, our computed percentages are likely to vary somewhat from the true or population percentages. One function which the statistician can perform for management is to form a judgment about the significance of the observed differences in percentages. That is, he can form an objective judgment concerning the likelihood that the observed differences might have happened just by chance.

6-5. TESTS OF INDEPENDENCE IN CROSS CLASSIFICATIONS

In the previous section we relied on judgment to determine whether relationships exist among the variables of classification. The χ^2 distribution provides a rigorous test of the hypothesis that there is relationship, that is, that the two variables of classification are independent. The test is referred to as a *contingency-table* test.

In an earlier illustration we tested the hypothesis that absenteeism was unrelated to sex. Suppose, however, that we had kept a record of absences due to illness and those due to some other cause. We could then construct a two-way classification as follows:

Sex	Absences		
	Due to illness	Other	Total
Male	1,800	1,700	3,500
Female	1,200	600	1,800
Total	3,000	2,300	5,300

In order to apply the χ^2 test we must arrive at a set of theoretical frequencies based upon the hypothesis of independence. We compute the theoretical frequencies by reference to the marginal totals only. We reason this way: 3,000/5,300 of the total absences were due to illness. If sex and cause of absence are independent, then 3,000/5,300 of the total absences by males (3,500) should be due to illness. A similar proportion

of female absences (1,800) should also be due to illness. We compute

$$\frac{3,000}{5,300}(3,500) = 1,981$$

which is the number of male absences due to illness that we should expect if we knew only the marginal totals. Since rows and columns must add to the same totals as the observed frequencies, we can fill in the remaining theoretical frequencies as follows:

Sex	Absences		
	Due to illness	Other	Total
Male	1,981	1,519	3,500
Female	1,019	781	1,800
Total	3,000	2,300	5,300

We are now ready to proceed with the test.

1. H_0: Sex and cause of absenteeism are independent. H_1: They are related.
2. $\alpha = 0.05$.
3. $\chi^2 = \sum \frac{(|f_o - f_c| - \frac{1}{2})^2}{f_c}$ with 1 degree of freedom. Note that once we computed the figure 1,981, the other three values in the table were automatically specified. Therefore we have only 1 degree of freedom. In general, if we have a contingency table with R rows and C columns, the degrees of freedom are $(R - 1)(C - 1)$.
4. *Region of rejection:* $\chi^2 \geq 3.84$.
5.

f_o	f_c	$\lvert f_o - f_c\rvert - \frac{1}{2}$	$(\lvert f_o - f_c\rvert - \frac{1}{2})^2/f_c$
1,800	1,981	180.5	16.4
1,200	1,019	180.5	32.0
1,700	1,519	180.5	21.4
600	781	180.5	41.7
5,300	5,300		111.5

   $\chi^2 = 111.5$ with 1 degree of freedom.
6. Reject the hypothesis. A larger proportion of women's absences are due to illness.

The above procedure is quite general and can be applied to any size of two-way classification. Furthermore, there are no assumptions about the distribution of frequencies in the cells. We call this a *distribution-*

free or *nonparametric* test. The same caution applies as for the use of χ^2 in the previous section; that is, the theoretical frequencies should be as large as 5 in each cell.*

The contingency-table test can be applied to three-way classifications, but the interpretation of the dependence (if one rejects the hypothesis of independence) is more difficult.

EXERCISE 6-9. Refer to Table 6-3 and test the hypothesis that ownership of grocery stores is independent of size.

EXERCISE 6-10. Refer to Table 6-4. (*a*) Test the hypothesis that preference is independent of sex. (*b*) Test the hypothesis that preference is independent of age. (*c*) What effect would this information have on your advertising policy if you were producing brand X cigarettes?

6-6. NOTE ON NONPARAMETRIC TESTS

In Chap. 5 nearly all the tests of hypotheses depended for their validity upon certain assumptions about the population. The most common assumption was that the population was normally distributed, or at least that samples drawn from the population would have approximately normally distributed means.

In this chapter the assumptions about the population are less rigid. Nonparametric tests, that is, tests in which no assumptions (or at most very mild assumptions) are made about the parameters of the population, have more general application than parametric tests because of the fewer assumptions. Many such tests have been constructed to replace the tests given in Chap. 5. An examination of a major portion of these tests would be out of place here, since entire books have been devoted to the subject,† but a simple example will demonstrate the principle. Consider the test of the hypothesis $\mu_1 = \mu_2$ with dependent samples. Section 5-4 gives an illustration. If the two samples do, in fact, come from the same population, we should expect the differences in Table 5-1 to be approximately half positive and half negative. The signs of the differences form the basis for a nonparametric test. In the given illustration there were 2 positive and 23 negative signs. One can employ the binomial distribution to show that the hypothesis of equal probability of positive and negative signs must be rejected. The χ^2 distribution may be used also. For example:

1. H_0: $p = \frac{1}{2}$. H_1: $p < \frac{1}{2}$ or $p > \frac{1}{2}$.
2. $\alpha = 0.01$.

* For a method of testing for fewer frequencies, see R. A. Fisher, *Statistical Methods for Research Workers*, 11th ed., pp. 96–97, Hafner Publishing Company, New York, 1950.

† For example, Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Inc., New York, 1956.

3. $\chi^2 = \sum \frac{(|f_o - f_e| - \frac{1}{2})^2}{f_e}$ with 1 degree of freedom.
4. *Region of rejection:* $\chi^2 \geq 6.635$.
5.

Signs	f_o	f_e	$\lvert f_o - f_e\rvert - \frac{1}{2}$	$(\lvert f_o - f_e\rvert - \frac{1}{2})^2/f_e$
Positive	2	12.5	10	8
Negative	23	12.5	10	8
Total				16

$\chi^2 = 16$ with 1 degree of freedom.

6. Reject the hypothesis.

In case there are differences of 0, these pairs may be ignored and the total number of cases reduced accordingly.

Even though nonparametric tests have wider applicability, it does not necessarily follow that they are "better" than the customary parametric tests. It can be shown that, if one knows something about the distribution of the variable under consideration, it is generally possible to construct a test which is "more powerful" than a nonparametric test. It may be remembered, from Chap. 3, that the power of a test is $1 - \beta$, that is, the probability that an alternative hypothesis will be accepted when the alternative is, in fact, true. One test is more powerful than another if, for the same value of α, it has a greater value of $1 - \beta$.

Thus, in the previous illustration, if one knows that the differences D between the paired observations are normally distributed, he is better off using the $\bar{D}$ test (Sec. 5-5) than the sign test. The departure from normality can be quite severe, and the $\bar{D}$ test will still be superior because of the central limit theorem.

One other factor needs to be noted. Unless the differences between the paired observations are symmetrically distributed, a sign test actually tests the hypothesis that the *medians* are the same rather than the hypothesis that the means are the same. If the differences are symmetrically distributed, then, of course, the means and medians coincide.

EXERCISE 6-11. The sample for a market research survey is split at random into 3 parts, and 3 interviewers are assigned the 3 parts at random. The following results to a yes-no question are observed:

Interviewer	Number of responses	
	Yes	No
A	75	55
B	90	75
C	60	65

Is there any substantial evidence here of "interviewer bias," that is, evidence that responses are dependent somewhat upon the interviewer?

EXERCISE 6-12. A random sample of 200 customers of store A and 400 customers of store B showed that 57 per cent of A's customers had TV sets and that 64 per cent of B's customers had TV sets. Test the hypothesis that the true proportions are equal, using (*a*) the χ^2 test, (*b*) the normal test on proportions.

EXERCISE 6-13. A firm with 500 employees expects a turnover of about 8 per week. In a particular week chosen at random there were 14 resignations. Is this consistent with the belief that turnover averages 8 per week?

EXERCISE 6-14. A merchant, in buying ear corn, examines 10 ears out of a truckload. He adopts this decision rule: If 2 or more are wormy, reject the load; otherwise accept. What risk is he running of accepting a load that is 20 per cent wormy?

EXERCISE 6-15. A company carries 3 lines of a product and has 5 salesmen. A 3-month record shows the following units sold by each salesman:

Line	Salesman				
	1	2	3	4	5
A	20	8	0	20	16
B	17	16	5	20	15
C	25	11	6	18	5

What are your conclusions?

7. SIMPLE LINEAR REGRESSION AND CORRELATION

7-1. INTRODUCTION

Some methods for examining the relationship between two variables were presented in Chap. 6. These methods were based on classification, or "sorting." For example, if a large grocery chain wants to find out whether sales can be influenced by the amount of local radio advertising, it might conduct an experiment along the following lines. A study period of 1 month is chosen, and results are to be judged by the ratio of the given month's sales to sales for the corresponding month for the previous year. This technique is designed to eliminate differences in size of store or community. Communities are divided at random into three groups. The first group will spend nothing on local radio advertising. The second group will spend $\frac{1}{2}$ of 1 per cent of gross sales (for the same month the previous year). The third group will spend 1 per cent of gross sales. There is to be no change in amount of newspaper and other advertising. Results are to be classified into three groups by the ratio of sales for the given month to sales for the corresponding month a year ago. The three classes are (1) those with ratios less than 1, (2) those with ratios from 1.00 to 1.10, and (3) those with ratios over 1.10. Assume that the results are as given in Table 7-1.

The table shows that an increase in sales appears to be related to an increase in advertising expenditure, but we wonder whether the results are statistically significant. The χ^2 test of contingency given in Chap. 6 may be used to test the hypothesis of independence, that is, the hypothesis that advertising expenditure and sales are unrelated. Rejection of this hypothesis implies existence of some relationship. To determine the nature of the relationship one must refer again to Table 7-1.

Table 7-1. Number of Communities Classified by Gain (or Loss) in Sales and Radio Advertising Expenditure for August, 1959

Ratio of sales in August, 1959, to August, 1958	Radio advertising expenditure			
	None	0.5 per cent of sales	1.0 per cent of sales	Total
Over 1.10	3	6	15	24
1.00 to 1.10	11	14	6	31
Under 1.00	11	6	5	22
Total	25	26	26	77

EXERCISE 7-1. Apply the χ^2 test to Table 7-1 and test for independence.

EXERCISE 7-2. (*a*) Would a generally increasing level of business activity affect the validity of the results? Why or why not? (*b*) Would an increasing or decreasing price level have any influence?

The difficulty with the contingency-table analysis is that it provides no quantitative measure of the effect of advertising. Suppose we consider another method of analysis. In this analysis we shall compute the average sales ratio for each level of advertising and the variance around that average. Suppose the results are as follows:

Advertising expenditure, per cent of sales	Number of observations	Average sales ratio	Variance
0	25	1.04	0.0065
0.5	26	1.07	0.0088
1.0	26	1.13	0.0145

Now, we can test the hypothesis that the average sales ratio is not affected by advertising, or we can place confidence limits on the amount of increase. This information, being quantitative, permits one to evaluate the advisability of utilizing local radio advertising.

EXERCISE 7-3. Place 90 per cent confidence limits on the difference in sales ratios (*a*) between 0 and 0.5 per cent expenditure, (*b*) between 0 and 1.0 per cent expenditure.

A technique for testing the hypothesis that all three means are alike, called the *analysis of variance*, is presented in Chap. 8.

Notice that in computing the means and either testing the hypothesis of no difference or placing confidence intervals we had to introduce some assumptions to validate the techniques. Our assumptions were that the average ratios were normally distributed (or nearly so) and that the variances were equal.

EXERCISE 7-4. Do you feel that these assumptions are valid? Why or why not?

Thus we see that increasingly useful methods require increasingly rigid assumptions about the data. This situation is not unusual in statistics.

If we are willing to make further restrictive assumptions about the nature of the relationship between sales and advertising expenditure, we can employ even more useful techniques, and it is these techniques which form the subject matter of this chapter.

Bivariate analysis, that is, the simultaneous analysis of two variables, may take one of two forms. First, there may be a natural pairing of variables, such as height and weight, and we may be interested in the joint relationship of the two variables. In the above illustration, for example, we might have observed the relationship between sales and advertising without having performed an experiment. If we draw a random sample of communities (or stores), we have, simultaneously, a random distribution of sales ratios and a random distribution of advertising expenditures. The two distributions are not necessarily independent, however. An analysis of the joint data may reveal something about the relationship between the two variables. It should be pointed out that this type of analysis, usually referred to as *correlation analysis*, is fraught with all kinds of dangers. Even though a relationship exists, there is no guarantee that it is a cause-and-effect relationship, and, in general, the technique of "observation" is a poor substitute for experimentation.

The second situation in bivariate analysis arises when one variable is selected arbitrarily and the results in terms of the other variable are observed. This is the experimental situation described above. As another example, we may vary the temperature in an industrial plant and observe the effect on production. In this case the temperature can be controlled, within reason, and hence is not a random variable. Production may depend somewhat on temperature, but each observation on production has a random element associated with it. This random element is not associated with the other variable (temperature). It is this type of situation, in which all the variation is associated with one of the variables, that lends itself to bivariate *regression analysis.*

7-2. THE LINE OF REGRESSION

Regression analysis is designed to examine the relationship of a variable Y to a variable X. In this text we shall consider the case in which such relationship is linear, that is, a relationship which can be expressed as a straight line.

The X variable is commonly called the "independent" variable. It

is the variable which is assumed to have no error associated with it, or the variable selected arbitrarily for the purpose of studying Y. The Y variable is called the "dependent" variable and contains a random element.

We assume that a relationship exists between X and Y in the universe such that each Y value in the universe can be expressed by the equation

$$Y_i = A + BX_i + \epsilon_i \tag{7-1}$$

where A and B are parameters of the population and ϵ is a random variable (independent of X and Y) with mean 0. Further, as indicated above, it is assumed that X is measured without error, or that the X values are selected arbitrarily by the researcher. These assumptions specify that all the variation is in Y, a condition which is somewhat unrealistic in many instances but which is necessary for the simplification of the methodology.

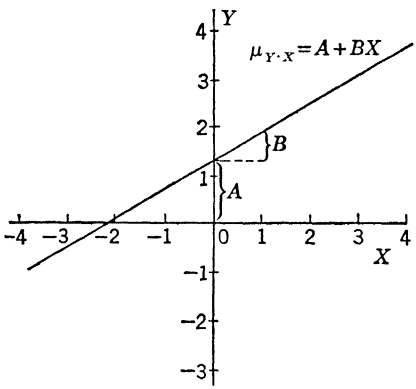

FIG. 7-1. Nature of the constants A and B in linear regression

The nature of the constants A and B is illustrated in Fig. 7-1. The equation $\mu_{Y\cdot X} = A + BX$ represents the line of average relationship between X and Y in the population. It is called the *line of regression of Y on X*. The parameter A is the Y *intercept*, or that value of Y where $X = 0$. The parameter B is the *slope* of the line of regression, or the amount the line rises for each unit increase in X. The line may, of course, slope downward to the right, in which case the value of B is negative.

In practice, the statistician never encounters a situation in which he can predict the exact value of Y from knowledge of X. Theoretically, however, one can form such distributions. For example, if one knows the radii of circles, he can state their circumferences exactly. There is no sampling error (or variation) in such a case, and it holds no interest for the statistician.

As a more typical illustration of the problem facing the statistical analyst, let us consider a problem in which we wish to predict the number of defective items produced by a machine run at varying speeds. We make the following assumptions:

1. Machine speed can be measured with little error.
2. For the machine under observation, the number of defectives is linearly related to speed.

3. The distribution of the residual ϵ is normal about the population line of regression, and its dispersion is constant for any value of X.

Another way of expressing these assumptions is to say that each observation Y_i can be stated in the following form:

$$Y_i = A + BX_i + \epsilon_i \tag{7-1}$$

where X_i (speed) is known, A and B are unknown parameters, and ϵ_i is a random variable such that its average value is 0, its variance is $\sigma^2_{Y \cdot X}$ for every value of X, and getting a particular error on one observation does not affect the value of the error on the next (or any other) observation.

We make the observations of Table 7-2 for a particular machine over a short interval of time. In order to reveal the nature of the relationship, we prepare the scatter diagram shown in Fig. 7-2. (The line drawn on the diagram will be explained later.) There appears to be some relationship between speed and defectives. Furthermore, the relationship appears to be linear.

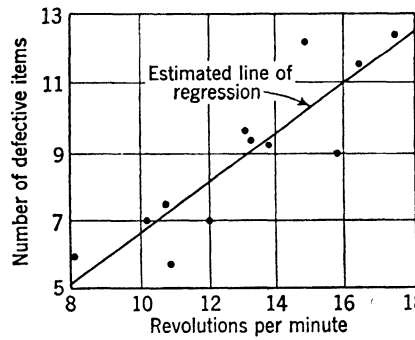

FIG. 7-2. Number of defectives at various machine speeds

Our objective is to estimate the line of regression for these data, that is, to estimate the values of A and B. Then, using the estimates a and b, each observation can be expressed as

$$Y_i = a + bX_i + e_i \tag{7-2}$$

Table 7-2. Speed of Machine and Number of Defective Items Produced per Hour

Speed, rpm	*Number of defective items*	*Speed, rpm*	*Number of defective items*
16.4	11.4	8.1	6.0
10.2	7.0	13.8	9.2
13.1	9.6	12.0	7.0
15.8	9.0	10.8	7.5
10.9	5.7	17.4	12.3
13.2	9.4	14.9	12.2

where the e_i are now deviations from a fitted (or approximate) line, that is, $e_i = Y_i - (a + bX_i)$. In this respect, the e_i correspond to $(Y_i - \bar{Y})$ in the univariate case. It should be noted that the random variable e is not the same variable as ϵ. The variable ϵ is a deviation from a population (or true) line of regression. The variable e is a devia-

tion from an estimated (or fitted) line of regression. The e_i may be considered to be composed of two parts: (1) a deviation of the fitted line from the true line, and (2) a deviation of the sample point (X_i, Y_i) from the true line.

Under the assumption that the relationship is linear, we proceed to fit a straight line to the scatter of points. Many methods could be employed for fitting the straight line, but we shall use the *method of least squares*, which will assure us that the sum of the squares of the vertical deviations from the fitted line will be less than the sum of the squares of the vertical deviations from any other line, no matter how computed. "Vertical" is to be understood to mean vertical with respect to the X axis. The least-squares line is found by solving simultaneously the following two equations, called *normal equations*, for the constants a and b:

$$\begin{aligned} \text{I.} \quad & na + b\Sigma X = \Sigma Y \\ \text{II.} \quad & a\Sigma X + b\Sigma X^2 = \Sigma XY \end{aligned} \tag{7-3}$$

We prepare a computation table (Table 7-3) to accumulate the sums required for the solution of the normal equations. The Y^2 column is not required immediately, but has been computed here for further reference.

Returning to the normal equations, we substitute the sums accumulated in the computation table:

Table 7-3. Computation Table for Regression Equation

X	Y	XY	X^2	Y^2
16.4	11.4	186 96	268.96	129 96
10.2	7.0	71.40	104.04	49.00
13 1	9.6	125 76	171.61	92.16
15.8	9.0	142.20	249.64	81.00
10.9	5.7	62.13	118.81	32.49
13.2	9.4	124.08	174.24	88 36
8.1	6.0	48.60	65.61	36.00
13.8	9.2	126.96	190.44	84 64
12.0	7.0	84.00	144 00	49.00
10 8	7.5	81.00	116.64	56.25
17.4	12.3	214.02	302.76	151.29
14.9	12.2	181.78	222.01	148.84
156.6	106.3	1,448.89	2,128.76	998.99

$$\begin{aligned} \text{I.} \quad & 12a + 156.6b = 106.3 \\ \text{II.} \quad & 156.6a + 2{,}128.76b = 1{,}448.89 \end{aligned}$$

We may eliminate a first and then solve for b. We may do this by multiplying the first equation by 156.6 and the second by 12 and subtracting.

$$\text{II.} \quad 1{,}879.2a + 25{,}545.12b = 17{,}386.68$$
$$\text{I.} \quad \underline{1{,}879.2a + 24{,}523.56b = 16{,}646.58}$$
$$1{,}021.56b = 740.10$$
$$b = \frac{740.10}{1{,}021.56} = 0.7245$$

Substituting this value of b into equation I, we have

$$106.3 = 12a + 156.6\ (0.7245)$$
$$12a = -7.1567$$
$$a = -0.60\dagger$$

Our estimate of the population line of regression is

$$\bar{Y}_X = -0.60 + 0.7245X$$

That is, our estimate of A is -0.60 and our estimate of B is 0.7245. The estimate of A has little significance in this particular case, but the *regression coefficient* b is of considerable interest. It represents an estimate of the average amount that the number of defectives is increased by each additional rpm. That is, within the range of speeds studied, each additional rpm increases the defectives by 0.72 unit per hour on the average.

It is important to note that this measure is applicable only to the range of speeds for which data are available. With an average speed of, say, 24 rpm, we have no information to indicate the expected number of defectives. We cannot assume that the linear relationship revealed by these data can be extended to include $X = 24$ rpm.

The line of regression can be used for prediction of Y values from known values of X. For example, suppose we know the speed for a given hour to be 14 rpm (a value which is well within the range of observed data). Our best estimate of the number of defectives is found by inserting $X = 14$ in the equation for the line of regression:

$$\hat{Y}_{14} = -0.60 + 0.7245(14)$$
$$= 9.54 \text{ defective items}$$

This is a "best estimate" in a sense to be defined later.

EXERCISE 7-5. A company produces small metal parts from compressed metal powder. One way of testing the product is to examine the parts for breaking strength. This destroys the product. It is believed that there may be some

† The constants a and b may be obtained directly by the following relationship:

$$b = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \tag{7-4}$$
$$a = \bar{Y} - b\bar{X} \tag{7-5}$$

These solutions may be obtained easily by solving (7-3) for a and b.

relationship between density (as measured by weight) and breaking strength. The following observations are made:

Weight, g	*Breaking strength, lb*
85.63	531
85.47	505
85.13	424
85.24	428
85.96	608
85.84	564
85.48	496
85.39	465
85.52	516
85.64	529

(*a*) Plot the data on a scatter diagram. Do you think there is a linear relationship between breaking strength and weight? (*b*) Find the line of regression. Note that your computations will be much simpler if you subtract 85 from each weight and 400 from each breaking strength. (*c*) If a part weighs 85.50 g, what is your estimate of its breaking strength? Remember to code the variables (subtract 85 from X and 400 from Y) before inserting into the estimating equation. (*d*) Do you think the relationship would remain linear for all values of X? Why? (*e*) If someone asked you to predict the breaking strength of a part that weighed 84.5 g, what answer should you give him? (*f*) Coding the variables by subtracting 85 from X and 400 from Y simplifies computations, but it requires that 85 be subtracted from each weight before the regression equation is used and that 400 be added to each predicted Y value. We can convert back to units of the original data by solving the following equation for $\bar{Y}_X$:

$$\bar{Y}_X - 400 = a + b(X - 85)$$

Do this and then repeat part (*c*).

7-3. NOTE ON THE METHOD OF LEAST SQUARES★

The method given above for fitting a straight line to a scatter of points is not the only method that might have been used. One might have drawn a freehand line through the points; then, from the graph, one could read off two points on the line, and from them he could solve for the equation of the line. Another method might be to divide the points into two groups by a vertical line and then to fit a straight line through the averages of the X's and Y's of each group. Many other methods might conceivably have been used. The student may wonder why the method of least squares was chosen.

A remarkable theorem in mathematical statistics, known as the

★ Denotes a section which can be omitted by the nonmathematical student.

Markoff theorem,* states that for the given conditions the least-squares line is the line of "best" fit in a well-defined sense. "Best" is used here to mean that the estimates of the constants A and B are the best linear unbiased estimates of those constants. An estimator is "unbiased" if its average value (or expected value) is equal to the population parameter it estimates. That is, if a is an estimate of A, and b is an estimate of B, a and b are unbiased if the average value of $a = A$ and the average value of $b = B$.

An estimate is a "linear" estimate if it is linear in the observations, that is, if it can be expressed as a sum of observed values (Y's) times constants. For example, $\bar{Y}$ is a linear estimate of μ since it can be expressed as a linear function of the Y's.

$$\bar{Y} = \frac{1}{n} Y_1 + \frac{1}{n} Y_2 + \cdots + \frac{1}{n} Y_n$$

We see that b is a linear estimate since it can be expressed as follows:

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma X^2 - n\bar{X}^2} = \frac{\Sigma(X - \bar{X})Y}{\Sigma X^2 - n\bar{X}^2}$$

$$= \frac{1}{\Sigma X^2 - n\bar{X}^2} [(X_1 - \bar{X})Y_1 + (X_2 - \bar{X})Y_2 + \cdots + (X_n - \bar{X})Y_n]$$

This is a linear function of the Y_i, since it is assumed that the X_i are known. The form of the estimator to be presented in the next section makes it easier to see that b is linear in the observations.

The constant a is linear in the observations, since it can be expressed as

$$a = \bar{Y} - b\bar{X}$$

and $\bar{Y}$ and b are both linear functions of the observations (Y's).

An estimate of a parameter is said to be the "best" linear unbiased estimator if its variance is less than the variance of any other linear unbiased estimator. Hence the variances of a and b are less than the variances of any other linear unbiased estimators of A and B.

On the basis of these definitions, the Markoff theorem can now be stated in the following form.

Consider a variable Y_i ($i = 1, 2, \ldots, n$) which can be expressed as a linear function of p unknown parameters ($b_1, b_2, \ldots, b_p$) with known coefficients ($X_{1i}, X_{2i}, \ldots, X_{pi}$) and a random variable ϵ_i; that is:

$$Y_i = b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi} + \epsilon_i \qquad (7\text{-}6)$$

* A. Markoff, *Wahrscheinlichkeitsrechnung*, Leipzig and Berlin, 1912. A statement of the theorem in English (with some extension) can be found in F. N. David, *Probability Theory for Statistical Methods*, pp. 161ff., Cambridge University Press, New York, 1949.

If the ϵ_i are such that

$$\begin{aligned} E\epsilon_i &= 0 \qquad \text{for all } i \\ E\epsilon_i^2 &= \sigma^2_{Y\cdot X} \qquad \text{(a constant)} \qquad \text{for all } i \\ E\epsilon_i\epsilon_j &= 0 \qquad \text{for all } i \neq j \end{aligned}$$

where E stands for expected, or average, value, then the best linear unbiased estimate of any linear function of the parameters $b_1, b_2, \ldots, b_p$ is the same linear function of the b's that minimize the function

$$Q = \Sigma\epsilon_i^2 = \Sigma(Y_i - \hat{b}_1X_{1i} - \hat{b}_2X_{2i} - \cdots - \hat{b}_pX_{pi})^2$$

That is, solutions are found which make this sum of squares a minimum. This result leads to the name "the method of least squares." It is assumed, of course, that solutions exist. The theorem provides further that an unbiased estimate of the variance $\sigma^2_{Y\cdot X}$ is given by

$$s^2_{Y\cdot X} = \frac{1}{n-p}\Sigma(Y_i - \hat{b}_1X_{1i} - \hat{b}_2X_{2i} - \cdots - \hat{b}_pX_{pi})^2 \qquad (7\text{-}7)$$

where $\hat{b}_i$ denotes the least-squares estimate of b_i.

This theorem will be referred to frequently in the material which follows. In order to fit the present problem into the framework of the Markoff theorem, let

$$\begin{aligned} X_{1i} &= 1 \qquad \text{(a constant for all } i\text{)} \\ X_{2i} &= X_i \qquad \text{(assumed to be known for all } i\text{)} \\ b_1 &= A \\ b_2 &= B \end{aligned}$$

All other X's and b's $= 0$.

Then,

$$Y_i = A + BX_i + \epsilon_i$$

This is the linear-regression model stated in Eq. (7-1).

According to the Markoff theorem the best linear unbiased estimates of A and B will be found by minimizing the quantity

$$Q = \Sigma(Y_i - a - bX_i)^2$$

with respect to a and b. This can be done easily by the differential calculus. The procedure is presented below for those whose background includes calculus:

$$\begin{aligned} \frac{\partial Q}{\partial a} &= -2\Sigma(Y_i - a - bX_i) = 0 \\ na + b\Sigma X_i &= \Sigma Y_i \\ \frac{\partial Q}{\partial b} &= -2\Sigma X_i(Y_i - a - bX_i) = 0 \\ a\Sigma X_i + b\Sigma X_i^2 &= \Sigma X_iY_i \end{aligned}$$

7-4. SOME SIMPLIFIED NOTATION

Certain functions of the sample values are used over and over again in regression and correlation analysis. Their frequency of occurrence makes it advisable to devise specialized symbols for them. Consider the following:

$$\begin{aligned} G_{xx} &= n\Sigma X^2 - (\Sigma X)^2 \\ G_{yy} &= n\Sigma Y^2 - (\Sigma Y)^2 \\ G_{xy} &= n\Sigma XY - \Sigma X\Sigma Y \end{aligned} \tag{7-8}$$

Using these forms, we can express the formula for b as

$$b = \frac{G_{xy}}{G_{xx}} \tag{7-9}$$

The simplification in notation really pays dividends when we consider more than two variables (as we do in Chap. 10).

An important feature of the G's is that they can be computed easily on modern calculating machines. Such machines will permit the accumulation of sums of squares and cross products such as ΣX^2, ΣY^2, and ΣXY. Many will permit one to accumulate ΣX and ΣY at the same time. After computation of these quantities one can compute the G's without having to copy down intermediate results. For example, to compute G_{xy}, multiply n by ΣXY and leave the result in the machine. Then use negative multiplication (or the minus bar with a nonautomatic machine) to subtract the product $\Sigma X\Sigma Y$. Generally speaking, one will avoid errors by utilizing the machine in such a way that as few figures as possible have to be recorded.

Some modern machines will permit the simultaneous accumulation of ΣX, ΣY, ΣX^2, and $2\Sigma XY$ if none of the X or Y values exceeds two digits in size. Often, the data can be coded by subtraction of a constant so that the computations can be handled in this manner. Coding by subtraction has no effect on the constant b. If the true means are substituted into the formula $a = \bar{Y} - b\bar{X}$, the correct value of a is also obtained.

The student will do well to familiarize himself with the characteristics of the machine with which he will be working. Reference to the manufacturer's manual of operating instructions will prove helpful.

As another illustration of regression using the abbreviated notation, consider the data of Table 7-4 constructed by drawing pairs of numbers at random from a table of random numbers.

Since X and Y were paired at random, one would expect to find no relationship between X and Y, that is, a regression coefficient of 0. The fact that $b = 0.4393$ illustrates an important sampling principle, namely, that one occasionally will find what appears to be a relationship when in fact none exists.

As another illustration of the same principle, suppose a population is composed of 32 pairs of observations X, Y. For example, suppose that we have 32 seeds and that we measure the length and hardness of each

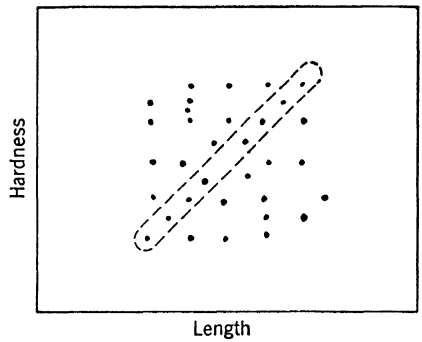

FIG. 7-3. Hypothetical pairing of length and hardness of seeds

seed. Then each pair of measurements (length and hardness) is plotted as a point in Fig. 7-3.

Table 7-4. Variables Paired by Drawing from a Table of Random Digits

X	Y	X	Y	X	Y
74	23	89	67	93	86
29	33	45	2	66	13
47	82	31	3	23	88
92	96	10	23	83	37

$$\Sigma X = 682 \qquad \Sigma Y = 553$$
$$\Sigma X^2 = 48{,}420 \qquad \Sigma Y^2 = 39{,}267$$
$$\Sigma XY = 35{,}672$$
$$G_{xx} = 115{,}916$$
$$G_{yy} = 165{,}395$$
$$G_{xy} = 50{,}918$$
$$b = \frac{G_{xy}}{G_{xx}} = \frac{50{,}918}{115{,}916} = 0.4393$$
$$a = \bar{Y} - b\bar{X} = \frac{553 - 0.4393(682)}{12}$$
$$= 21.1$$
$$\bar{Y}_X = 21.1 + 0.4393X$$

EXERCISE 7-6. Verify the computations of Table 7-4.

It is clear from the scatter diagram that length and hardness are unrelated. That is, knowledge of the length of a seed contributes no information about its hardness. Suppose, however, that we draw a sample of 8 from the population of 32 seeds and find that our sample of

8 consists of the points enclosed by the dotted lines. A line of regression fitted to these points will show perfect relationship between Y and X. The probability of obtaining these particular 8 points is approximately 0.000000095,* which, perhaps, is small enough to be ignored. However, the probability of obtaining 8 points which show some relationship between Y and X is substantial. This illustration shows the necessity for judging the significance of regression in the light of probabilities. This matter is elaborated upon in the next section.

7-5. TESTS OF HYPOTHESES AND CONFIDENCE INTERVALS IN REGRESSION

Let $\sigma^2_{Y \cdot X}$ denote the variance of the ϵ_i, that is, the variance of the observations around the true line of regression. Then, according to the extended Markoff theorem (Sec. 7-3), an unbiased estimate of σ^2 is provided by

$$s^2_{Y \cdot X} = \frac{1}{n-2} \Sigma(Y_i - a - bX_i)^2 \tag{7-10}$$

$$= \frac{1}{n-2} (\Sigma Y_i^2 + na^2 + b^2 \Sigma X_i^2 - 2b\Sigma X_i Y_i + 2ab\Sigma X_i - 2a\Sigma Y_i)$$

$$= \frac{1}{n-2} [\Sigma Y_i^2 + a(na + b\Sigma X_i) + b(a\Sigma X_i + b\Sigma X_i^2) - 2a\Sigma Y_i - 2b\Sigma X_i Y_i]$$

But the quantities in parentheses are the left-hand sides of the normal equations. Substituting

$$\text{I.} \qquad na + b\Sigma X_i = \Sigma Y_i$$

$$\text{II.} \qquad a\Sigma X_i + b\Sigma X_i^2 = \Sigma X_i Y_i$$

one obtains

$$s^2_{Y \cdot X} = \frac{1}{n-2} (\Sigma Y_i^2 - a\Sigma Y_i - b\Sigma X_i Y_i) \tag{7-11}$$

These quantities are readily available from the computation table and the calculations already completed.

An alternative form, using the G notation, is

$$s^2_{Y \cdot X} = \frac{1}{n(n-2)} \left[G_{yy} - \frac{(G_{xy})^2}{G_{xx}} \right]$$

$$= \frac{G_{xx}G_{yy} - (G_{xy})^2}{n(n-2)G_{xx}} \tag{7-12}$$

$\frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{32 \cdot 31 \cdot 30 \cdot 29 \cdot 28 \cdot 27 \cdot 26 \cdot 25}$.

Using this latter form with the illustrative data on temperature and defects, we have

$$G_{xx} = 12(2{,}128.76) - (156.6)^2 = 1{,}021.56$$
$$G_{yy} = 12(998.99) - (106.3)^2 = 688.19$$
$$G_{xy} = 12(1{,}448.89) - 156.6(106.3) = 740.10$$

Hence
$$s^2_{Y \cdot X} = \frac{1{,}021.56(688.19) - (740.10)^2}{12(10)(1{,}021.56)} = 1.267$$

The square root of the above quantity, $s_{Y \cdot X}$, is sometimes called the *standard error of estimate.* The term *standard deviation from regression* would seem to be more meaningful.

So far, nothing has been specified concerning the distribution of the ϵ_i, that is, the residuals from the true regression line. If these residuals are normally distributed, then b is normally distributed with mean B and variance

$$\frac{n\sigma^2_{Y \cdot X}}{G_{xx}}$$

That is,

$$Eb = B \tag{7-13}$$

$$E(b - B)^2 = \sigma_b{}^2 = \frac{n\sigma^2_{Y \cdot X}}{G_{xx}} \tag{7-14}$$

where E stands for expected, or average, value. To test the hypothesis that $B = B_0$, one can use the t ratio:

$$t = \frac{b - B_0}{s_b} \qquad \text{with } n - 2 \text{ } df \tag{7-15}$$

where
$$s_b = \sqrt{\frac{ns^2_{Y \cdot X}}{G_{xx}}} \tag{7-16}$$

Confidence limits for the regression coefficient B are given by

$$b - ts_b < B < b + ts_b \tag{7-17}$$

where s_b is defined by (7-16) and t is read from a t table with the appropriate probability level and with $n - 2$ degrees of freedom. In the example given above, the 0.10 level of t with 10 degrees of freedom is 1.81, and $s_b = 0.124$, so that 90 per cent confidence limits on B are given by

$$0.7245 - 1.81(0.124) < B < 0.7245 + 1.81(0.124)$$
$$0.50 < B < 0.95$$

Therefore, with 90 per cent confidence, we assert that each degree of increase in temperature results in from 0.50 to 0.95 additional defective units per hour.

We may be interested in setting confidence limits or testing hypotheses about the average population value of Y for a fixed value of X. Such an estimate involves the errors in estimating both a and b. We need the following standard deviation, which might be called the standard error of the average value of Y for a given X.

$$s_{\bar{Y}_X} = s_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{n(X - \bar{X})^2}{G_{xx}}} \tag{7-18}$$

Then the ratio

$$t = \frac{\bar{Y}_X - \mu_{Y \cdot X}}{s_{\bar{Y}_X}} \tag{7-19}$$

is distributed as t with $n - 2$ degrees of freedom.

For example, suppose we wish the 95 per cent confidence-interval estimate of $\mu_{Y \cdot X}$ for $X = 10$ in the speed-versus-defectives illustration. We compute

$$\bar{X} = \frac{\Sigma X}{n} = \frac{156.6}{12} = 13.05$$

$$s_{\bar{Y}_X} = 1.13 \sqrt{\frac{1}{12} + \frac{12(10 - 13.05)^2}{1{,}021.56}} = 0.50$$

$$\bar{Y}_X = -0.60 + 0.7245(10) = 6.645$$

$$P_c[\bar{Y}_X - t_{(0.05,10)} s_{\bar{Y}_X} < \mu_{Y \cdot 10} < \bar{Y}_X + t_{(0.05,10)} s_{\bar{Y}_X}] = 0.95$$

$$6.645 - 2.23(0.50) < \mu_{Y \cdot 10} < 6.645 + 2.23(0.50)$$

$$5.53 < \mu_{Y \cdot 10} < 7.76$$

This confidence interval is narrowest where $X = \bar{X}$, because then $s_{\bar{Y}_X} = s_{Y \cdot X} \sqrt{1/n}$. When n is small, as in this illustration, the confidence interval becomes wide rapidly as the value of X departs from $\bar{X}$. We say then, with 90 per cent confidence, that when the coded temperature equals 10, the true average number of defects per hour is between 5.53 and 7.76.

One situation remains to be discussed. Suppose we wish to place confidence limits on the expected variation in Y (not $\bar{Y}_X$) for a given value of X. We compute a standard error of Y for a given X by the following formula:

$$s_{\hat{Y}} = s_{Y \cdot X} \sqrt{1 + \frac{1}{n} + \frac{n(X - \bar{X})^2}{G_{xx}}} \tag{7-20}$$

Exercise 7-7. Refer to Exercise 7-5. (*a*) Find the standard error of estimate. (*b*) Place 90 per cent confidence limits on the true breaking strength for all parts having a weight of 85.5 g. (*c*) What assumptions were made in (*b*)? Do you think they are justified?

Exercise 7-8. In order to establish a workable standard for accepting or rejecting parts on the basis of weight alone, one would want to examine more

than 10 pieces. Suppose that an examination of 400 pieces yields a regression equation of

$$\bar{Y}_X = -16{,}600 + 200X$$

with $s_{Y \cdot X} = 10$. Suppose also that the observations appear to be distributed normally around the line of regression. If a breaking strength of less than 450 lb is considered unsatisfactory, what is the approximate probability that the breaking strength of a part is less than 450 lb if it weighs (*a*) 85.3 g, (*b*) 85.2 g, (*c*) 85.5 g?

7-6. NATURE OF THE CORRELATION COEFFICIENT

Previous sections were devoted to the analysis of situations in which it is desired that one variable, Y, be predicted or estimated from knowledge of another variable, X. The value of Y is assumed to be dependent in some degree upon X. The greater the dependence, the smaller the variation around the line of regression. This principle serves as the basis for tests of hypotheses concerning the regression coefficients.

We now give our attention to a situation in which neither variable need be considered independent. We are interested now in some measure of the relationship between two variables, X and Y, which does not require that they be designated independent and dependent. Such designation may be made as a matter of convenience, but the choice must not influence the results. Certainly, no cause-and-effect relationship can be implied from the computations alone. One such measure is the *coefficient of correlation*.

Let us consider a bivariate normal population, that is, one in which the X variable is normally distributed, with mean μ_X and standard deviation σ_X, and the Y variable is also normally distributed with mean μ_Y and standard deviation σ_Y. The two distributions need not be independent, however. That is, high values of one variable may be associated with high values of the other variable, or low values of one with high values of the other. The mathematical expression for such a distribution is

$$f(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(X-\mu_X)^2}{\sigma_X{}^2} - \frac{2\rho(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(Y-\mu_Y)^2}{\sigma_Y{}^2}\right]\right\} \quad (7\text{-}21)$$

Note that the function contains five parameters: μ_X, μ_Y, σ_X, σ_Y, and ρ. The fifth parameter, ρ, is the coefficient of correlation for the bivariate normal population. Its range is -1 to $+1$.

Some brief comment may be in order concerning the properties of the bivariate normal distribution, although the beginning student need not be expected to manipulate the imposing formula given above. It is

easiest to think of the bivariate normal distribution as a three-dimensional surface, which might appear as shown in Fig. 7-4.

This particular bivariate normal surface has some correlation. That is, the long axis of the mound-shaped surface is not parallel to either the X or the Y axis. If we take any cross section of the mound-shaped solid parallel to the XY plane, the cross section will be an ellipse. As the correlation approaches either $+1$ or -1, the cross section will approach a straight line. The diagrams in Fig. 7-5 illustrate the concept.

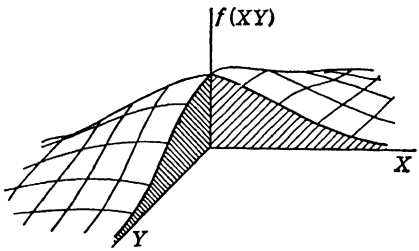

FIG. 7-4. Bivariate normal distribution

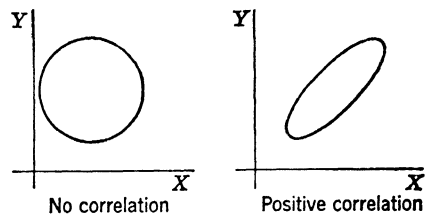

FIG. 7-5. Ellipses from correlation surfaces

If we take a vertical slice through the solid, either parallel to the X axis or parallel to the Y axis, the slice through the bivariate normal solid will be a normal distribution (see cutaway section of the correlation solid). In fact, a vertical slice in any direction will expose a normal cross section.

7-7. ESTIMATION OF THE CORRELATION COEFFICIENT

Let us consider the case in which a sample is drawn from a bivariate normal distribution. Note that a single drawing results in two sample values, an X value and a Y value. The X value is a random observation from the X population and the Y value is a random observation from the Y population, but the two are not necessarily independent. We wish to compute r as an estimator for ρ.

The estimated coefficient of correlation, with which we shall be concerned, is defined by

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} \tag{7-22}$$

$$r = \frac{G_{xy}}{\sqrt{G_{xx}G_{yy}}} \tag{7-23}$$

This coefficient may vary from -1 through 0 to $+1$. Perfect positive correlation, that is, where an increase in X determines exactly an increase in Y, yields a coefficient of $+1$, and perfect negative correlation, where an increase in X determines exactly (or is determined exactly by) a decrease

in Y, yields a coefficient of -1. This measure, computed from sample data, is an estimate of the population parameter ρ, and, as we shall see subsequently, part of the job of the statistician is to use r to test hypotheses about ρ.

For illustration, let us consider scores made by 20 students on the mid-term and final examinations in elementary statistics (Table 7-5).

We need the following intermediate computations:

$$G_{xx} = 20(44{,}900) - (920)^2 = 51{,}600$$
$$G_{yy} = 20(118{,}850) - (1{,}500)^2 = 127{,}000$$
$$G_{xy} = 20(71{,}600) - 920(1{,}500) = 52{,}000$$

Then, by use of (7-23), the correlation coefficient is computed as follows:

$$r = \frac{52{,}000}{\sqrt{51{,}600(127{,}000)}} = 0.642$$

7-8. TESTS OF HYPOTHESES CONCERNING CORRELATION COEFFICIENTS

Like any other statistic computed from sample data, r is expected to vary from its parameter ρ. The distribution of r is not symmetric, however. The limits on the variation of r (-1 and $+1$) cause the distribution of r to be skewed to the left when ρ is near $+1$ and to be skewed to the right when ρ is near -1. When $\rho = 0$, the distribution of r is symmetric, although not quite normal. Figure 7-6 shows the shape of the distribution of r for $\rho = 0.8$ and $\rho = 0$.

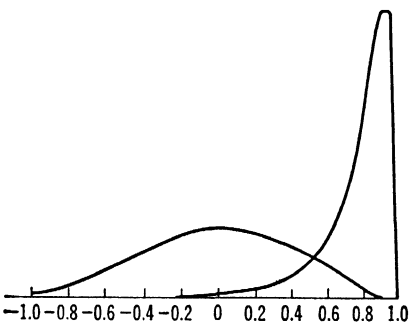

FIG. 7-6. Distribution of the correlation coefficient for $\rho = 0.8$ and $\rho = 0$

The symmetrical shape of the distribution when $\rho = 0$ makes it possible to test the hypothesis that $\rho = 0$ by the use of the t distribution:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{with } n-2 \text{ } df \tag{7-24}$$

In our illustration of mid-term and final examination scores,

$$t = \frac{0.642\sqrt{18}}{\sqrt{1-0.4112}} = \frac{2.72}{0.767} = 3.55$$

Referring to a table of t with $n - 2 = 18$ degrees of freedom, we see clearly that we should not expect to draw a sample at random having a

correlation coefficient of 0.642 from a population with $\rho = 0$. Hence, we conclude that there is correlation between X and Y in the population from which the sample was drawn.

Note that this test can only be used against the hypothesis that $\rho = 0$. Hence, if this hypothesis is rejected, we simply accept the hypothesis that there is *some* relationship. It reveals nothing about the extent of the relationship.

Table 7-5. Mid-term and Final Examination Scores in Elementary Statistics

Student	Mid-term, X	Final, Y	XY	X^2	Y^2
1	35	60	2,100	1,225	3,600
2	60	80	4,800	3,500	6,400
3	55	60	3,300	3,025	3,600
4	35	80	2,800	1,225	6,400
5	35	75	2,625	1,225	5,625
6	50	90	4,500	2,500	8,100
7	30	60	1,800	900	3,600
8	60	105	6,300	3,600	11,025
9	50	60	3,000	2,500	3,600
10	20	30	600	400	900
11	55	90	4,950	3,025	8,100
12	45	75	3,375	2,025	5,625
13	40	80	3,200	1,600	6,400
14	60	80	4,800	3,600	6,400
15	40	45	1,800	1,600	2,025
16	60	80	4,800	3,600	6,400
17	50	80	4,000	2,500	6,400
18	55	95	5,225	3,025	9,025
19	50	100	5,000	2,500	10,000
20	35	75	2,625	1,225	5,625
Total	920	1,500	71,600	44,900	118,850

Sometimes we wish to test the hypothesis that ρ has some value other than 0, or to set confidence limits on ρ. Since the distribution of r is not symmetric, we employ a transformation to normalize, approximately, the distribution of the correlation coefficient. The transformation is

$$z_r = \tfrac{1}{2} \log_e \frac{1 + r}{1 - r} \tag{7-25}$$

where $\log_e$ is a natural logarithm, or $2.3026 \log_{10}$.† The standard deviation of z may be expressed approximately as

$$\sigma_z = \frac{1}{\sqrt{n - 3}} \tag{7-26}$$

† Elsewhere in this book the abbreviation "log" always denotes "$\log_{10}$," or the common logarithm.

To test r against a hypothesis ρ, we compute

$$Z = \frac{z_r - z_\rho}{\sigma_z} \tag{7-27}$$

which is approximately normally distributed.

Table VI in the Appendix gives the value of z corresponding to various values of r, avoiding the necessity of working with logarithms.

Suppose we wish to test our computed r against the hypothesis that $\rho = 0.9$. We refer to Appendix Table VI and find the following z values:

$z_\rho = 1.4722$ for $\rho = 0.9$
$z_r = 0.7667$ (by linear interpolation) for $r = 0.642$

Therefore,

$$Z = \frac{z_r - z_\rho}{\sigma_z} = \frac{0.7667 - 1.4722}{1/\sqrt{17}} = -2.91$$

It is unlikely, then, that our sample was drawn from a population with $\rho = 0.9$.

The significance of the difference between two sample correlation coefficients may be tested by

$$Z = \frac{z_1 - z_2}{\sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}} \tag{7-28}$$

where $\sigma_{z_1} = \frac{1}{\sqrt{n_1 - 3}}$ and $\sigma_{z_2} = \frac{1}{\sqrt{n_2 - 3}}$

Confidence limits may be placed on z as follows:

$$P_c(z_r - Z_{\alpha/2}\sigma_z < z_\rho < z_r + Z_{\alpha/2}\sigma_z) = 1 - \alpha \tag{7-29}$$

where z_r equals the z value corresponding to the sample correlation coefficient and $Z_{\alpha/2}$ equals the normal deviate beyond which lies $\alpha/2$ of the area under the normal curve. The conversion to r may be accomplished by looking in the *body* of Table VI for the value of z and then checking the stub and column heading to find the corresponding value of r. As an illustration, we compute a 90 per cent confidence-interval estimate of ρ from the previous data:

$$P_c(z_r - Z_{\alpha/2}\sigma_z < z_\rho < z_r + Z_{\alpha/2}\sigma_z) = 1 - \alpha$$
$$0.7667 - 1.645(0.2425) < z_\rho < 0.7667 + 1.645(0.2425)$$
$$0.3678 < z_\rho < 1.1656$$

We refer to Table VI to convert these to correlation coefficients:

$$0.35 < \rho < 0.82$$

7-9. RELATIONSHIP BETWEEN REGRESSION AND CORRELATION

It will be noted that regression and correlation techniques are based upon different assumptions about the data. Regression assumptions begin with the model $Y_i = A + BX_i + \epsilon_i$, where the X's are assumed to be *fixed quantities*. Correlation techniques are based upon the assumption of a joint probability distribution of both X and Y, so that X is a random variable also.

Practical situations do not always fit the assumptions, and the analyst may find that, even though the correlation assumptions apply, he needs a predicting equation. This is a common occurrence, and, fortunately, the assumption that X is a known quantity is not a critical one. One can obtain useful results from regression techniques even though X and Y are both random variables.

The relationship between regression and correlation is illustrated in Fig. 7-7. The diagram shows the line of regression, a line representing the average of all the Y's, and an actual Y value. The quantity $\Sigma(Y_i - \bar{Y})^2$ is a sum of squares of deviations of the Y values from the mean of the Y's. The quantity $\Sigma(\bar{Y}_X - \bar{Y})^2$ is a sum of squares of deviations between the points on the line of regression and the mean of the Y values. The following relationship holds:

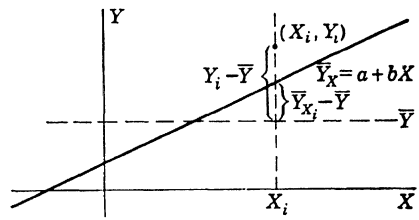

FIG. 7-7. Subdivision of the error in regression estimation

$$r^2 = \frac{\Sigma(\bar{Y}_{X_i} - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} \tag{7-30}$$

Thus we see that the square of the correlation coefficient is the fraction of the total sum of squares in Y that is accounted for by the regression line. A correlation coefficient of $r = 0.5$ means that only 25 per cent of the variation in Y is accounted for by the regression line. The other 75 per cent is accounted for by other factors.

EXERCISE 7-9. A company is trying to determine whether it can screen prospective machine operators effectively by a manual-dexterity test. It gives the test to 92 beginning employees. Six months later the employees are given a score based upon their efficiency in operating the machines. Let X be the score on the dexterity test and Y the efficiency index. The following figures are observed:

$$\Sigma X = 1{,}420 \qquad \Sigma X^2 = 24{,}300$$
$$\Sigma Y = 3{,}000 \qquad \Sigma Y^2 = 101{,}000$$
$$\Sigma XY = 48{,}200$$

(*a*) Find the coefficient of correlation. (*b*) Place 90 per cent confidence limits on it. (*c*) Will the test be of any assistance in screening employees? Why?

EXERCISE 7-10. In order to use the screening test proposed in Exercise 7-9, it is necessary to compute the line of regression. (*a*) Find the regression equation. (*b*) Find the standard error of estimate. (*c*) If an employee scores 23 on the dexterity test would you expect him to score as low as 25 on the efficiency test? Why or why not?

7-10. CURVILINEARITY

It must be remembered that the techniques presented in this chapter apply only to cases in which there is a straight-line relationship between X and Y. This is a serious limitation, but frequently the relationship between X and Y is approximately linear within narrow ranges of X. This fact gives the methods presented here more general application than one might at first suppose. Figure 7-8 illustrates the principle. The relationship between X and Y is curvilinear, but if one is experimenting with values of X between a and b or between c and d, the relationship will appear linear. Figure 7-8 is a typical illustration of what occurs to apparently linear relationships when one extends the range of the X variable. Suppose, for example, that X is machine speed and Y is units produced per hour. Increasing machine speed may at first increase production, but an excessive increase in speed is almost sure to reduce output because of breakdowns, spoiled units, and other factors.

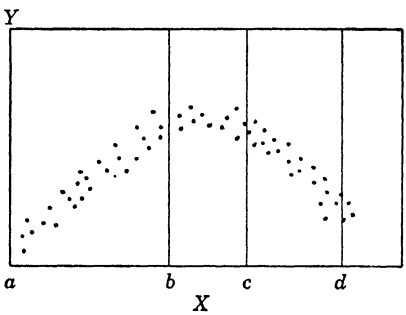

FIG. 7-8. Curvilinear relationship

When one observes a curvilinear relationship within the range of observation, there are techniques for fitting curves to the scatter of points. However, a freehand curve drawn through the scatter of points may give a close enough approximation to the true relationship to serve as a basis for the managerial decision. The range of scatter about such a freehand curve provides an indication of the accuracy with which one can use the curve for prediction.

EXERCISE 7-11. The following values were observed on the speed and temperature of a motor:

X (speed, rpm)	*Y (temperature, °F)*
1,500	150
1,600	159
1,700	172
1,800	187
1,900	196
2,000	214

(*a*) Plot the data and observe the curvilinear relationship. (*b*) Take logarithms of the Y values and plot. (*c*) Compute the regression of the logarithms of Y on X. (*d*) Estimate temperature for a speed of 1,750 rpm. (*e*) Place 90 per cent confidence limits on the true value for $X = 1{,}750$. Note that limits are placed on the logarithms and the antilogarithms taken to find the limits in degrees. (*f*) What assumptions are necessary to validate the limits in (*e*)?

EXERCISE 7-12. Here is a set of data for easy computation.

X	Y	X	Y
0	4	3	7
1	5	4	7
1	6	4	8
2	6		

(*a*) Compute the linear regression of Y on X. (*b*) What value of Y do you estimate for $X = 3.5$? (*c*) Place 99 per cent confidence limits around this individual estimate. (*d*) Compute the correlation between X and Y.

8. ANALYSIS OF VARIANCE AND INDUSTRIAL EXPERIMENTATION

Some apology seems in order for the ambitious title of this chapter. Entire books have been written on limited phases of the problems of industrial experimentation alone. In this chapter we shall restrict ourselves to the examination of some statistical tools which are useful in many areas of statistical analysis and to the presentation of some concepts which are indispensable in experimentation.

8-1. THE NATURE OF EXPERIMENTAL DATA

Statistical data fall into two categories. First, there is the class of data represented by observations on a population, or a segment thereof, where there has been no attempt to modify or "control" any of the possible influencing factors. Examples are sample surveys of buying habits, incomes, and opinions. The second class of data results from varying certain factors in order to determine what effect, if any, they have on the data.

It is very difficult to assign cause and effect by analysis of observational data. The fact that cities with large numbers of crimes also have large numbers of churches does not mean that churches cause crimes. The large numbers of each are due primarily to large numbers of people. In a plant one may observe that college graduates perform better on a job than high school graduates. One cannot say on the basis of this evidence alone that their better performance is due to their college training. One might just as well say that people who go to college have more ambition than those who don't and that it is the difference in ambition which makes them more successful on the job. Or one might attribute higher

rate of success to greater intelligence, or expanded social contacts, or a combination of factors. The point is that the observational data collected will neither confirm nor deny any of these hypotheses.

Consider this illustration. A doctor who is company physician for a large industrial concern examines 20,000 prospective employees. He is interested in vascular disorders—in particular, high blood pressure. He asks each prospective employee whether he uses little salt, a moderate amount of salt, or a lot of salt on his food. Then he has another doctor take the blood pressure of the individuals. The second doctor does not know the response the prospective employee has given to the question, so that his observation of pressure will not be biased by this information. The blood pressures are averaged for each of the three groups of respondents, and it is found that those who use a lot of salt have higher blood pressure, on the average, than those who use little salt. What can the doctor conclude?

If he concludes that excessive use of salt is a cause of high blood pressure, he is using information other than that provided by his survey. From the standpoint of logic, the following hypotheses are just as admissible: (1) people with high blood pressure have a craving for salt; or (2) people with high blood pressure reply that they use more salt whether they actually do or not. However, the doctor may have other information which leads him to suspect that excessive salt is a cause of high blood pressure. It is not likely that he can ever substantiate his hypothesis by observation, so he conducts an experiment. Half of his patients suffering from high blood pressure are chosen at random and asked to reduce their salt intake. In all other respects their treatment does not differ from the "standard" treatment. At the end of a specified period of time it is found that, on the average, those with reduced-salt diets have significantly lower blood pressure. He is now willing to state that reduced salt is significant in the *treatment* of high blood pressure. But he still has not substantiated his hypothesis that excessive salt *causes* high blood pressure.

Suppose he gets the cooperation of the Army and conducts an experiment on a grand scale. Pairs of Army companies are chosen at various locations, and one member of each pair is chosen for a low-salt diet and the other for a high-salt diet. The consumption of salt is reduced in the mess halls of the low-salt companies, and all food is excessively salted in the high-salt companies. After a sufficient lapse of time the men are tested and their average blood pressures compared. The doctor can come closer now to a test of his hypothesis. He has not reached his objective yet, however, because he is dealing exclusively with men, and then only with those whose age and physical condition meet army specifications. Also, the experiment can be conducted only for a small portion

of the life span of an individual—perhaps not for a long enough time for the effect, if any, to become apparent.

We may summarize by saying that (1) observational data seldom reveal cause and effect, but may point the way to logical experiments; (2) properly designed experiments can make it possible to assign cause; and (3) when human beings are the subject of experiments, it is frequently difficult, if not impossible, to get the information one really desires out of the experiment.

From the above discussion it is seen that observational data refer to an existing population, and no attempt is made to modify or change that population during the course of observations. Experimental data, on the other hand, arise from observing a population which has been modified or changed with some particular objective in view.

The layman tends to think of experimentation as a laboratory technique in which one factor is varied, with all other factors held "constant." Historically there is some justification for this point of view, since much of modern science was developed by precisely these methods. However, it is never possible to control exactly all factors except the one whose effect is being examined. Technology is simply not that precise. The variation which is not controlled in an otherwise controlled experiment is termed *experimental error*. It is measured by the variance of the observations under presumably identical conditions.

The process of varying one factor may serve adequately for certain laboratory experiments, but often the results cannot be extended, or *extrapolated*, to the real world, in which innumerable things vary. Thus, if one chooses to conduct a "controlled" experiment in the real world rather than in the laboratory ("control" here is a matter of degree), his results will be more variable. It is convenient in this case to class as experimental error all variation which cannot be accounted for by the variation in the experimental factors, even though some of it might have been accounted for had the experiment been conducted in the laboratory. Thus, taking the experiment out of the laboratory tends to inflate the experimental error and to make it more difficult to distinguish the real effects of the experimental procedure. It has the advantage, however, of generalizing the results to the environment in which the experiment was conducted.

When an effect can be attributed to two or more factors and the portion due to each cannot be determined, we say that the effects of the factors have been *confounded*. This is the situation which arises by the nature of things in observational data. Unless care is taken, it will arise also in experimental data. In determining treatment differences, we should like to know the amount of the difference if the treatments had all been applied to the same unit of experimental material. This is usually impos-

sible to accomplish, however, because the treatment may destroy or "use up" the experimental material or change it in some way so that it cannot be used again. Therefore, we must estimate treatment differences by examining their effects on different experimental units. It is clear, then, that treatment difference is confounded with experimental-unit difference. This difficulty can be circumvented by replication and randomization. *Replication* is the repetition of the same treatment on different experimental units, and *randomization* is the use of a random process to assign experimental units to treatments.

Consider this example. Suppose that two machines are being tested and that the criterion of their effectiveness is the quality of the product. A batch of, say, 1,000 items is produced on each machine, and it is found that machine A is "better" than machine B. We cannot conclude that machine A is actually better than machine B unless we can rule out the possibility that the "operator effects" may have been different. In the absence of this knowledge we must consider the 1,000 items from each machine as a single observation on production. No estimate of experimental error is available until some more batches are run with other operators. The repetition of experiments with various operators has the advantage of "averaging out" operator effects. That is, in a series of trials it would be unlikely that machine A would receive all the good operators and machine B all the bad. Therefore, one is more certain of his conclusion that the observed difference is a real difference. We have, in fact, reduced the confounding with operators by an averaging process. Random assignment of operators to machines makes it possible to generalize the results to the population of operators from which the sample of operators was drawn.

Another way of eliminating confounding is to subdivide "similar" units of experimental material and to apply various treatments to these pieces. "Similarity" is a matter of degree, of course. As an illustration, consider Exercise 5-11 in Chap. 5. We wish to find out whether the method of frying sardines, in the open air or away from air, has any effect on the percentage of free fatty acids obtained. In this case a sample of sardines is a sample from some larger batch, such as a boatload. The sample is then divided into 2 parts and each part is assigned to a treatment at random, say, by the toss of a coin. The sample is the unit of similar experimental material—at least assumed to be similar or homogeneous, since it came from the same boatload of sardines. Applying the 2 treatments to the 2 portions of the same sample avoids confounding with batches of sardines. Averaging over the 7 samples (from 7 different batches) gives us some assurance that the observed difference is not due to the peculiar characteristics of a particular batch. It also provides us with an estimate of experimental error.

In the sardine example one might wonder at the necessity for assigning the 2 parts of the sample to the 2 treatments by a random process. There are two reasons for randomization in experimentation: (1) it eliminates all bias, conscious or otherwise, on the part of the experimenter; (2) it provides an estimate of experimental error because the rules for estimation of error are based strictly on the assumption of randomness. Therefore, every statistical experimental design contains random choice at some stage of the design. There may be, of course, nonstatistical designs which contain no random procedures. They do not have a self-contained estimate of the error, however.

The topics we have discussed—cross classification, experimentation with similar materials, replication, and randomization—form the building blocks of experimental designs. Some illustrations will clarify the concepts.

Suppose we are considering the replacement of our present machines by 1 of 3 possible new makes of machines, which we shall designate M_1, M_2, and M_3. We want to find out which machine is "best" by some logical criterion—the exact criterion is not important for this discussion. We may set up 1 of each of the 3 machines, select 3 employees, say, E_1, E_2, and E_3, assign them to the 3 machines at random, and observe what happens. Our results will not be conclusive for at least two reasons: (1) the machines selected for the experiment may have peculiar characteristics which distinguish them from the classes they are supposed to represent; that is, M_1 may be different from other machines of the M_1 type, and so forth; (2) the employees may have characteristics which distinguish them from each other and from the class of all employees. We see then that our results will be confounded with the particular machines selected and with the particular employees chosen.

There are various ways in which we can reduce the confounding. The obvious way to reduce confounding with particular machines is to repeat the experiment with other machines of the same kind. This will average out individual machine effects and will have the further desirable consequence of providing an estimate of the variability among machines of a particular type.

To reduce confounding with employees, we can use the same technique, repeating the experiment with several employees working on each type of machine. If we wish to generalize our results to the population of all employees, it would seem sensible to choose the machine operators at random from the entire group and to assign them at random to the machines.

Another technique that we might use is to permit each employee selected to try every machine under test. This employs the principle of cross classification and experimentation within similar blocks of experi-

mental material. The term "experimental material" must be interpreted rather broadly. In this case it refers to the machine operator himself, and his characteristics are assumed to remain constant from trial to trial. This means, of course, that there must be no carry-over of training from one machine to another, so that the order of use is not important. One way of controlling this sort of "order bias" is to make sure that the operators are thoroughly trained in the use of the particular test machine before results are recorded. Another technique is to vary the order in which the machines are tried, so that differences due to order tend to balance out. Even then, some attention must be given to carry-over in the interpretation of results.

It is apparent that a relatively simple idea can require a moderately complex experimental design if one is to be fairly certain that results will be conclusive. The important thing is that controls must be built into the design if the experiment is to serve as a basis for assigning cause and effect. The assignment of cause and effect without such built-in controls is extremely hazardous.

The following sections deal with some tools for the analysis of experimental data. The usefulness of the tools is not limited to experimental data, however, as the next example illustrates.

8-2. ANALYSIS OF A SIMPLE EXPERIMENT

We begin by discussing the assumptions underlying the simple analysis of variance. Assume that the jth observation in the ith class can be represented as follows:

$$Y_{ij} = \mu + T_i + \epsilon_{ij} \qquad (i = 1, 2, \ldots, t; j = 1, 2, \ldots, n_i) \quad (8\text{-}1)$$

where μ is a general constant entering into each observation and T_i is a constant applying only to the ith class. We impose the condition that $\Sigma T_i = 0$. The symbol T is used because it is easy to think of a class as representing a particular treatment. The "treatments," however, might represent any logical subdivisions of the experimental data. The quantity ϵ_{ij} is a random variable whose average value is 0 and whose variance is a constant, say, σ^2, for every class. The ϵ_{ij} are also assumed to be independent of each other. That is, the value of ϵ_{ij} on one observation does not affect the value of any other ϵ_{ij}. These are the only assumptions necessary to validate estimates of the treatment effects (the T_i). However, in order to make tests of hypotheses, it also is necessary to specify that the ϵ_{ij} are normally distributed. This assumption is not a critical one, however, and it has been shown that the usual tests of hypotheses are approximately correct without the normality assumption.

It is interesting to note the similarity between the above model and the

linear-regression model (7-1). As a matter of fact, one can estimate the parameters in (8-1) by essentially the same techniques as presented in Chaps. 7 and 10.

It may be helpful to discuss the simple analysis of variance in the light of a specific illustration. Mr. O'Callaghan, foreman of a machine shop, has 5 workers under his jurisdiction. He is concerned with the time that they devote to the coffee break, so he times them (the rascal!). His observations, together with the total and mean for each employee, are shown in Table 8-1.

The number of observations is not the same for each individual, but this does not matter. We may assume that employees A, B, and D were absent on certain days, so we have fewer observations for them. Least-squares estimates of the quantities $\mu + T_i$ are found by taking the simple means of the columns.* That is, our "best estimate" of the average amount of time Mr. B devotes to the coffee break is 21 min, assuming, of course, that the recorded observations are representative of his typical behavior.

Table 8-1. Time, in Minutes, Spent on the Coffee Break by 5 Employees of the Machine Shop

Observations	A	B	C	D	E	Total
1	13	23	24	18	21	
2	15	19	22	17	22	
3	11	19	28	21	17	
4	15	23	26	15	17	
5	16		24		20	
6			27		21	
Sum	70	84	151	71	118	494
n_i	5	4	6	4	6	25
Avg.	14.0	21.0	25.2	17.8	19.7	19.8

Mr. O'Callaghan observes that the average time spent on the coffee break varies from 14 to 25.2 min. He asks himself whether this variation is too great to be accounted for by random variation. In other words, he wishes to test the hypothesis that the column means are equal. This hypothesis can be stated formally as $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. Our test of this hypothesis is accomplished by comparing the variance among the column means with the variance within the columns. The rationale here is that if the variance among the column means is significantly greater than the variance within the columns, the added variance must

* The student with some mathematical background can prove this easily by minimizing the quantity $\Sigma_{i,j}(y_{ij} - \mu - T_i)^2$ with respect to μ and T_i and imposing the condition that $\Sigma \hat{T}_i = 0$, where $\hat{T}_i$ is an estimate of T_i.

be due to real differences among the columns, rather than to chance factors.

The variance within the columns is a pooled variance (see Chap. 5), formed by weighting the 5 separate variances by their respective degrees of freedom and then averaging. It is the total sum of squares within the columns divided by the total degrees of freedom within the columns. We may express the formula as

$$\text{Variance within columns} = \frac{\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2}{\Sigma n_i - t} \tag{8-2}$$

where $\bar{Y}_i$ is the mean of the ith column and t is the number of columns. The variance of the column means can be expressed as

$$\text{Variance among column means} = \frac{\sum_i n_i(\bar{Y}_i - \bar{Y})^2}{t - 1} \tag{8-3}$$

where $\bar{Y}$ is the grand mean. Note that the square of the difference between the column mean and the grand mean is multiplied by the number of observations in the column before summing. Since there are t column means, we divide by $t - 1$ degrees of freedom. We test the hypothesis that the column means are equal by dividing (8-3) by (8-2). This quantity is distributed as F with $t - 1$ and $\Sigma n_i - t$ degrees of freedom.

Formulas (8-2) and (8-3) are definitional formulas. Ordinarily we do not use them in the actual computation any more than we should use

$$s^2 = \frac{\Sigma(Y_i - \bar{Y})^2}{n - 1}$$

to compute the variance in the single sample case. We usually compute the sums of squares [the numerators of (8-2) and (8-3)] and then divide by degrees of freedom. We first compute the total sum of squares in the entire table, ignoring the classification into columns. It is clear that this sum of squares is

$$\text{Total SS} = \sum_{ij} Y_{ij}^2 - \frac{\left(\sum_{ij} Y_{ij}\right)^2}{n} \tag{8-4}$$

where $n = \Sigma n_i$. Next, we compute the sum of squares among column means as*

$$\text{SS among column means} = \sum_i \frac{\left(\sum_j Y_{ij}\right)^2}{n_i} - \frac{\left(\sum_{ij} Y_{ij}\right)^2}{n} \tag{8-5}$$

* The algebra leading to these formulas is given in Sec. 8-4.

That is, we square each column total and divide by the number of observations in the column, sum, and subtract the same correction factor as in (8-4). The sum of squares within columns is found as

$$\sum_j Y_{1j}^2 - \frac{\left(\sum_j Y_{1j}\right)^2}{n_1} + \sum_j Y_{2j}^2 - \frac{\left(\sum_j Y_{2j}\right)^2}{n_2} + \cdots + \sum_j Y_{tj}^2 - \frac{\left(\sum_j Y_{tj}\right)^2}{n_t} = \sum_{ij} Y_{ij}^2 - \sum_i \frac{\left(\sum_j Y_{ij}\right)^2}{n_i} \quad (8\text{-}6)$$

This is precisely what we obtain by subtracting (8-5) from (8-4). This is an important principle in the analysis of variance. We describe it by saying that sums of squares are *additive*. It is a convenient property for checking purposes.

The computations for our illustration are as follows:

$$\text{Total SS} = (13)^2 + (15)^2 + (11)^2 + \cdots + (21)^2 - \frac{(494)^2}{25}$$
$$= 10{,}224 - 9{,}761.44 = 462.56$$

$$\text{SS among column means} = \frac{(70)^2}{5} + \frac{(84)^2}{4} + \frac{(151)^2}{6} + \frac{(71)^2}{4} + \frac{(118)^2}{6} - \frac{(494)^2}{25} = 10{,}125.08 - 9{,}761.44 = 363.64$$

$$\text{SS within columns} = 462.56 - 363.64 = 98.92$$

As a convenient way of presenting these sums of squares, along with their degrees of freedom and variances (mean squares),* we set up a summary table such as Table 8-2.

Table 8-2. Summary Table for Coffee Break Data

Due to	SS	*df*	MS
Employees (columns)	363 64	4	90.91
Error (within columns)	98 92	20	4.95
Total	462.56	24	

Since the within-columns mean square is to be used as the standard against which we shall measure our employees' mean square, we may label it "error." This implies that it is an estimate of experimental error or statistical error. In fact, it can be shown that it is an unbiased estimate of the quantity σ^2 in the definition of the model at the beginning of this section.

* The term *mean square* (MS), rather than *variance*, is traditionally used for this type of analysis.

We construct the F test as follows:

$$F = \frac{90.91}{4.95} = 18.37 \qquad \text{with 4 and 20 } df$$

The 0.05 level of F for these degrees of freedom is 2.87, and the 0.01 level is 4.43. Since our computed level *exceeds* both of these figures, it is said to be significant at the 0.01 level. In other words, the between-columns variance is significantly greater than the within-columns variance, and our hypothesis of the equality of the column means is rejected.

Mr. O'Callaghan decides to speak to Mr. C about the excessive time taken for the coffee break. He is not likely to suggest that Mr. A increase his break time, although statistically this might have the same effect, that is, of reducing the discrepancies among the means. Actually the analysis of variance only shows that some differences among the means exist. It does not show where those differences are. One must return to the observed means for this judgment.

Since the error mean square is an unbiased estimate of the error variance, one can immediately find the standard error of each column mean. If we let E represent the error mean square, then the standard error of the ith column mean is

$$s_{\bar{Y}_i} = \sqrt{\frac{E}{n_i}} \tag{8-7}$$

It may be remembered from Chap. 4 that subtracting a constant from the sample figures does not affect the variance. Therefore, one might have subtracted a convenient constant, say, 10, from each observation in Table 8-1 before doing the computations. The mean squares would have been identical with those computed from the original data.

Note that the data used in the analysis of variance in the above illustration are observational data rather than experimental data. The analysis is valid in either case, provided that the assumptions relating to the model (8-1) are met.

Another interesting thing to note about the F test is that the hypothesis of equality of the column means is rejected only when the F ratio is *large*. Small values of F simply indicate that the column means are not more diverse than one would expect if they did, in fact, come from the same population. Thus, if we let α be the level of significance chosen, we see that the region of rejection is $F \geq F_\alpha$, where F_α indicates the value of F having α of the area under the F curve to the right of it.

Exercise 8-1. A time study is undertaken in order to establish the average time required to assemble a small mechanism. Six assemblers are chosen at random from the entire group of assemblers. Three of these are chosen at random and told that they are to be timed. The other 3 are not aware of the

time study. Five observations are taken on each assembler, with the following results, recorded in seconds.

Assemblers aware of study			Assemblers unaware of study		
A_1	A_2	A_3	A_4	A_5	A_6
46	49	44	66	85	86
66	45	52	88	95	59
32	33	48	43	83	68
49	53	52	59	85	70
38	30	62	56	99	56

(*a*) What are your estimates of the average time for each employee? (*b*) What are your estimates of the average times for the 2 groups, those aware of the time study and those unaware of the study? (*c*) Test the hypothesis that the true averages of the 6 assemblers are equal.

Exercise 8-1 is an example of a *nested* classification. That is, the whole class of assemblers has been divided into 2 classes, those aware of the experiment and those not aware. Each of these classes is represented by 3 persons, and there are 5 observations for each of these assemblers.

The test of the hypothesis suggested in part (*c*) of Exercise 8-1 is not an adequate analysis for a nested classification. We are probably more interested in comparing the means of the 2 groups of assemblers than we are in testing the hypothesis that all 6 means are equal. We can subdivide our total sum of squares as follows:

1. Sum of squares between group means
2. Sum of squares among assemblers within the first group
3. Sum of squares among assemblers within the second group
4. Sum of squares among observations within assemblers

The second and third sums of squares may be combined and called the sum of squares among assemblers within groups. Let us see how the sums of squares are computed.

The total sum of squares is computed in the usual manner and represents the sum of squares of deviations around the grand mean of the entire 30 observations. To discuss the other sums of squares some supplementary notation will be helpful. Let $[T]$ denote the grand total of the table, $[G_1]$ the total of the first group (15 observations on the first 3 employees), $[G_2]$ the total for the second group, $[A_1]$ the total for the first assembler, and so forth. Then the sum of squares between group means is the weighted sum of squares of deviations of the two group means from the grand mean. If $\bar{G}_1$ is the first group mean, $\bar{G}_2$ is the second group mean, and $\bar{T}$ is the grand mean, the computation can be done as

$$15(\bar{G}_1 - \bar{T})^2 + 15(\bar{G}_2 - \bar{T})^2$$

However, by the rules discussed earlier, we can arrive at the same result by

$$\frac{[G_1]^2}{15} + \frac{[G_2]^2}{15} - \frac{[T]^2}{30}$$

EXERCISE 8-2. Compute in both ways and show that the results are the same.

Similarly, the sum of squares among assemblers within the first group can be computed as

$$\frac{[A_1]^2 + [A_2]^2 + [A_3]^2}{5} - \frac{[G_1]^2}{15}$$

EXERCISE 8-3. What is the corresponding form of this computation using means of individual assemblers?

The computation for sum of squares among assemblers in the second group is the same as the above, with a slight change in the subscripts.

The sum of squares among observations within assemblers is computed as

$$\sum_j (Y_{1j})^2 - \frac{[A_1]^2}{5} + \cdots + \sum_j (Y_{6j})^2 - \frac{[A_6]^2}{5}$$

where Y_{ij} indicates the jth observation in the ith column.

EXERCISE 8-4. Compute all the sums of squares and verify that the components add up to the total sum of squares.

The next problem is to associate a number of degrees of freedom with each sum of squares computed above. We observe that there are 2 groups of employees, hence 1 degree of freedom for the sum of squares between group means. There are 3 assemblers in each group, hence 2 degrees of freedom for the sum of squares among assemblers *within each group*. Similarly, there are 4 degrees of freedom among each set of 5 observations. The analysis-of-variance summary table takes the form of Table 8-3.

Table 8-3. Analysis of Variance Summary Table for Time-study Data

Due to	SS	*df*	MS
Groups		1	(G)
Assemblers within groups		4	(A)
Within first group		2	(A_1)
Within second group		2	(A_2)
Observations within assemblers		24	(O)
Total		29	

EXERCISE 8-5. Fill in the sums of squares and mean squares in the above summary table.

What are the hypotheses to be tested and how do we test them? Several hypotheses can be tested. First, it may be of interest to find out whether the averages vary significantly within the first group. In this case a valid experimental error for testing the equality of the 3 means within the first group is provided by the variance among observations within the first group. If there is reason for believing that the observations within the second group are no more variable or no less variable than those within the first group, then the entire mean square for variation among observations can be used as the denominator of the F test. Combining errors in this manner has the advantage of increasing degrees of freedom and hence increasing the sensitivity of the test. One can test the hypothesis that the variances among observations are uniform over the 2 groups, and this test will be presented later. For the present, let us assume homogeneity of the variances and proceed with the test by computing

$$F = \frac{(A_1)}{(O)} \qquad \text{with 2 and 24 } df$$

where (A_1) and (O) are the appropriate mean squares from the summary table.

EXERCISE 8-6. Perform the above test and draw a conclusion about uniformity of the means within the first group of employees.

EXERCISE 8-7. Perform a similar test of the equality of means within the second group.

If one uses the ratio $(A)/(O)$ from the summary table with 4 and 24 degrees of freedom, he is testing the hypothesis that the 3 means within the first group are equal and that the 3 means within the second group are equal, but not that the 2 groups are equal to each other. That is, letting the means be denoted by $\mu_1, \mu_2, \ldots, \mu_6$, this F ratio tests the hypothesis that $\mu_1 = \mu_2 = \mu_3$ and that $\mu_4 = \mu_5 = \mu_6$, but μ_1 need not equal μ_4, and so forth.

Another hypothesis to be tested is that the mean of the first group is equal to the mean of the second group. Clearly we shall use the mean square (G) as the numerator of the F ratio, but what should we use as experimental error? The mean square listed as (O) in the table will not do because it contains variation only among observations for each assembler individually. It does not include variation among individuals within the groups. On the other hand, the mean square represented as (A) seems ready-made for this purpose, since it contains variation due to differences in individuals as well as the observational error on individuals

themselves. Hence the required test is accomplished by

$$F = \frac{(G)}{(A)} \qquad \text{with 1 and 4 } df$$

EXERCISE 8-8. Perform the above test and state your conclusions.

In this illustration, testing the hypotheses has little real meaning since it is probably taken for granted that individuals will vary and that assemblers who know they are being timed will react differently than they would ordinarily. In this case we should probably be interested in determining a standard or average of some sort. If we wish to use the average in order to estimate costs of producing a large contract order, we shall undoubtedly use the group of individuals that did not know about the study. On the other hand, if we are interested in establishing standards which are in the nature of objectives, the other group will be of major interest to us. Also, from a scientific management point of view, the amount of the difference between the groups will hold some interest. Here then we are concerned primarily with *means* and their *standard errors.*

Suppose we are interested in the average of the group that has not been told about the experiment. Our estimate of the average performance of this group is the average of the means of the 3 individuals in the group. The mean square (A) forms the basis for our estimation of the standard error, *under the assumption that the variances within the two groups are homogeneous.* If this assumption is untenable, then we must use the mean square (A_2). Assuming homogeneity of (A_1) and (A_2) (which itself can be tested by the F test), how do we find the variance of a group mean? We recall that (A) is computed by squaring deviations of assembler means from group means, *weighting by number of observations,* and dividing by degrees of freedom. The weighting by number of observations has the effect of putting the variance (A) on an individual-observation basis. That is, it makes the variance among means comparable to the variance among observations. In order to have the variance apply to means, rather than to individual observations, we must divide the variance (A) by the number of observations in the second group (15 observations in all). The square root is the standard error of the group mean: $\text{SE} = \sqrt{(A)/15}$.

EXERCISE 8-9. (*a*) Compute the above standard error. (*b*) What is the standard error of the first group mean? (*c*) Find the 2 group means and place 90 per cent confidence limits on the difference between them, using the combined error (A).

It will be noted that in the discussion of this illustration no reference has been made to a mathematical model. The approach to the analysis

has been intuitive, based upon the ability to partition the sum of squares in a meaningful manner. This approach was taken deliberately. Actually a major portion of the common analyses of variance were developed in this manner, without reference to a model. It may be helpful, however, to examine the mathematical model exemplified by this illustration. Consider the model

$$Y_{ijk} = \mu + G_i + A_{ij} + e_{ijk} \tag{8-8}$$

where G_i is the special effect associated with the ith group, A_{ij} is the special effect associated with the jth assembler in the ith group, and e_{ijk} is an observational error associated with the kth observation on the jth assembler in the ith group. It will be noted that the e_{ijk} are random errors with average value of 0 and constant variance of σ^2. Also, since the assemblers were drawn at random, the A_{ij} must be considered a random variable whose average value is 0 and whose variance is, say, $\sigma_A{}^2$. This is a variance among the true means of all the possible assemblers in the group. When we compute the mean-square error among the group means, we include both observational error and "assembler" error. It can be shown that what we are actually estimating is $\sigma^2 + 5\sigma_A{}^2$, since there are 5 observations on each assembler and since the variance is on a "per observation" basis.

The computed group mean, which is of some interest to us in the time study, is an estimate of the quantity $\mu + G_i$. As we have seen earlier, we estimate the variance of this group mean by the quantity $(A)/15$. This is, in fact, an estimate of the quantity $(\sigma^2 + 5\sigma_A{}^2)/15$, as seen from the above argument.

The analysis described above can be generalized easily. Suppose there are a classes of factor A. Within these classes we experiment with b classes of factor B. Within each of these we experiment with c classes of factor C. Finally, we take d observations on each individual. The analysis of variance is summarized in Table 8-4. To test for equality of the C means, we use the F ratio C/D, since the only difference between the expected or true mean squares is the term $d\sigma_B{}^2$. For similar reasons we test for equality of B means by using the F ratio B/C and for equality of A means by A/B.

It is important to note that the validity of the analysis is dependent upon the assumption that the classes of B have been selected at random from a large set of such possible classes, and the same, of course, must hold for the C classes. Failure to select at random in any of the stages may change the appropriate error for a particular F test. These are complications which it seems wise not to pursue in this brief introduction to the analysis of variance.

Table 8-4. Analysis of Variance for a Nested Classification

Due to	*df*	MS	MS is an estimate of
Between A groups	$a - 1$	A	$\sigma^2 + d\sigma_C^2 + cd\sigma_B^2 + bcd\sigma_A^2$
Between B groups, within A groups	$a(b - 1)$	B	$\sigma^2 + d\sigma_C^2 + cd\sigma_B^2$
Between C groups, within B groups	$ab(c - 1)$	C	$\sigma^2 + d\sigma_C^2$
Error (observational)	$abc(d - 1)$	D	σ^2
Total	$abcd - 1$		

It will be remembered that the choice of an appropriate experimental error depended upon the assumption of homogeneity of variances. In general, this assumption can be tested as a hypothesis. A common test for homogeneity of variances is Bartlett's test.* Suppose there are k variances to be compared, denoted by $s_1^2, s_2^2, \ldots, s_k^2$, each with $n_i - 1$ degrees of freedom. Then the quantity

$$\chi^2 = \frac{2.3026[\log \bar{s}^2 \Sigma(n_i - 1) - \Sigma(n_i - 1) \log s_i^2]}{1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{n_i - 1} - \frac{1}{\Sigma(n_i - 1)}\right]} \tag{8-9}$$

is distributed as χ^2 with $k - 1$ degrees of freedom. The quantity $\bar{s}^2$ is the weighted average of the variances, using degrees of freedom as weights. It is a pooled estimate of the variance and is the same quantity that we have denoted by s_p^2 in earlier chapters.

Exercise 8-10. Test for homogeneity of observational errors among the 6 assemblers in the previous illustration, using Bartlett's test.

8-3. THE RANDOMIZED-BLOCKS EXPERIMENT

The randomized-blocks design receives its name from agronomic experiments in which several "treatments," or "varieties," are tested on a "block" of land. Repetitions, or replications, of the basic experiment are accomplished by using more than one block of land. The randomization occurs in the assignment of treatments to plots within the block. Despite its agricultural origin, this design is used widely in nearly all experimental fields. The fundamental principle is that a comparison among all treatments is made within a small block of experimental material, thus reducing the effect of environmental variation.

The model for this experiment can be expressed as follows:

$$Y_{ij} = \mu + T_i + B_j + \epsilon_{ij} \tag{8-10}$$

* M. S. Bartlett, "Some Examples of Statistical Methods of Research in Agriculture and Applied Biology," *Suppl. J. Roy. Statist. Soc.*, vol. 4, pp. 137–183, 1937.

where μ = general constant
T_i = effect due to the ith treatment ($i = 1, 2, \ldots, t$)
B_j = effect due to the jth block ($j = 1, 2, \ldots, b$)
ϵ_{ij} = a random variable with the same properties as in the simple random design of the previous section

For convenience in the solution, it is assumed that $\Sigma T_i = \Sigma B_j = 0$. Note that the model assumes additivity of block and treatment effects. That is, it assumes that there is no joint effect of the T_i and B_i beyond the sum of their simple effects. More will be said about this topic later.

We can subdivide the total sum of squares (and the corresponding degrees of freedom) into the following components: the sum of squares among block means, the sum of squares among treatment means, and the sum of squares for error. The block means generally hold little interest for us, since they are indicative of differences in the blocks of experimental material. The treatment means, however, are of a great deal of interest, and the simple arithmetic mean of all observations for the ith treatment is an unbiased estimate of the quantity $\mu + T_i$.

Suppose that a processor of dairy products is interested in comparing 4 storage procedures. The variable of interest is an index of the bacteria count after 72 hr of storage. Since the milk which he receives is variable with respect to bacteria count, he will wish to experiment with several lots or batches. He chooses to use 5 batches, which constitute his blocks of experimental data. He takes a portion of each batch, divides it into 4 parts, and assigns the treatments at random to the portions. The experimental results appear in Table 8-5.

Table 8-5. Bacteria Count (Coded) in Milk after Storage for 72 Hours under 4 Storage Methods

Batch	Storage treatment				Total
	S_1	S_2	S_3	S_4	
1	2	2	5	3	12
2	7	9	12	8	36
3	5	11	11	9	36
4	4	4	6	6	20
5	2	4	6	4	16
Total	20	30	40	30	120
Mean	4	6	8	6	6

The sums of squares are computed as follows:

$$\text{Total SS} = 2^2 + 7^2 + \cdots + 4^2 - \frac{120^2}{20} = 904 - 720 = 184$$

$$\text{SS among column means} = \frac{20^2 + 30^2 + 40^2 + 30^2}{5} - \frac{120^2}{20}$$

$$= 760 - 720 = 40$$

$$\text{SS among row means} = \frac{12^2 + 36^2 + 36^2 + 20^2 + 16^2}{4} - \frac{120^2}{20}$$

$$= 848 - 720 = 128$$

$$\text{Error SS} = \text{total SS} - \text{SS among column means} - \text{SS among row means}$$

$$= 184 - 40 - 128 = 16$$

Note that the additive property of the sums of squares makes it possible for us to compute the error sum of squares by subtraction. We can compute it directly by the formula

$$\text{Error SS} = \sum_{ij} (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2 \qquad (8\text{-}11)$$

where $\bar{Y}_i$ is the ith treatment mean and $\bar{Y}_j$ is the jth block mean.

The analysis-of-variance summary is shown in Table 8-6. Note that the number of degrees of freedom for error is the product of the degrees of freedom for batches and for treatments. This is always true in the randomized-blocks design. The error mean square can be shown to be an unbiased estimate of the error variance of the ϵ_{ij} [Eq. (8-10)]. This quantity, then, is used as a standard against which to judge the mean square for treatments. The test of the hypothesis $T_1 = T_2 = T_3 = T_4$ is accomplished by the F test.

Table 8-6. Analysis of Variance Summary of the Milk-storage Data

Due to	SS	*df*	MS
Batches (blocks)	128	4	
Treatments	40	3	13.33
Error	16	12	1.33
Total	184	19	

$$F = \frac{13.33}{1.33} = 10 \qquad \text{with 3 and 12 } df$$

The F value required for significance at the 0.01 level is 5.95, so we conclude that there is a real difference among the treatment means. Since there are 5 observations for each treatment, the standard error of a treatment mean is

$$s_{\bar{T}} = \sqrt{\frac{E}{b}} = \sqrt{\frac{1.33}{5}} = 0.52$$

It is typical of a randomized-blocks experiment that the treatments are chosen arbitrarily, rather than at random from a large population of possible treatments. In either case, the appropriate F test is the treatments' mean square divided by the error mean square. The most critical assumption in validating the analysis is that blocks and treatments do not interact, that is, that the true differences in treatment effects are independent of the block in which the comparison is made. If this assumption seems questionable, techniques are available whereby it may be tested as a hypothesis.*

It should be noted that the "blocks" represent replicates of the experiment. Selection of pieces of experimental material by random selection from a designated population permits generalization of the results to that population. Arbitrary selection of blocks may limit seriously the scope of possible generalization. The tests of hypotheses are valid, however, regardless of the manner in which blocks were selected, as long as the assumptions of the model are met. There is little point, though, in testing hypotheses about a population which has no correspondence to the world in which the results are to be applied.

EXERCISE 8-11. A motion study is conducted to determine the best placement of parts bins at an assembly table. Four arrangements are under study. Ten assemblers are selected at random from the hundred assemblers in the plant. All assemblers in the test are given adequate time to become familiar with the particular placements chosen for the experiment. Then a test is conducted, and the number of assemblies put together in an hour is recorded, with the following results:

Assembler	Placement			
	A	*B*	*C*	*D*
1	8	11	23	16
2	34	35	41	39
3	28	52	64	50
4	6	14	11	8
5	24	38	35	38
6	16	23	20	20
7	16	26	27	24
8	34	49	48	50
9	34	38	43	39
10	26	43	39	43

* For example, see J. W. Tukey, "One Degree of Freedom for Nonadditivity," *Biometrics*, vol. 5, pp. 232–242, 1949.

(*a*) Make an analysis of variance of the above data. (*b*) Compute the means of the different placements and their standard errors. (*c*) What is the population about which you are generalizing? (*d*) What is the importance of having chosen the 10 assemblers at random? (*e*) How would you organize an experiment to determine whether the order of the trials is significant?

8-4. TWO-FACTOR EXPERIMENTS

Sometimes a researcher may be interested in examining the effect of two classes of treatments, or "factors." For example, in the milk-storage experiment, he may be interested in studying both the effect of different types of containers and the effect of different storage temperatures. He is interested not only in the individual effect of each of these factors but also in their joint effect. That is, certain temperatures may work better in conjunction with particular containers than with others. This differential effect is called *interaction.*

To take another example, suppose that an experiment is conducted on the life of cutting tools and that the factors of interest are the speed of the lathe and the type of metal. We assume the data of Table 8-7. There are 2 observations at each speed for each type of metal. It is convenient in thinking of interaction to consider a table of averages such as Table 8-8. If the differences among the figures in the last column of Table 8-8 are greater than can be accounted for by random variation, then we say there is significant interaction between metals and speeds.

Table 8-7. Life of Cutting Tools, in Units Produced, by Type of Metal and Speed of Lathe

Speed of lathe	Type of metal		Total
	Metal A	Metal B	
Speed 1	15	10	
	16	9	
	31	19	50
Speed 2	12	6	
	10	8	
	22	14	36
Speed 3	10	3	
	11	2	
	21	5	26
Total	74	38	112

Table 8-8. Average Life of Cutting Tools by Type of Metal and Speed of Lathe

Speed of lathe	Type of metal		Metal A minus metal B
	Metal A	Metal B	
Speed 1	15.5	9.5	6.0
Speed 2	11.0	7.0	4.0
Speed 3	10.5	2.5	8.0

Following the procedure already established, we shall consider the analysis with respect to a model. In the cutting-tools illustration we assume that the individual pieces of metal are assigned at random to the cutting tools, so that the order of the observations within each cell is unimportant. It is, then, a simple random experiment. For this design, we establish the following model:

$$Y_{ijk} = \mu + M_i + S_j + (MS)_{ij} + \epsilon_{ijk} \tag{8-12}$$

where M_i = effect of the ith class of metals
S_j = effect of the jth class of speeds
$(MS)_{ij}$ = effect of the interaction between metals and speeds in the ijth cell
ϵ_{ijk} = a random variable with the usual properties

We assume that $\Sigma M_i = \Sigma S_j = \Sigma(MS)_{ij} = 0$.

We have four components of the total sum of squares to consider: the sum of squares due to M, the sum of squares due to S, the sum of squares due to interaction, and the error sum of squares. For the cutting-tools experiment we perform the computations as follows:

$$\text{Correction factor} = C = \frac{(112)^2}{12} = 1{,}045.33$$

Since this term is used in each of the following computations, it is convenient to compute it first.

$$\begin{aligned}\text{Total SS} &= 15^2 + 16^2 + 12^2 + \cdots + 3^2 + 2^2 - C \\ &= 1{,}240.00 - 1{,}045.33 = 194.67\end{aligned}$$

$$\begin{aligned}\text{SS due to metals} &= \frac{74^2 + 38^2}{6} - C \\ &= 1{,}153.33 - 1{,}045.33 = 108.00\end{aligned}$$

$$\begin{aligned}\text{SS due to speeds} &= \frac{50^2 + 36^2 + 26^2}{4} - C \\ &= 1{,}118.00 - 1{,}045.33 = 72.67\end{aligned}$$

We compute the sum of squares due to interaction as the sum of squares among subgroups minus the sum of squares due to metals minus the sum of squares due to speeds, where the subgroups are the observations treated alike:

$$\text{SS for subgroups} = \frac{31^2 + 19^2 + 22^2 + 14^2 + 21^2 + 5^2}{2} - C$$

$$= 1{,}234.00 - 1{,}045.33 = 188.67$$

$$\text{SS for interaction} = 188.67 - 108.00 - 72.67 = 8.00$$

The error sum of squares can be obtained by subtraction, that is,

$$\text{Total SS} - \text{SS for metals} - \text{SS for speeds} - \text{SS for interaction}$$
$$= 194.67 - 108.00 - 72.67 - 8.00 = 6.00$$

It can be obtained also by computing the sum of squares within each subgroup and summing over all subgroups, that is,

$$15^2 + 16^2 - \frac{31^2}{2} + \cdots + 3^2 + 2^2 - \frac{5^2}{2} = 6.00$$

The student can easily verify this result.

Table 8-9. Analysis of Variance Summary for Cutting-tools Experiment

Due to	SS	*df*	MS
Metals	108.00	1	108.00
Speeds	72.67	2	36.33
Interaction	8.00	2	4.00
Error	6.00	6	1.00
Total	194.67	11	

The analysis-of-variance summary is presented in Table 8-9. Note that under the assumptions for the model the error mean square is an unbiased estimate of σ^2, the variance of the ϵ_{ijk}. We test for interaction by

$$F = \frac{4.00}{1.00} = 4.00 \qquad \text{with 2 and 6 } df$$

This F value is not significant at either the 0.05 or the 0.01 level. We conclude, therefore, that there is no very clear evidence of interaction between speeds and metals. We test for the significance of speeds and metals separately in the same manner, finding $F = 108.00$ for metals and $F = 36.33$ for speeds. Both are significant at the 0.01 level.

If, in such an experiment, we find that interaction is significant, it is a bit pointless to test the main effects (speeds and metals). The fact

that there is significant interaction between M and S seems on the face of it to preclude saying that M or S is not significant.

Occasionally, one may have a single observation on each treatment combination and may then ask whether a test of the main effects is possible. *If one can assume that interaction is zero*, the main-effect mean squares can be tested against interaction mean square, because then we have essentially the randomized-blocks model. If, however, interaction is not negligible, there is no very good test available. Sometimes the data are transformed by use of logarithms, square roots, or some other device, in order to reduce interaction so that a test can be made. Such techniques generally require considerable knowledge concerning the nature of the physical, biological, or social laws which determine the observations.

The factorial experiment can be fitted easily into the randomized-blocks design. In this case all the treatment combinations are randomized over a block of experimental material, and this basic arrangement is repeated as often as is necessary to achieve the accuracy required. In this case the model becomes

$$Y_{ijk} = \mu + C_i + A_j + B_k + (AB)_{jk} + \epsilon_{ijk} \tag{8-13}$$

where C_i = block effect for the ith block
A_j = effect of the A factor for the jth class of A
B_k = effect of the B factor for the kth class of B
$(AB)_{jk}$ = effect of the interaction in the jkth cell

If there are c blocks, a classes of A, and b classes of B, the subdivision of degrees of freedom (and corresponding sums of squares) is as shown in Table 8-10.

Table 8-10. Subdivision of the Degrees of Freedom for a Factorial Experiment Organized by Randomized Blocks

Due to	*df*
Blocks	$c - 1$
A	$a - 1$
B	$b - 1$
AB	$(a - 1)(b - 1)$
Error	$(c - 1)(ab - 1)$
Total	$abc - 1$

Table 8-11 presents operating times (in seconds) of 3 machine operators at 2 room temperatures on 4 lots of raw material. The lots of raw material constitute the blocks in the randomized-blocks design. The sums of squares are computed as follows:

$$\text{Correction factor} = C = \frac{233^2}{24} = 2{,}262.04$$

$$\text{Total SS} = 5^2 + 4^2 + \cdots + 14^2 - C = 248.96$$

$$\text{SS among operators} = \frac{77^2 + 63^2 + 93^2}{8} - C = 56.34$$

$$\text{SS among temperatures} = \frac{120^2 + 113^2}{12} - C = 2.04$$

$$\text{SS among lots} = \frac{39^2 + 58^2 + 56^2 + 80^2}{6} - C = 141.46$$

$$\text{SS for subgroups} = \frac{41^2 + 36^2 + 32^2 + 31^2 + 47^2 + 46^2}{4} - C = 59.71$$

SS for operators × temperatures' interaction = subgroup SS − SS for operators − SS for temperature = 59.71 − 56.32 − 2.04 = 1.33

Error SS (by subtraction)

The only significant F ratio is that obtained from the operators' mean square.

Table 8-11. Operating Time, in Seconds, of 3 Machine Operators at 2 Room Temperatures on 4 Lots of Material

Lot	Temperature	Operators			Total
		1	2	3	
1	High	5	4	10	19
	Low	4	6	10	20
		9	10	20	39
2	High	10	8	12	30
	Low	8	9	11	28
		18	17	23	58
3	High	11	8	10	29
	Low	8	8	11	27
		19	16	21	56
4	High	15	12	15	42
	Low	16	8	14	38
		31	20	29	80
Total	High	41	32	47	120
	Low	36	31	46	113
		77	63	93	233

Table 8-12. Analysis of Variance Summary for Operators, Temperatures, and Lots

Due to	SS	*df*	MS
Lots	141.46	3	
Operators	56.34	2	28.17
Temperatures	2.04	1	2.04
Operators × temperatures	1.33	2	0.67
Error	47.79	15	3.19
Total	248.96	23	

EXERCISE 8-12. The manner in which the lots and operators of Table 8-11 were selected was not specified. What effect does their method of selection have upon the generalizations which can be made?

EXERCISE 8-13. Suppose in Table 8-11 that operators and lots were both chosen at random. What is the high-temperature mean and what is its standard error?

8-5. FURTHER DISCUSSION OF PRINCIPLES*

The validity of the computational forms presented in the previous two sections can be established quite easily by some simple algebraic manipulation. Although it is not necessary from the standpoint of the arithmetic to know the basis for the formulas, one must be able to follow through some of the derivations in order to understand the statistical principles involved.

On many occasions we have used the term "expected value" to mean the average value over the entire population. At this time it seems appropriate to examine "expectation" as a mathematical operation which will be helpful to us in the material which follows.

It may be helpful to refer again to Sec. 4-2, in which a random observation drawn from a population was defined by the expression

$$Y_i = \mu + e_i$$

where μ is the mean of the population and e_i is a random variable whose average value over the whole population is 0. In other words, the expected value of e_i equals 0 ($Ee_i = 0$). Also, since μ is, by definition, the average of the Y_i values over the entire population, we can say that the expected value of Y_i equals μ ($EY_i = \mu$).

The operation of expectation need not be limited to the variables Y_i themselves but may be extended to functions of the variables. For example, EY_i^2 means the average value of Y_i^2 over the entire population

* This section may be omitted without loss in continuity.

and $E(Y_i - 10)^2$ means the average value of the quantity $(Y_i - 10)^2$. The variance of a population may be defined as follows:

$$\sigma^2 = E(Y_i - \mu)^2 \tag{8-14}$$

That is, the variance of any population is the average value of the squares of deviations from the true mean.

Now, we are ready to state some rules for manipulating expected values:

1. The expected value of a constant is the constant itself:

$$Ec = c$$

2. The expected value of a constant times a variable is the constant times the expected value of the variable:

$$Ecy = cEy$$

3. The expected value of a sum is the sum of the expected values:

$$E(y_1 + y_2 + y_3) = Ey_1 + Ey_2 + Ey_3$$

4. The expected value of a product of two variables is the product of the expected values if the two variables are independent (or even uncorrelated):

$$Ey_1y_2 = Ey_1Ey_2 \qquad (y_1 \text{ and } y_2 \text{ are independent})$$

Here, "independence" is the concept defined in Sec. 2-3. That is, y_1 and y_2 are independent if $P(y_1|y_2) = P(y_1)$ or if $P(y_2|y_1) = P(y_2)$.

Using the above definitions and rules, we can easily prove some statements we have previously taken for granted. For example, to prove that $\bar{Y}$ is an unbiased estimate of μ, we write

$$\begin{aligned} E\bar{Y} &= E\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) \\ &= \frac{1}{n}E(Y_1 + Y_2 + \cdots + Y_n) \qquad \text{(by rule 2)} \\ &= \frac{1}{n}(EY_1 + EY_2 + \cdots + EY_n) \qquad \text{(by rule 3)} \\ &= \frac{1}{n}(\mu + \mu + \cdots + \mu) \qquad \text{(by definition of expected value)} \\ &= \frac{n}{n}\mu = \mu \end{aligned} \tag{8-15}$$

Also, we have said that the variance of a mean is equal to σ^2/n. This can be proved as follows:

$$\begin{aligned} \sigma_{\bar{Y}}^2 &= E(\bar{Y} - \mu)^2 \qquad \text{(by definition of the variance of a mean)} \\ &= E\left(\frac{Y_1 + Y_2 + \cdots + Y_n}{n} - \mu\right)^2 \qquad \text{(by definition of } \bar{Y}) \end{aligned}$$

$$\sigma_{\bar{Y}}^2 = E\left(\frac{Y_1 - \mu}{n} + \frac{Y_2 - \mu}{n} + \cdots + \frac{Y_n - \mu}{n}\right)^2$$

(by rearrangement of terms)

$$= \frac{1}{n^2} E[(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_n - \mu)]^2$$

(by factoring and rule 2)

$$= \frac{1}{n^2} E[(Y_1 - \mu)^2 + (Y_2 - \mu)^2 + \cdots + (Y_n - \mu)^2 + 2(Y_1 - \mu)(Y_2 - \mu) + \cdots + 2(Y_{n-1} - \mu)(Y_n - \mu)]$$

(by expansion)

$$= \frac{1}{n^2} [E(Y_1 - \mu)^2 + E(Y_2 - \mu)^2 + \cdots + E(Y_n - \mu)^2 + 2E(Y_1 - \mu)(Y_2 - \mu) + \cdots + 2E(Y_{n-1} - \mu)(Y_n - \mu)]$$

(by rule 3)

$$= \frac{1}{n^2} [\sigma^2 + \sigma^2 + \cdots + \sigma^2 + 0]$$

[by definition of the variance and by rule 4, that is, $E(Y_1 - \mu)(Y_2 - \mu) = E(Y_1 - \mu)E(Y_2 - \mu) = (\mu - \mu)(\mu - \mu) = 0$, by the independence assumption of the random sample of Y's]

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad (8\text{-}16)$$

Now, we are ready to consider a proof that is fundamental to the derivations which follow, namely, s^2 is an unbiased estimate of σ^2. We have

$$s^2 = \frac{\Sigma(Y_i - \bar{Y})^2}{n - 1}$$

$$(n - 1)s^2 = \Sigma(Y_i - \bar{Y})^2$$

$$E(n - 1)s^2 = E\Sigma[(Y_i - \mu) - (\bar{Y} - \mu)]^2$$

(by adding and subtracting μ)

$$= E[\Sigma(Y_i - \mu)^2 - 2\Sigma(Y_i - \mu)(\bar{Y} - \mu) + n(\bar{Y} - \mu)^2]$$

But $\Sigma(Y_i - \mu) = n(\bar{Y} - \mu)$, so we can combine the second and third terms:

$$E(n - 1)s^2 = E\Sigma(Y_i - \mu)^2 - nE(\bar{Y} - \mu)^2$$

$$= n\sigma^2 - n\frac{\sigma^2}{n}$$

(by rule 3 and the definition of the variance of Y and of $\bar{Y}$)

$$= (n - 1)\sigma^2 \qquad (8\text{-}17)$$

Therefore the expected value of s^2 is equal to σ^2.

Now, suppose we draw a sample of size n from an infinite population with mean μ and variance σ^2. (An infinite population is assumed in all of the above proofs, although some are true for finite populations as well.) Then suppose we group the n observations at random into t classes, with n_1 in the first group, n_2 in the second group, and so forth, so that $n_1 + n_2 + \cdots + n_t = n$. We shall let Y_{ij} denote the jth observation in the ith group, $\bar{Y}_{i.}$ denote the mean of the ith group, and $\bar{Y}_{..}$ the mean of the entire sample of n observations. Then a deviation of an observation from the over-all mean (which we shall call the *grand mean*) can be expressed as the deviation of the observation from the group mean plus the deviation of the group mean from the grand mean. That is,

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..}) \tag{8-18}$$

That is just an algebraic identity. If we square both sides, we obtain

$$(Y_{ij} - \bar{Y}_{..})^2 = (Y_{ij} - \bar{Y}_{i.})^2 + 2(Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

Now, we sum on the j subscript. That is, we add up the squares within each of the groups:

$$\Sigma_j(Y_{ij} - \bar{Y}_{..})^2 = \Sigma_j(Y_{ij} - \bar{Y}_{i.})^2 + 2(\bar{Y}_{i.} - \bar{Y}_{..})\Sigma_j(Y_{ij} - \bar{Y}_{i.}) + n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

But $\Sigma_j(Y_{ij} - \bar{Y}_{i.}) = 0$, so the middle term disappears. Hence, adding over the i subscript, that is, adding up the sums of squares over all the groups, we obtain

$$\Sigma_{ij}(Y_{ij} - \bar{Y}_{..})^2 = \Sigma_{ij}(Y_{ij} - \bar{Y}_{i.})^2 + \Sigma_i n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \tag{8-19}$$

It will be seen that the left-hand side is the total sum of squares, the first term on the right is the sum of squares within columns [the numerator of (8-2)], and the last term is the sum of squares among column means [the numerator of Eq. (8-3)]. Thus we have verified that the sums of squares of the components add to the total sum of squares.

From the standpoint of statistical analysis it is important to examine the expected values of the sums of squares in (8-19). Since, by assumption, all n observations were drawn independently from the same population, the expected value of the left-hand side is given by (8-17), that is, $(n - 1)\sigma^2$.

Now, consider the first term on the right. We have

$$\begin{aligned} E\Sigma_{ij}(Y_{ij} - \bar{Y}_{i.})^2 &= \Sigma_i\Sigma_j E(Y_{ij} - \bar{Y}_{i.})^2 \qquad \text{(by rule 3)} \\ &= \Sigma_i(n_i - 1)\sigma^2 \qquad \text{[by (8-17)]} \\ &= (\Sigma n_i - t)\sigma^2 = (n - t)\sigma^2 \end{aligned} \tag{8-20}$$

where t is the number of classes (columns).

Now, consider the second term on the right:

$$
\begin{aligned}
E\Sigma_i n_i(Y_{i.} - Y_{..})^2 &= E\Sigma_i n_i[(\bar{Y}_{i.} - \mu) - (\bar{Y}_{..} - \mu)]^2 \\
&\qquad \text{(adding and subtracting } \mu\text{)} \\
&= E\Sigma_i n_i[(\bar{Y}_{i.} - \mu)^2 - 2(\bar{Y}_{..} - \mu)(\bar{Y}_{i.} - \mu) + (\bar{Y}_{..} - \mu)^2] \\
&= E[\Sigma_i n_i(\bar{Y}_{i.} - \mu)^2 - n(\bar{Y}_{..} - \mu)^2] \\
&\qquad \text{(the middle term combines with the last upon summation)} \\
&= \Sigma_i n_i E(\bar{Y}_{i.} - \mu)^2 - nE(\bar{Y}_{..} - \mu)^2 \\
&= \Sigma_i n_i \frac{\sigma^2}{n_i} - n\frac{\sigma^2}{n} \qquad \text{[by (8-16)]} \\
&= t\sigma^2 - \sigma^2 \qquad \text{(because } \Sigma_i \sigma^2 = t\sigma^2\text{)} \\
&= (t-1)\sigma^2 \qquad (8\text{-}21)
\end{aligned}
$$

Therefore, the expectation of (8-19) can be written:

$$(n-1)\sigma^2 = (n-t)\sigma^2 + (t-1)\sigma^2 \qquad (8\text{-}22)$$

This shows that the degrees of freedom add up in the same manner as the sums of squares. That is, $(n-t) + (t-1) = n-1$. Also, it shows that the expected value of the within-columns mean square is σ^2, which is also the expected value of the among-columns mean square. These are, in fact, two independent estimates of the variance σ^2. Remember that this is under the assumption that all n observations came from the same population. Therefore, the ratio of among-columns mean square to between-columns mean square is distributed as F.

Now, if we assume that the observations in the ith column have the mean $\mu + T_i$, instead of μ, we must reexamine the expectation of (8-19). That is, we now consider the case in which the means of the columns differ. We assume that the variances around these means still equal σ^2 and recall that for convenience we have assumed $\Sigma T_i = 0$. We shall also assume that all the groups (columns) have the same number of observations, say, m, so that the total number of observations is $n = ml$. This assumption is made to simplify the algebra. With equal group numbers the formula for the subdivision of sums of squares is

$$\Sigma_{ij}(Y_{ij} - \bar{Y}_{..})^2 = \Sigma_{ij}(Y_{ij} - \bar{Y}_{i.})^2 + m\Sigma_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \qquad (8\text{-}23)$$

Consider the expected value of the first term on the right:

$$
\begin{aligned}
E\Sigma_{ij}(Y_{ij} - \bar{Y}_{i.})^2 &= E\Sigma_{ij}[(Y_{ij} - \mu - T_i) - (\bar{Y}_{i.} - \mu - T_i)]^2 \\
&= E\Sigma_{ij}(Y_{ij} - \mu - T_i)^2 - mE\Sigma_i(\bar{Y}_{i.} - \mu - T_i)^2 \\
&= mt\sigma^2 - \frac{tm\sigma^2}{m} = t(m-1)\sigma^2 \qquad (8\text{-}24)
\end{aligned}
$$

Now, consider the second term on the right:

$$\begin{aligned}
Em\Sigma_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 &= mE\Sigma_i[(\bar{Y}_{i.} - \mu - T_i) - (\bar{Y}_{..} - \mu - T_i)]^2 \\
&= m\Sigma_i E(\bar{Y}_{i.} - \mu - T_i)^2 \\
&\quad - 2mE\Sigma_i(\bar{Y}_{..} - \mu - T_i)(\bar{Y}_{i.} - \mu - T_i) \\
&\quad + mE\Sigma_i(Y_{..} - \mu - T_i)^2 \\
&= m\frac{t\sigma^2}{m} - 2mtE(\bar{Y}_{..} - \mu)^2 + mE\Sigma_i[(\bar{Y}_{..} - \mu) - T_i]^2 \\
&= t\sigma^2 - 2mtE(\bar{Y}_{..} - \mu)^2 + mtE(\bar{Y}_{..} - \mu)^2 + m\Sigma_i T_i^2 \\
&= t\sigma^2 - mt\frac{\sigma^2}{mt} + m\Sigma_i T_i^2 \\
&\quad \text{(note that } E(\bar{Y}_{..} - \mu)^2 \text{ is the variance of a mean of} \\
&\quad\ mt \text{ observations. The } T_i\text{'s sum to zero)} \\
&= (t-1)\sigma^2 + m\Sigma_i T_i^2 \qquad (8\text{-}25)
\end{aligned}$$

We note, then, that the within-columns mean square is an estimate of σ^2 [from (8-24)] and the among-columns mean square is an estimate of $\sigma^2 + [m/(t-1)]\Sigma_i T_i^2$ [from (8-25)]. Therefore, any difference (other than random variation) between these two mean squares must be attributed to the factor $\Sigma_i T_i^2$. Thus, the F test is a test of the hypothesis that $T_1 = T_2 = \cdots = T_t = 0$.

We shall now derive the computing formulas from (8-19). The total sum of squares is

$$\begin{aligned}
\Sigma_{ij}(Y_{ij} - \bar{Y}_{..})^2 &= \Sigma_{ij}(Y_{ij}^2 - 2\bar{Y}_{..}Y_{ij} + \bar{Y}_{..}^2) \\
&= \Sigma_{ij}Y_{ij}^2 - 2\bar{Y}_{..}\Sigma_{ij}Y_{ij} + n\bar{Y}_{..}^2 \\
&= \Sigma_{ij}Y_{ij}^2 - 2\bar{Y}_{..}n\bar{Y}_{..} + n\bar{Y}_{..}^2 \\
&= \Sigma_{ij}Y_{ij}^2 - n\bar{Y}_{..}^2 = \Sigma_{ij}Y_{ij}^2 - \frac{(\Sigma_{ij}Y_{ij})^2}{n}
\end{aligned}$$

which is the computing formula (8-4). For the sum of squares among column means we have

$$\begin{aligned}
\Sigma_i n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 &= \Sigma_i n_i(\bar{Y}_{i.}^2 - 2\bar{Y}_{..}\bar{Y}_{i.} + \bar{Y}_{..}^2) \\
&= \Sigma_i n_i\bar{Y}_{i.}^2 - 2n\bar{Y}_{..}^2 + n\bar{Y}_{..}^2 \\
&= \Sigma_i\frac{(\Sigma_j Y_{ij})^2}{n_i} - \frac{(\Sigma_{ij}Y_{ij})^2}{n}
\end{aligned}$$

which is the computing formula listed as (8-5).

Although the above discussion has been limited to the simple analysis of variance, similar methods applied to the two-way classification and more complex designs will yield the appropriate mean squares and computing formulas. An extension of the algebra to these cases is not of vital concern here, however.

8-6. A FURTHER NOTE ON STATISTICAL DESIGNS

It should be clear to the student that there is no necessity to limit one's investigations to two factors. For example, in the machines-operators illustration one might be interested in time of day as another variable. In this event one must consider:

Three main effects:

Time of day
Operator
Temperature

Three two-factor interactions:

Time × operators
Time × temperatures
Operators × temperatures

One three-factor interaction:

Time × operators × temperatures

The combination of factors can be continued indefinitely. With many factors it becomes difficult or impossible to compare all treatments within one block of experimental material. This situation leads to *confounding*, a process by which selected comparisons are made within each block and information abandoned on other comparisons. It is frequently possible to confound the higher-order interactions and to retain full information

Table 8-13. A Systematic Latin Square with One Randomized Form

Lots	Operators				
	1	2	3	4	5
1	*A*	*B*	*C*	*D*	*E*
2	*E*	*A*	*B*	*C*	*D*
3	*D*	*E*	*A*	*B*	*C*
4	*C*	*D*	*E*	*A*	*B*
5	*B*	*C*	*D*	*E*	*A*

Systematic Plan

Lots	Operators				
	1	2	3	4	5
1	*D*	*E*	*C*	*A*	*B*
2	*E*	*A*	*D*	*B*	*C*
3	*C*	*D*	*B*	*E*	*A*
4	*B*	*C*	*A*	*D*	*E*
5	*A*	*B*	*E*	*C*	*D*

Randomized Plan

on main effects and on some or all of the two-factor interactions, so that the result is not at all detrimental to the interpretation of the data.

Another device used frequently in the analysis of complex factorials is to assume that the higher-order interactions are actually zero, so that the computed interactions (from analysis of the data) constitute a valid estimate of experimental error. This assumption, when valid, avoids the

necessity for replication and, in fact, may make it possible to estimate main effects and important interactions from only a portion of the total possible treatment effects. The latter process is called *fractional replication.*

We have seen how randomized blocks can be used to compare treatments within a common lot of experimental material. There are situations in which one wishes to control two sets of environmental conditions simultaneously. The result is a *Latin-square* design. For example, suppose that one wishes to test 5 types of cutting tools and that, in order to control the influence of operators and lots of material, he arranges the experiment in such a manner that each cutting tool is used once on each lot of material and once by each operator. The Latin-square arrangement is shown in Table 8-13.*

For the analysis, one takes out the effect of operators and lots and tests treatments against residual. The skeleton analysis of variance is shown in Table 8-14.

Table 8-14. Subdivision of Degrees of Freedom for Latin Square

Due to	*df*
Operators	4
Lots	4
Treatments	4
Residual (error)	12
Total	24

It was pointed out in Sec. 8-2 that estimates of the parameters in analysis-of-variance models can be accomplished by customary least-squares techniques. This observation might lead one to think that there is no place in statistics for a separate study of the analysis of variance. Certainly this is not true. The partitioning of sums of squares into meaningful components is worthy of study in its own right and may be considered one of the principal advances in statistical methodology in the twentieth century. Furthermore, it is inextricably intertwined with the principles of experimental design. For example, the randomized-blocks design presented in this chapter has all the treatments repeated in every block. This assures us that comparisons among all the treatments can be made with equal precision, and we say that the design is "balanced." Even if one decides to confound certain comparisons in order to reduce size of blocks, he retains balance among the rest of the compari-

* For the actual process of randomization the reader is referred to Ronald A. Fisher and Frank Yates, *Statistical Tables for Biological and Medical Research,* 4th ed., Oliver & Boyd, Ltd., Edinburgh, 1953.

sons. This retention of balance, which simplifies mechanical partitioning of sums of squares, is properly a part of the study of experimental design, but the partitioning process is, in fact, the analysis of variance.

It should be clear to the reader at this point that a single chapter on analysis of variance is an inadequate treatment of the subject matter for the person who is required to use the technique as a tool. The aim in this chapter has been to show the elementary principles of the analysis and to impress the reader with the power of the tool. The reader interested in a more adequate treatment of the subject is referred to the following references:

George W. Snedecor, *Statistical Methods*, 5th ed., Iowa State College Press, Ames, Iowa, 1956.

Bernard Ostle, *Statistics in Research*, Iowa State College Press, Ames, Iowa, 1954.

William G. Cochran and Gertrude M. Cox, *Experimental Designs*, 2d ed., John Wiley & Sons, Inc., New York, 1957.

EXERCISE 8-14. Suppose that the data resulting from an experiment organized according to the plan of Table 8-14 are as follows:

Lots	Operators				
	1	2	3	4	5
1	$D = 5$	$E = 16$	$C = 13$	$A = 8$	$B = 7$
2	$E = 12$	$A = 4$	$D = 11$	$B = 10$	$C = 5$
3	$C = 12$	$D = 13$	$B = 10$	$E = 20$	$A = 5$
4	$B = 3$	$C = 8$	$A = 5$	$D = 10$	$E = 14$
5	$A = 7$	$B = 7$	$E = 15$	$C = 11$	$D = 9$

Perform an analysis of variance and test the hypothesis that the averages for the cutting tools are equal.

EXERCISE 8-15. Subtract 4 from each observation in Table 8-11 and recompute. Check answers with those already given.

EXERCISE 8-16. Independent observations on 2 samples are as follows:

Sample 1	*Sample 2*
4	11
9	9
8	12
6	13
7	8
5	7
	9
	12

Test the hypothesis that $\mu_1 = \mu_2$ by the analysis of variance.

EXERCISE 8-17. Refer to Exercise 8-16. Test the hypothesis that $\mu_1 = \mu_2$ by the t test. Note that $t^2 = F$.

EXERCISE 8-18. Assume that you have the following paired observations:

Pair	Test 1	Test 2
1	18	16
2	25	22
3	29	24
4	28	21
5	19	19

Test the hypothesis that $\mu_1 = \mu_2$ by the analysis of variance.

EXERCISE 8-19. Refer to Exercise 8-18. Test the hypothesis that $\mu_1 = \mu_2$ by the t test for paired observations. Note that $t^2 = F$.

9. ANALYSIS OF TIME SERIES

9-1. INTRODUCTION

In some ways, time-series analysis is the most exasperating area of statistical analysis, and the least satisfying from the theoretical viewpoint. A large part of the difficulty is due to the nature of time-series observations. They are not random drawings from a population, but are, instead, ordered observations over time. Another way of expressing the situation is to say that time has entered as another parameter.

A time series is simply a one-way classification with time as the variable of classification. Examples are (1) number of automobiles sold, by months, by a dealer (sales in dollars might also be used), (2) hourly temperatures in a locker storage plant (here a recording is made for a particular instant in time), and (3) ratio of current liabilities to current assets at year end, for a series of years. From these examples, it may be seen that the variable may be either counted (discrete) or measured (continuous). The recording may represent either an interval of time (in the case of sales) or the situation at a given instant in time (e.g., hourly temperatures). There is, of course, the case in which a reading, or recording, is made for every instant of time. An example is a continuous recording drum for temperatures.

The philosophy underlying all time-series analysis is that the variable is a function of time and that classifying the data according to time of occurrence is meaningful.

9-2. THE CONVENTIONAL COMPONENTS OF TIME SERIES

Usually the classification of data by time periods serves only as a starting point for further analysis. Often an attempt is made to break down the data into component parts to assist (1) in understanding what has happened in the past and (2) in predicting the future. Unfortunately,

such analysis often proves quite ineffective, so that the second objective is seldom reached. Much more is said about this point later. Let us, for the moment, concentrate on description of past events.

Historically, it has been assumed that a time series is composed of four elements: (1) trend, (2) seasonal, (3) cycle, and (4) random. The *trend* may be thought of as a tendency for the data to increase or decrease over a long period of time. There is no implication that the increase or decrease is constant. It may be, of course, in which case the trend forms a straight line, but it may, in fact, follow complex curvilinear patterns. For example, population in the United States has certainly tended to increase over the past century and a half, but it is well known that this trend has not followed a straight line (see Fig. 9-1).

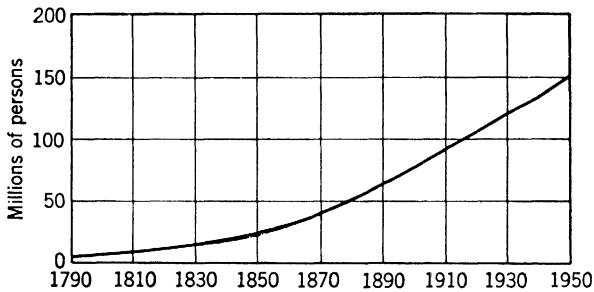

FIG. 9-1. Population of the United States (*Statistical Abstract of the United States*)

A *seasonal* pattern is one which recurs regularly over time. For example, there is a more or less regular pattern throughout each year in employment in the construction industry. This regular pattern constitutes the seasonal element in these particular data. The word "seasonal" is unfortunate, perhaps, because it seems to imply a relationship to the seasons of the year. Actually, a seasonal pattern may be weekly (in the case of department-store sales) or daily (in the case of pedestrian traffic in a business district).

Cycle is usually considered to be a fairly long term movement which is less persistent than trend. It recurs at longer intervals than the seasonal pattern, and the periods of recurrence tend to be somewhat irregular. It is difficult to introduce the idea of a cycle without reference to that much-used and ill-defined concept "the business cycle." A cycle may, in fact, have no relationship to the business cycle, but if the student feels the need for leaning on an old familiar crutch, he may think of the business cycle as a special case of cycles in general.

A *random* movement, conceptually, is one which contributes period-to-period variation. The name "random" implies that each period's contribution is completely unpredictable. In practice, that part of the data which the statistician attributes to random movement is seldom

really random. It is usually data which the statistician cannot "explain" in terms of trend, seasonal pattern, and cycle.

Sometimes a fifth component of time series is proposed. It may be observed that a series may change its level or direction suddenly. For example, a large industry in a community may shut down, with resulting immediate drops to new levels of employment and gross business income. This abrupt type of change is sometimes called a *structural* change.

9-3. TIME-SERIES MODELS USING THE FOUR CONVENTIONAL COMPONENTS

There are two conventional time-series models, using the four components described above. *The additive model* assumes that the four components can be added together to get the value of the data for a given point in time.

Let Y_i = the data at the ith time period
T_i = the trend at the ith time period
S_i = the seasonal at the ith time period
C_i = the cycle at the ith time period
R_i = the random at the ith time period

then the additive model is as follows:

$$Y_i = T_i + S_i + C_i + R_i \tag{9-1}$$

A little reflection will convince one that, given n observations for n periods in time, one cannot solve uniquely for all the constants T_i, S_i, and C_i. This approach would, in fact, amount to trying to solve n equations for $3n$ parameters. Even if one assumes that seasonal is constant from year to year and that trend contains only two parameters (as in the case of a straight line), one still must solve n equations for $n + 14$ parameters (assuming monthly data). This is again a hopeless task. Clearly, then, some compromise must be made, and the necessity for compromise gives rise to the myriad of methods for finding the several components.

An example of an additive model is given in Table 9-1. It has been assumed that there is a straight-line trend which increases by one unit per quarter, that seasonal has a constant pattern throughout the four seasons of the year, and that cycle has a known pattern also. The random element was introduced by drawing numbers from a table of random normal numbers.

Ordinarily, of course, we are given only the last column, Y_i. We try to arrive at estimates of the figures in the T_i, S_i, and C_i columns. Suppose we first try to estimate trend, using only the data of the last column. We shall assume that the trend is a straight line and will proceed to fit a

Table 9-1. Illustration of an Additive Time-series Model

Year and quarter		Trend, T_i	Seasonal, S_i	Cycle, C_i	Random, R	Data, Y_i $(T_i + S_i + C_i + R_i)$
1952	1	140	5	−13	−2	130
	2	141	−10	−12	8	127
	3	142	−3	−10	6	135
	4	143	8	−5	−6	140
1953	1	144	5	−2	−15	132
	2	145	−10	−1	−5	129
	3	146	−3	0	2	145
	4	147	8	2	1	158
1954	1	148	5	4	−4	153
	2	149	−10	8	−1	146
	3	150	−3	12	5	164
	4	151	8	16	−5	170
1955	1	152	5	12	1	170
	2	153	−10	9	0	152
	3	154	−3	5	3	159
	4	155	8	4	1	168
1956	1	156	5	3	1	165
	2	157	−10	1	−3	145
	3	158	−3	0	−2	153
	4	159	8	−2	1	166
1957	1	160	5	−4	−2	159
	2	161	−10	−5	6	152
	3	162	−3	−10	4	153
	4	163	8	−12	4	163

line through the data. There are, of course, many ways of doing this, the easiest of which is to plot the data and simply draw a straight line which appears to fit the pattern of the data. We shall use a slightly more complex method, called the *method of averages.* We proceed as follows:

1. Assign X values to the time periods, letting the first quarter of 1952 have the value 1, the second quarter the value 2, and so forth through the 6 years of data.
2. Divide the data into two equal parts of 3 years each (refer to Table 9-2).
3. Average the X's and the Y's in each part to find $\bar{X}_1$, $\bar{Y}_1$, $\bar{X}_2$, and $\bar{Y}_2$.
4. Let the trend line be denoted by $Y = a + bX$, where a and b are the constants described in Sec. 7-2.

Table 9-2. Computation of Trend by the Method of Averages

Year and quarter	X_i	Y_i	Estimated trend, $\hat{T}_i$	Actual trend, T_i
1952 1	1	130	137.3	140
2	2	127	138.5	141
3	3	135	139.8	142
4	4	140	141.0	143
1953 1	5	132	142.2	144
2	6	129	143.4	145
3	7	145	144.7	146
4	8	158	145.9	147
1954 1	9	153	147.1	148
2	10	146	148.3	149
3	11	164	149.5	150
4	12	170	150.8	151
Average	6.5	144.08		
1955 1	13	170	152.0	152
2	14	152	153.2	153
3	15	159	154 4	154
4	16	168	155.7	155
1956 1	17	165	156.9	156
2	18	145	158.1	157
3	19	153	159.3	158
4	20	166	160.5	159
1957 1	21	159	161.8	160
2	22	152	163.0	161
3	23	153	164.2	162
4	24	163	165.4	163
Average	18.5	158.75		
Grand average	12.5	151.42		

$$b = \frac{\bar{Y}_2 - \bar{Y}_1}{\bar{X}_2 - \bar{X}_1} = \frac{14.67}{12}$$

$$= 1.222$$

$$a = \bar{Y} - b\bar{X}$$

$$= 151.42 - 1.222(12.5)$$

$$= 136.1$$

$$\text{Trend} = 136.1 + 1.222X$$

(quarterly data, origin fourth quarter, 1951)

*

5. The quantities computed in (3) above constitute two points through which the trend line will pass. The points can be described by the coordinates $(\bar{X}_1, \bar{Y}_1)$ and $(\bar{X}_2, \bar{Y}_2)$.
6. Substitute these quantities into the trend equation for X and Y to form the two equations

$$\bar{Y}_1 = a + b\bar{X}_1$$
$$\bar{Y}_2 = a + b\bar{X}_2$$

7. Solve simultaneously to find a and b. The solutions are

$$b = \frac{\bar{Y}_2 - \bar{Y}_1}{\bar{X}_2 - \bar{X}_1}$$
$$a = \bar{Y} - b\bar{X}$$

where $\bar{Y}$ is the average of all Y's and $\bar{X}$ is the average of all X's. The computations are shown following Table 9-2.

8. Write the trend equation and label it. That is, state that the data are quarterly and the origin (the time when X is zero) is at the fourth quarter of 1951.
9. Compute the trend by inserting the constant $a = 136.1$ in the adding or calculating machine and adding one "trend increment," b, for each quarter thereafter.

The computed trend as well as the actual trend is shown in Table 9-2. Note that the computed trend is too low at the beginning and too high at the end. In an actual problem, of course, one does not know the real value of the trend.

Our next step is to estimate the seasonal constants. We reason that, if we subtract trend from $Y_i = T_i + S_i + C_i + R_i$, we shall have only seasonal, cycle, and random left. Then we can eliminate cycle and random by averaging. We shall, of course, have a "trend error," which we shall ignore.

Table 9-3. Computation of Absolute Seasonal by Averaging

Year	Quarter			
	1	2	3	4
1952	−7.3	−11.5	−4.8	−1.0
1953	−9.8	−14.4	0.3	12.1
1954	5.9	−2.3	14.5	19.2
1955	18.0	1.2	4.6	12.3
1956	8.1	−13.1	−6.0	5.5
1957	−2.8	−11.0	−11.2	−2.4
Average	2.02	−8.92	−0.43	7.62
Adjusted average	2.0	−9.0	−0.5	7.5

1. Average the data by quarters.
2. Add the quarterly averages together. If the sum is 0, the averages are the seasonal amounts. If the sum is not 0, divide the sum by the number of periods in a year (4 in this case), and add this amount algebraically to the quarterly averages.
3. Round to whatever accuracy is desired. The result is the seasonal in absolute amounts (same units as the original data).

The computations of Table 9-3 compare with the actual seasonal pattern as follows:

Quarter	Actual	Estimated
1	5	2.0
2	−10	−9.0
3	−3	−0.5
4	8	7.5

Now we require an estimate of the cycle. We reason that, if we subtract trend and seasonal from $Y_i + S_i + C_i + R_i$, we shall have only cycle and random left. This time we have a seasonal error left, which,

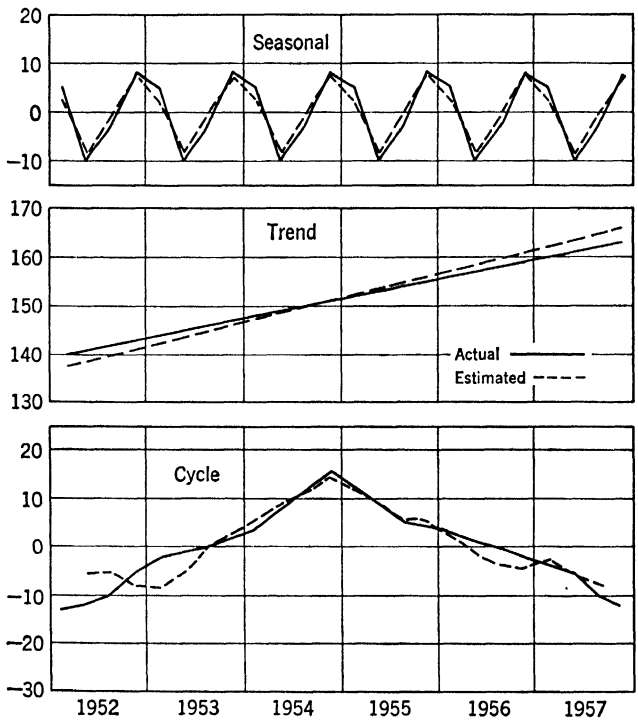

FIG. 9-2. Actual and estimated components of artificially constructed time series

again, we shall ignore. We can eliminate the random element by some sort of smoothing process. The smoothing process we shall employ is that of moving averages. That is, we take the average of the first three figures; the average of the second, third, and fourth; the average of the third, fourth, and fifth; and so on. The computations are shown in

Table 9-4. Computation of Cycle by Smoothing of Data Minus Trend Minus Seasonal

Year and quarter		Data	Trend, $\hat{T}_i$	Seasonal, $\hat{S}_i$	Crude cycle, $Y_i - \hat{T}_i - \hat{S}_i$	3-point moving average	Actual cycle
1952	1	130	137.3	2.0	−9.3		−13
	2	127	138.5	−9.0	−2.5	−5.4	−12
	3	135	139.8	−0.5	−4.3	−5.1	−10
	4	140	141.0	7.5	−8.5	−8 3	−5
1953	1	132	142.2	2.0	−12.2	−8.7	−2
	2	129	143.4	−9.0	−5.4	−5.6	−1
	3	145	144.7	−0.5	0.8	0	0
	4	158	145.9	7.5	4.6	3.1	2
1954	1	153	147.1	2.0	3.9	5.1	4
	2	146	148.3	−9.0	6.7	8.5	8
	3	164	149.5	−0.5	15.0	11.1	12
	4	170	150.8	7.5	11.7	14.2	16
1955	1	170	152.0	2.0	16.0	11.8	12
	2	152	153.2	−9.0	7.8	9.6	9
	3	159	154.4	−0.5	5.1	5.9	5
	4	168	155.7	7.5	4.8	5.3	4
1956	1	165	156.9	2.0	6.1	2.3	3
	2	145	158.1	−9.0	−4.1	−1.3	1
	3	155	159.3	−0.5	−5.8	−4.0	0
	4	166	160.5	7.5	−2.0	−4.2	−2
1957	1	159	161.8	2.0	−4.8	−2.9	−4
	2	152	163.0	−9.0	−2.0	−5.8	−5
	3	153	164.2	−0.5	−10.7	−7.5	−10
	4	163	165.4	7.5	−9.9		−12

Table 9-4, as well as a comparison with the actual cycle (which, of course, is unknown in a real problem).

The estimated components of the time series (trend, cycle, and seasonal) are compared with the actual components in Fig. 9-2. It will be observed that trend was underestimated at the beginning and overestimated at the end. This error is reflected in the estimation of cycle. It is difficult to distinguish between trend and cycle, and sometimes the

Table 9-5. Forecast of Artificially Constructed Series for 1958

Year and quarter	Trend	Seasonal	Cycle	Total
1958 1	166.6	2.0	−10	159
2	167.8	−9.0	−10	149
3	169.1	−0.5	−10	159
4	170.3	7.5	−10	168

problem is avoided by not trying to distinguish the two. More will be said about this later.

Now, suppose one wishes to forecast the series for the 4 quarters of the next year. We project trend and seasonal ahead for 1 year and guess at the cyclical pattern. Suppose we anticipate that cycle has about reached the bottom of the downward swing. We may guess that cycle will hold steady at −10, which is about the average of the last two crude cycle figures. Then our estimates for 1958 will be as shown in Table 9-5. The estimation of cycle is likely to be subject to the greatest error. More will be said about forecasting this component later.

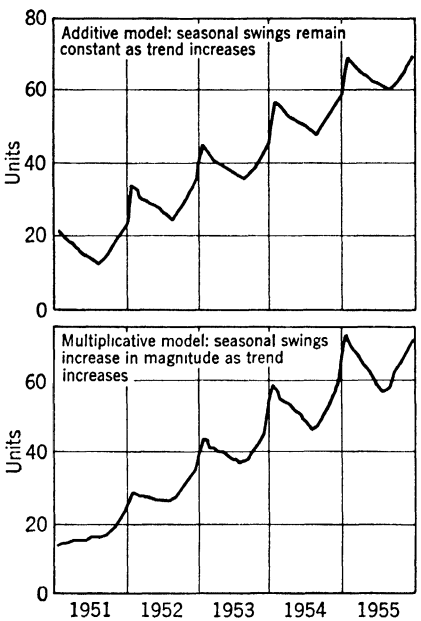

FIG. 9-3. Comparison of additive and multiplicative time-series models

EXERCISE 9-1. Suppose we have a monthly-trend equation for sales as follows: $Y = 1{,}460 + 1.4X$ (monthly data, origin Jan. 15, 1952). (*a*) In the absence of any other information, what value of the series do you predict for August, 1958? (*b*) For December, 1958? (*c*) If you know that August sales are usually off by 50 and that December sales are up by 200, what are your predictions for August and December, 1958?

The multiplicative model assumes that the four components are related as follows:

$$Y_i = T_i \times S_i \times C_i \times R_i \qquad (9\text{-}2)$$

Here one of the components is in units of the original data (usually trend),

and the remaining components are one-based relatives, that is, ratios whose average value is 1, or unity. The same problem arises if one tries to find a unique solution. It is clear that taking the logarithms of both sides of (9-2) will make the expression linear in the four components. That is,

$$\log Y_i = \log T_i + \log S_i + \log C_i + \log R_i \qquad (9\text{-}3)$$

This model can be solved in the same manner as (9-1). The only difference is that the solutions (estimates of the Y_i) are in terms of logarithms and must be converted back to natural numbers by taking antilogarithms in order to achieve the final result.

Generally one distinguishes the multiplicative model from the additive model given earlier when it is observed that seasonal swings increase with the trend. Two situations are shown in Fig. 9-3. In the upper curve the *amount* of seasonal variation remains constant and calls for solution by the linear model. In the second the *ratio of seasonal to trend* remains constant and calls for solution by the multiplicative model. In this illustration the magnitude of the seasonal swings is proportional to the general level of the trend.

Most of the remainder of this chapter is devoted to estimation of the time-series components when one or the other of the above models is assumed.

EXERCISE 9-2. The following are figures on Portland cement production, in millions of barrels, in the United States over a period of 7 years.

Month	1950	1951	1952	1953	1954	1955	1956
Jan.	15.2	17.4	17.0	18.9	17.8	20.2	21.4
Feb.	13.1	15.2	16.5	17.3	16.9	17.6	19.6
Mar.	14.3	18.7	18.1	20.2	20.1	22.3	23.4
Apr.	18.1	20.2	19.8	21.8	21.7	24 8	26.1
May	19.9	21.9	21.8	23.3	23.3	27.0	29.6
June	20.0	22.0	20.7	22.7	22.8	26.8	28.8
July	20.7	22.4	21.3	24.1	25.5	27.3	29.5
Aug.	21.9	22.5	23.6	24.3	25.7	27.9	30.1
Sept.	20.9	22.3	23.0	23.8	25.5	27.0	28.6
Oct.	22.5	22.8	24.2	24.7	25.9	27.9	29.1
Nov.	20.2	20.7	22.0	22.5	23.8	24.9	25.9
Dec.	19.1	19.9	20.9	20.2	22.3	23.1	24.4
Total	225.9	246.0	248.9	263.8	271.3	296.8	316.5

Plot the above monthly data on graph paper and answer the following questions: (*a*) Do you think the additive model or the multiplicative model is appropriate? Why? (*b*) Would you try to separate trend and cycle? Why?

9-4. FITTING OF STRAIGHT LINES BY LEAST SQUARES

In the previous section a straight line was fitted by the method of averages. It should be clear that there are other ways of fitting straight lines. Also, it should be clear that a straight line will not always be a proper representation of trend, so that we must devote some attention to the selection of a suitable trend equation as well as to its method of computation.

We turn our attention first to an alternative way of fitting a straight line. If the trend component can be expressed properly by a straight line of the form

$$Y_i = A + BX_i + \epsilon_i \tag{9-4}$$

where A and B are constants, X_i is the time variable (coded in any desired manner), and ϵ_i is an independent random variable with 0 mean

Table 9-6. Computation of Trend by the Method of Least Squares, Arbitrary Data

Year	Annual total		XY	X^2
	X	Y		
1952	−2.5	532	−1,330 0	6.25
1953	−1.5	564	−846 0	2.25
1954	−0.5	635	−316.5	0.25
1955	0.5	649	324.5	0.25
1956	1.5	629	943.5	2.25
1957	2.5	627	1,567.5	6.25
Total	0	3,634	343.0	17.50

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad \text{(see Sec. 7-3)}$$

$$= \frac{6(343)}{6(17.5)} = 19.6$$

$$a = \bar{Y} - b\bar{X} = 605.67 - 19.6(0) = 605.67$$

$$\text{Trend} = 605.67 + 19.6X$$

(annual units, origin Jan. 1, 1955)

and constant variance, then the method of least squares will yield best linear unbiased estimates of the constants A and B. The situation then is exactly as discussed in Sec. 7-3.

In practice, the assumptions of (9-4) are seldom met. For one thing, the raw data, to which one fits the trend line, have a cycle component, so that ϵ_i contains cycle as well as random. Therefore the ϵ_i do not meet the criteria of being independent random variables with constant vari-

ance. Since the necessary assumptions are not met, we cannot say that the least-squares trend line is "better" than any other trend line. However, statisticians have become accustomed to using least squares, and the method at least has the advantage of objectivity, so it is frequently used.

Since seasonal variation is a repetitive pattern from year to year, it contributes nothing to the estimation of trend. We may as well, then, use annual totals of data rather than monthly or quarterly data in the computation of trend. Also, since the values that we assign to the time variable X are arbitrary, we may as well place the value 0 (the origin) in the middle of the series. This procedure simplifies the computation.

The calculation of trend by the method of least squares for the artificially constructed data of the previous section is shown in Table 9-6. It will be noted that, if we have an odd number of years, the X variable is somewhat simpler to handle, since the middle year has the value 0.

The resulting trend equation is in annual units. We should like to convert it to quarterly units. We may do so by dividing the computed constant a by 4 and the value of b by 16. (We divide a and b by 4 because the data are sums of 4 quarters, and we divide b by 4 again so that X's will be in quarters as well.) If we wish to convert to a monthly-trend equation, we divide a by 12 and b by 144. The quarterly-trend equation is

$$\text{Trend} = 151.42 + 1.225X$$

(quarterly data, origin Jan. 1, 1955)

We have not disturbed the origin by the above procedure. In order to compare the trend equation with that computed earlier by the method of averages, we shall shift the origin back $12\frac{1}{2}$ quarters to the middle of the fourth quarter of 1951. (Follow this through carefully.) One accomplishes the shift by substituting $X - 12.5$ for X in the trend equation and simplifying. The results are

$$\begin{aligned}\text{Trend} &= 151.42 + 1.225(X - 12.5)\\ &= 151.42 - 15.31 + 1.225X\\ &= 136.11 + 1.225X\end{aligned}$$

(quarterly data, origin fourth quarter, 1951)

EXERCISE 9-3. (*a*) Using the data of Exercise 9-2, compute a straight-line trend, by the method of least squares, to the annual totals, change to monthly units, and shift the origin to January, 1950. (*b*) Compute the trend values for each month.

9-5. NONLINEAR TRENDS

There are many trends other than straight lines which may be used to fit particular types of data. Figure 9-4 shows some of the common types.

Polynomial trends are of the form

$$Y = A + BX + CX^2 + DX^3 + \cdots + JX^k \qquad (9\text{-}5)$$

The straight line is a special case, having only the first two terms on the right of the equality sign. With three terms on the right, we have a cubic, and so forth. Typical forms are shown in Fig. 9-4. Generally speaking, it is unwise to fit a high-degree polynomial to the data because, if one does, he is almost sure to be mixing trend and cycle. Also, a glance at Fig. 9-4 will show that none of the polynomials, other than the straight

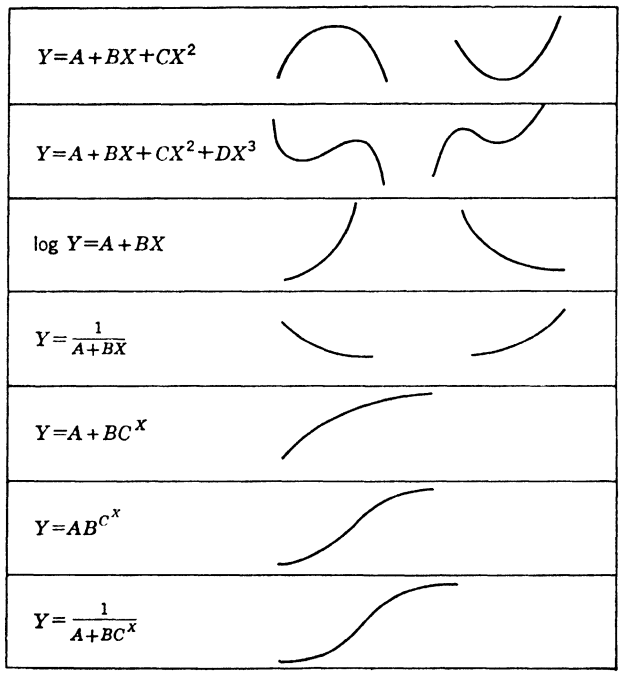

Fig. 9-4. Typical forms of some trend equations

line, can be extended or projected very far without going off the page. Keep in mind that only a portion of the curve is used to represent the trend.

One can, of course, force a polynomial to fit the data quite closely by adding enough terms. A well-known theorem in algebra states that a polynomial of degree k can be passed through any $k + 1$ points on a plane. Accomplishing this, or anything near to it, does not contribute any information about trend. This becomes evident when we recall that we lose 1 degree of freedom for error for every parameter we estimate from the data. Thus, if there are n observations and we lose n degrees of freedom in fitting a polynomial of degree $n - 1$, we have 0 degrees of freedom left for error.

All polynomials can be fitted by the method of least squares. A poly-

nomial of degree k requires $k + 1$ equations, which can be written as follows:

$$\begin{aligned}
&\text{I.} && \Sigma Y = na + b\Sigma X + c\Sigma X^2 + \cdots + j\Sigma X^k \\
&\text{II.} && \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + \cdots + j\Sigma X^{k+1} \\
&\text{III.} && \Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + \cdots + j\Sigma X^{k+2} \\
& && \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&k+1. && \Sigma X^kY = a\Sigma X^k + b\Sigma X^{k+1} + c\Sigma X^{k+2} + \cdots + j\Sigma X^{2k}
\end{aligned} \tag{9-6}$$

These equations may be solved by any of the well-known methods (see, for example, the square-root method of Chap. 10). If the origin is placed at the middle of the time scale, then all the sums of odd powers of X become 0, greatly simplifying the solution. Little more need be said about higher-degree polynomial trends. The truth of the matter is that they are not very useful.

Fitting trends after transformation of the data is a common practice. A straight line that will not fit the original data may fit the logarithms of the data or the reciprocals. This class of trends can be typified by the following two:

$$\log Y = A + BX \tag{9-7}$$

$$\frac{1}{Y} = A + BX \tag{9-8}$$

Typical traces of these curves are shown in Fig. 9-4.

The actual fitting is quite simple. To fit (9-7) one need only take the logarithms of the Y's, call them a new series of data, and solve for estimates of A and B in the usual manner, either by the method of averages or by least squares. If least squares is used, it assures us that the sum of the squares of the deviations from the logarithms is a minimum. In terms of the original data, however, the sum of squares of deviations is not necessarily a minimum. An alternative procedure is simply to plot the original data on semilogarithmic paper (i.e., paper with the vertical scale in logarithms and the horizontal scale in arithmetic units) and then to draw a straight line freehand through the data. The actual values are obtained by reading points off the graph—a process which is accurate enough for most purposes.

The term *growth curves* refers to a group of curves that are supposed to simulate the typical pattern of growth, within fixed bounds, of organisms, animals, or human beings. They are used sometimes to fit population-growth patterns (none too successfully). Three such curves are

$$Y = A + BC^x \tag{9-9}$$

$$Y = AB^{C^x} \tag{9-10}$$

$$Y = \frac{1}{A + BC^x} \tag{9-11}$$

Typical traces of these curves are shown in Fig. 9-4.

The first of the above curves is called the *modified exponential*, the second is the *Gompertz curve*, and the third (which is just the first fitted to the reciprocals of the Y's) is the *logistic curve*. We can illustrate the fitting of all these curves by fitting the first to a set of hypothetical data. They cannot be fitted by least squares except by approximation.

The widespread use of high-speed computers makes it completely feasible to fit curves such as (9-9), (9-10), and (9-11) by least squares, using approximation procedures. Such methods usually require trial solutions which are reasonably close to the least-squares solutions, in order to assure convergence of the computing procedure. A fairly rough method of estimation may yield initial values of sufficient accuracy for this purpose, however. It must be realized that there is nothing magic about least squares as a criterion for fitting curves of the nonlinear type. Perhaps its chief attributes are objectivity and general acceptability.

As soon as one departs from least squares, he encounters a large number of possible fitting procedures, the properties of which are mostly unknown. A simple procedure will be given here for fitting curves of the type (9-9), and it will be shown that (9-10) and (9-11) can be fitted by the same technique by simple transformation of the observations. The suggested procedure is easy to understand and reasonably simple to follow. Also, certain theoretical work done on fitting of exponential curves has shown that the method compares favorably with the method of least squares.* Whether or not least squares is a good criterion depends upon the distribution of the errors around the fitted line. It will be noted that errors have been ignored in the statement of models (9-9) through (9-11).

Consider Eq. (9-9). We shall assume that the origin of X is placed at the first value of Y, so that

$$Y_1 = A + BC$$
$$Y_2 = A + BC^1$$

and so forth. In general,

$$Y_X = A + BC^{X-1} \tag{9-12}$$

It is clear that

$$Y_{X+1} = A + BC^X \tag{9-13}$$

Also, if we multiply (9-12) by C, we have

$$CY_X = AC + BC^X \tag{9-14}$$

Solving (9-14) for BC^X and substituting into (9-13) yield

* I am indebted to R. F. White for pointing out the advantages of this method over certain other commonly used methods of fitting.

$$\begin{aligned} Y_{X+1} &= A - AC + CY_X \\ &= A(1 - C) + CY_X \end{aligned} \tag{9-15}$$

We can let $A(1 - C) = A'$, a new constant. Then we can see that each Y value is a linear function of the previous Y value. That is, we have a straight-line form for Y_{X+1}, as follows:

$$Y_{X+1} = A' + CY_X \tag{9-16}$$

We can solve for the constants A' and C by the familiar method of least squares, which may be theoretically inexact but which works quite well in practice. The two normal equations are

$$\begin{aligned} &\text{I.} && ma' + c\Sigma Y_X = \Sigma Y_{X+1} \\ &\text{II.} && a'\Sigma Y_X + c\Sigma Y_X^2 = \Sigma Y_X Y_{X+1} \end{aligned} \tag{9-17}$$

Note that the only difference between these normal equations and those listed in Chap. 7 is in notation. Comparing (9-17) with (7-3), we have replaced a by a', b by c, Y by Y_{X+1}, and X by Y_X. It will be noted also that, since each Y value must be multiplied by the next Y value, we must drop off 1 observation at the end of the series, so that m in (9-17) is 1 less than the number of time periods in the time series. We can solve for c directly by the formula

$$c = \frac{m\Sigma Y_X Y_{X+1} - \Sigma Y_X \Sigma Y_{X+1}}{m\Sigma Y_X^2 - (\Sigma Y_X)^2} \tag{9-18}$$

After solving Eq. (9-18) for c, we return to (9-9) and estimate A and B by the method of least squares. The normal equations are:

$$\begin{aligned} &\text{I.} && na + b\Sigma c^X = \Sigma Y_X \\ &\text{II.} && a\Sigma c^X + b\Sigma c^{2X} = \Sigma c^X Y_X \end{aligned} \tag{9-19}$$

Here, n is the number of time periods, since we can compute c^X for each of these periods. The computations for a set of hypothetical data are shown in Table 9-7 and the arithmetic which follows:

$$c = \frac{m\Sigma Y_X Y_{X+1} - \Sigma Y_X \Sigma Y_{X+1}}{m\Sigma Y_X^2 - (\Sigma Y_X)^2} = \frac{11(85{,}341) - 961(975)}{11(84{,}133) - (961)^2} = 0.9145$$

From the second set of normal equations (9-19) we compute

$$b = \frac{n\Sigma c^X Y_X - \Sigma c^X \Sigma Y_X}{n\Sigma c^{2X} - (\Sigma c^X)^2} = \frac{12(666.48) - 7.6942(1{,}055)}{12(5.3940) - (7.6942)^2} = -21.64$$

$$a = \frac{\Sigma Y_X}{n} - b\frac{\Sigma c^X}{n} = \frac{1{,}055}{12} + \frac{21.64(7.6942)}{12} = 101.8$$

Thus the computed trend equation is

$$Y = 101.8 - 21.64(0.9145)^X$$

Substituting the appropriate values of X into this equation yields the last column of Table 9-7. This is the fitted trend.

In following through the computations of Table 9-7, the student may encounter differences in final digits, depending upon the rule of rounding that is adopted.

The constant a represents an asymptote which the fitted curve approaches, but never reaches. That is, no matter how large the value of X, the fitted value of Y will not exceed 101.8.

Table 9-7. Fitting of Modified Exponential Trend to Hypothetical Data

Year	X	Y_X	Y_{X+1}	$Y_X Y_{X+1}$	Y_X^2	c^X	$c^X Y_X$	c^{2X}	$\hat{Y}$
1945	0	80	82	6,560	6,400	1.0000	80.00	1.0000	80.2
1946	1	82	84	6,888	6,724	0.9145	74.99	0.8363	82.0
1947	2	84	85	7,140	7,056	0.8363	70.25	0.6994	83.7
1948	3	85	87	7,395	7,225	0.7648	65.01	0 5849	85.2
1949	4	87	88	7,656	7,569	0.6994	60.85	0.4892	86.7
1950	5	88	89	7,832	7,744	0.6396	56 28	0.4091	88.0
1951	6	89	90	8,010	7,921	0.5849	52.06	0 3421	89.1
1952	7	90	91	8,190	8,100	0.5349	48.14	0.2861	90.2
1953	8	91	92	8,372	8,281	0.4892	44.52	0.2393	91.2
1954	9	92	93	8,556	8,464	0.4474	41.16	0.2002	92.1
1955	10	93	94	8,742	8,649	0.4091	38.05	0.1674	92.9
1956	11	94				0.3741	35.17	0.1400	93.7
Sum		961	975	85,341	84,133				
		1,055				7.6941	666.48	5.3940	

The curve denoted by formula (9-10) can be fitted in the same manner as the modified exponential by first taking the logarithms of the data. Thus,

$$\log Y = \log A + C^X \log B \tag{9-20}$$

The computations are carried out on the logarithms of the data, finding $\log a$, c, and $\log b$. The fitted curve is computed in terms of logarithms also. As a last step, the antilogarithms of the fitted points are taken to convert the curve to units of the original data.

It is obvious that (9-11) can be fitted in the same manner as (9-9) after taking reciprocals of the original data. As a last step, reciprocals of the fitted points are taken to convert the data to original units.

The curves described above are only a few of the infinitely many that could be fitted. For example, consider the curve

$$Y = A + BX^r + CX^s \tag{9-21}$$

and let r and s take on fractional values. It is immediately obvious that an unending set of such curves can be constructed.

Generally speaking, the more complex curves are avoided in applied work. One reason is that the true trend values are affected by such a myriad of economic laws, social behavior patterns, laws, and natural phenomena that any attempt at expressing the sum total of these by a simple formula is fruitless. Another reason is that, regardless of the amount of data available, one can never be sure that he has the best-fitting curve. In fact, as discussed earlier, one can always find a curve which goes through every point. This curve contributes nothing to an understanding of the long-term effect which we identify with trend. One is much better off if he draws a freehand curve which seems to average out the ups and downs in some kind of approximate fashion. This is precisely what is done in many applied situations, and one must confess that the method has considerable merit.

We can conclude this section by saying that trend fitting is at best an elusive process. It can only contribute a rough idea of long-range forces. For this reason it does not appear to merit more attention by the business analyst. He can better spend his time in studying the underlying economic, social, and technological forces which contribute to the general pattern (fitted freehand) than in trying to improve the fit of a curve which, no matter how well it fits the past, can be sure not to fit the future.

EXERCISE 9-4. (*a*) Obtain population figures for the United States, by decades, since 1790, and fit the Gompertz curve and the logistic curve to these data. (*b*) Which appears to be the better fit? (*c*) Would you care to use either curve to predict population in 1980? (*d*) in 2100? Why?

9-6. COMPUTATION OF SEASONALS

Perhaps no other phase of time-series analysis is so vital to the day-to-day planning of business operations as is seasonal analysis. Purchases are made and employees hired with an eye to seasonal patterns a few months hence. It means nothing to say that sales are up 20 per cent this month unless this figure is adjusted for normal seasonal variation.

One method of seasonal computation has been presented in Sec. 9-3, namely, a method of *absolute seasonals*. That is, seasonal variation has been computed as a positive or negative deviation (in the same units as the data) from what would be expected, or "normal," for the month if only trend and cycle were considered. It has also been shown that if the multiplicative model (9-2) holds, seasonal can be computed as above, after taking logarithms of the original data. Seasonal variation is then in units of logarithms, which is considered a disadvantage by people not trained in analytical methods. We consider now some methods by which a seasonal index can be computed without resort to logarithms.

The method of ratios to trend assumes that seasonal variation for a given

month is a constant fraction of trend. This is similar to saying that December, let's say, is 20 per cent above "normal."

The concept of the method is quite simple. Since we consider the model $Y = TCSR$, we argue that we can eliminate trend by dividing each data value by its corresponding trend value to form a "ratio to trend" composed of CSR (again ignoring trend errors). Each of these is a one-based relative, that is, a pure number with a base of unity. We eliminate cycle and random by averaging the ratios to trend, leaving only the seasonal component. Some slight adjustment is usually necessary to force the resulting ratios to average unity.

To illustrate the method of ratios to trend, let us take total retail sales in the United States. We have good reason for believing that, if the annual level of sales increases, the seasonal swing will widen accordingly.

A special problem arises whenever we work with data that are reported in dollars. The dollar itself fluctuates in value, so that part of the fluctuation in dollar volume is due to change in the price of the dollar. It is customary to correct for changing values of the dollar by dividing through by an appropriate index of prices. This process is called *deflating the data*, or *devaluating the series*. Presumably, when we consider retail sales, we are interested in a measure of the "volume" of sales in physical units, rather than the dollar volume of sales. For example, suppose we are selling vacuum cleaners. We measure how well we are doing compared with our competition, not in terms of dollar volume of sales, but in terms of number of vacuum cleaners sold. With an increasing price level, it is conceivable that we could be increasing our dollar sales and still be losing a share of the market to our competition. The adjustment for price changes has the approximate effect of making dollar volumes comparable from period to period.

An appropriate price index is not always available, and one sometimes must compromise by using an index not ideally suited to his purposes. In our case, however, we are working with retail sales, and a retail price index is available. The retail price index is based upon the average for the years 1935 to 1939, taken as 100. Thus the figure 207.8 for June, 1955, purports to show that average retail prices in June, 1955, were 107.8 per cent above the average for the period from 1935 to 1939. If we deflate by such an index, we hope to arrive at a set of figures representing *what the value of retail sales would have been if there had been no price increase since the base period.*

The raw data, the price index, the deflated series, the trend, and the ratios to trend are shown in Table 9-8. The method of fitting trend is unimportant. Since in this case trend was fitted to only 6 years' data, there has been no attempt to distinguish between trend and cycle.

The average ratios to trend for each month of the year are computed

Table 9-8. Computation of Ratios to Trend, Total Retail Sales, United States, in Millions of Dollars

Year and month		Sales	Retail price index	Devalued sales	Trend	Ratios to trend
1951	1	$12,187	202.4	$6,021	$6,109	0.986
	2	11,192	204.9	5,462	6,131	0.891
	3	12,932	205.8	6,284	6,153	1.021
	4	11,898	205.6	6,787	6,175	0.937
	5	12,736	206.5	6,168	6,197	0.995
	6	12,660	206.4	6,134	6,218	0.986
	7	12,364	206.6	6,985	6,240	0.959
	8	13,268	206.1	6,438	6,262	1.028
	9	13,101	207.4	6,317	6,284	1.005
	10	13,858	209.0	6,631	6,306	1.052
	11	13,391	210.3	6,368	6,328	1.006
	12	15,375	210.8	7,294	6,350	1.149
1952	1	11,844	210.9	5,616	6,271	0.881
	2	11,744	208.9	5,622	6,393	0.879
	3	12,736	208.7	6,103	6,415	0.951
	4	13,396	209 7	6,388	6,437	0.992
	5	14,350	210.3	6,824	6,459	1.057
	6	13,814	210.6	6,559	6,481	1.012
	7	13,396	211.8	6,325	6,503	0.973
	8	13,448	211.8	6,349	6,525	0.973
	9	13,620	211.1	6,452	6,546	0.986
	10	14,819	210 7	7,033	6,568	1.071
	11	14,008	210.4	6,658	6,590	1.010
	12	16,910	209.6	8,068	6,612	1.220
1953	1	13,054	209.0	6,246	6,634	0.942
	2	12,329	207.8	5,933	6,656	0.891
	3	13,956	208.2	6,703	6,678	1.004
	4	14,167	207.9	6,814	6,699	1.017
	5	14,665	208.2	7,044	6,721	1.048
	6	14,578	209.7	6,952	6,743	1.031
	7	14,385	210.1	6,847	6,765	1.012
	8	14,176	210.1	6,747	6,787	0.994
	9	14,082	210.3	6,696	6,809	0.983
	10	14,951	210.0	7,120	6,831	1.042
	11	13,955	208.9	6,680	6,852	0.975
	12	16,444	209 1	7,864	6,874	1.144

SOURCE: U.S. Department of Commerce, *Survey of Current Business.*

Table 9-8. Computation of Ratios to Trend, Total Retail Sales, United States, in Millions of Dollars (*Continued*)

Year and month		Sales	Retail price index	Devalued sales	Trend	Ratios to trend
1954	1	$12,339	209.5	$5,890	$6,896	0.854
	2	12,065	208.9	5,775	6,918	0.835
	3	13,540	208.3	6,500	6,939	0.937
	4	14,324	208.1	6,883	6,962	0.989
	5	14,246	208.7	6,826	6,984	0.977
	6	14,658	209.0	7,013	7,005	1.001
	7	14,390	209.7	6,862	7,027	0.977
	8	13,896	209.0	6,649	7,049	0.943
	9	14,139	208.2	6,791	7,071	0.960
	10	14,665	207.6	7,064	7,093	0.996
	11	14,531	207.6	7,000	7,115	0.984
	12	17,872	207.6	8,609	7,137	1.206
1955	1	13,279	207.3	6,406	7,158	0.895
	2	12,762	207.5	6,150	7,180	0.857
	3	14,704	207.5	7,086	7,202	0.984
	4	15,622	207.9	7,514	7,224	1.040
	5	15,468	207.7	7,447	7,246	1.028
	6	15,734	207.8	7,572	7,268	1.042
	7	15,398	208.6	7,382	7,290	1.013
	8	15,622	208.1	7,507	7,311	1.027
	9	15,905	208.9	7,614	7,333	1.038
	10	15,824	208.7	7,582	7,355	1.031
	11	15,894	208.2	7,634	7,377	1.035
	12	19,268	208.1	9,259	7,399	1.251
1956	1	13,866	207.6	6,679	7,421	0.900
	2	13,686	207.7	6,589	7,443	0.885
	3	15,864	208.2	7,620	7,464	1.021
	4	15,029	208.8	7,198	7,486	0.962
	5	16,257	209.8	7,749	7,508	1.032
	6	16,724	211.9	7,892	7,530	1.048
	7	15,382	213.6	7,201	7,552	0.954
	8	16,187	212.5	7,617	7,574	1.006
	9	15,583	213.1	7,313	7,596	0.963
	10	16,130	213.4	7,559	7,618	0.992
	11	16,493	213.8	7,714	7,639	1.010
	12	19,380	213.9	9,060	7,661	1.183

in Table 9-9. They have been carried to four decimal places. When rounded to three decimal places in the usual manner they fail, by 0.007, to add to 12.000. This is sufficient accuracy for any practical use. However, businessmen, particularly accountants, are trained to believe that everything should "balance out." We are pleased to accommodate

Table 9-9. Computation of Seasonal Index by Method of Ratios to Trend for Total Retail Sales, United States

Month	1951	1952	1953	1954	1955	1956	Average	Seasonal index
Jan.	0.986	0.881	0.942	0.854	0.895	0.900	0.9097	91.0
Feb.	0.891	0.879	0.891	0.835	0.857	0.885	0.8730	87.3
Mar.	1.021	0.951	1.004	0.937	0.984	1.021	0.9863	98.7
Apr.	0.937	0.992	1.017	0.989	1.040	0.962	0.9895	99.0
May	0.995	1.057	1.048	0.977	1.028	1.032	1.0233	102.4
June	0.986	1.012	1.031	1.001	1.042	1.048	1.0200	102.0
July	0.959	0.973	1.012	0.977	1.013	0.954	0.9813	98.2
Aug.	1.028	0.973	0.994	0.943	1.027	1.006	0.9952	99.6
Sept.	1.005	0.986	0.983	0.960	1.038	0.963	0.9892	99.0
Oct.	1.052	1.071	1.042	0.996	1.031	0.992	1.0307	103.1
Nov.	1.006	1.010	0.975	0.984	1.035	1.010	1.0033	100.4
Dec.	1.149	1.220	1.144	1.206	1.251	1.183	1.1922	119.3

them, so we round up seven figures. Those nearest the next larger number were rounded upward, and the results were multiplied by 100 to form the seasonal index in the last column. There is no reason, other than custom, for multiplying by 100.

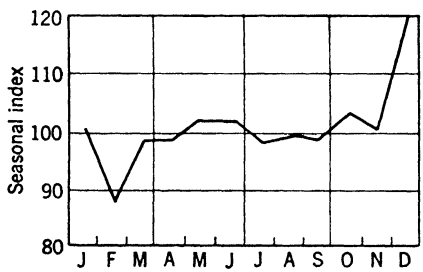

FIG. 9-5. Seasonal index of gross retail sales, United States, 1951–1956

A graph of the seasonal index is shown in Fig. 9-5. The only pronounced pattern relates to Christmas sales and the postholiday slump. Keep in mind that we are considering gross retail sales, so that seasonal movements are masked by many influences. Generally, when one studies one particular type of sales, say, gasoline sales, the seasonal pattern is more clearly defined. The seasonal pattern of gross retail sales is influ-

enced by such events as Christmas, Easter (which changes position from year to year), vacation travel, August furniture sales, January white sales, and the opening of school. The manager usually derives the greatest value from examining seasonal patterns within his own business, which directly influence his hiring, buying, and inventory policy.

EXERCISE 9-5. (*a*) Using the data of Exercise 9-2 and the trend computed in Exercise 9-3, compute a seasonal index for cement production by the method of ratios to trend. (*b*) Estimate production in 1957, by months. (*c*) Look up the actual production figures for 1957 from the *Survey of Current Business* and compare them with your estimates. (*d*) Why are you so far wrong? (*e*) What lesson is contained in this exercise?

The method of ratios to moving averages has some theoretical as well as practical advantages. In this method it is not necessary to isolate trend. The procedure is as follows:

1. Compute a 12-month moving average of the data. This obviously eliminates seasonal variation. The 12-month moving average is not "centered" at the middle of the month. For example, the average of the first 12 months falls at July 1; the average of the second 12 months falls at August 1; and so forth. In order to get a series centered at the middle of the month, we take a 2-month moving average of the 12-month moving average. This refinement is not necessary in most cases.
2. Compute ratios to moving average by dividing the data by the centered moving average for each month.
3. Average the ratios to moving average and adjust so that they total 12.000.
4. Multiply by 100 to form the seasonal index.

Note that the procedure is the same as for ratios to trend, with the exception that the moving average forms the basis for the ratio rather than trend.

The process of finding first a 12-month moving average and then a 2-month moving average of these averages seems tedious. It is. A simpler procedure is to take one average which is itself centered. Let m_7 denote the moving average to be centered opposite the seventh month. It can be computed as follows:

$$m_7 = \frac{Y_1 + 2Y_2 + 2Y_3 + 2Y_4 + 2Y_5 + 2Y_6 + 2Y_7 + 2Y_8 + 2Y_9 + 2Y_{10} + 2Y_{11} + 2Y_{12} + Y_{13}}{24}$$

That is, we do not form the noncentered moving average at all. We observe that the desired average is found from the sum of two moving totals, the first covering the months 1 to 12 and the second the months 2 to 13. The actual computing procedure is shown in Table 9-10.

$$T_8 = T_7 - Y_1 - Y_2 + Y_{13} + Y_{14}$$
$$T_9 = T_8 - Y_2 - Y_3 + Y_{14} + Y_{15}$$
$$T_{10} = T_9 - Y_3 - Y_4 + Y_{15} + Y_{16}$$

and so forth. These computations are easily run on a calculating machine or on an adding machine (by using the subtotal device).

Table 9-10. Procedure for Finding 12-month Centered Moving Average

Year and month	Data	24-month moving total	12-month centered moving average
19__ 1	Y_1		
2	Y_2		
3	Y_3		
4	Y_4		
5	Y_5		
6	Y_6		
7	Y_7	$Y_1 + 2\sum_{2}^{12} Y_i + Y_{13} = T_7$	$T_7/24$
8	Y_8	$Y_2 + 2\sum_{3}^{13} Y_i + Y_{14} = T_8$	$T_8/24$
9	Y_9	$Y_3 + 2\sum_{4}^{14} Y_i + Y_{15} = T_9$	$T_9/24$
10	Y_{10}		
11	Y_{11}		
12	Y_{12}		
19__ 1	Y_{13}		
2	Y_{14}		
3	Y_{15}		

Many other methods of computing seasonal variation are available to the business analyst. No attempt is made to catalogue them here, since the objective of this book is merely to give the reader an understanding of statistical procedures and to acquaint him with usual methods.

9-7. CYCLES

In an illustration early in this chapter, cycle was estimated by smoothing the data after removal of trend and seasonal variation. Although this is one of the more popular methods of cycle estimation, it is admittedly imperfect, primarily because a cycle estimated in this manner cannot be projected into the future. It puts the economist in the position of the

famous bird who flies backward—he knows where he has been, but doesn't know where he is going.

Sometimes a cycle is fitted to the data by some function of the sines and cosines of the time scale. Such a fitting assumes that the cyclical pattern is consistent, both in amplitude and period, or that it is changing in amplitude and/or period by some known law. Unfortunately this is almost never the case.

The blunt truth is that one can never predict an economic cycle from a historical study of the series of data itself. He must go outside the series to explain the cycle phenomenon. Sometimes the analyst can relate the cycle in the series he is studying to the cycle in another series. If the other series leads (i.e., anticipates the turning points in) the series in which he is interested, everything is fine, and his discovery assures him of a presidency in his company. Unfortunately, one often cannot find another series which is closely related to the one in which he is interested and which leads the series sufficiently to be of any value. The result is that the prediction of cycles remains in the realm of "expert judgment," or opinion. If a sufficient number of prominent economists predict an increasing business cycle, then one has some faith in the combined judgment, and he predicts accordingly. As a matter of fact, if enough people believe in an increasing business cycle and buy in anticipation of it, then the cycle will surely move upward because of the demand generated for goods. This in itself illustrates the fallacy of trying to predict a cycle from an analysis of the given series only.

EXERCISE 9-6. (*a*) Obtain monthly data on production of bricks from the *Survey of Current Business* for the years 1950 to 1957, inclusive, and compute a seasonal index by the method of ratios to moving average. (*b*) Plot the data. (*c*) Do you think the variation from a straight line is due to trend or to cycle? (*d*) Is it necessary to know for this method of computing seasonal? (*e*) Adjust the data for seasonal variation; that is, divide each production figure by the appropriate seasonal index figure.

9-8. INDEX NUMBERS

Index numbers are popular, although somewhat abused, devices for summarizing large masses of data. In effect, they relate a variable in one period to the same variable at another period, called the *base period.* To illustrate, suppose we consider the sales data of Table 9-11. Index A is found by dividing each year's sales by the sales in 1949 and multiplying by 100. Thus the index figure for 1952 means that sales for 1952 are 150 per cent of sales in 1949. Index B is found by using the average sales for 1949 and 1950 (i.e., $135,000) as a base. This base is divided into each year's sales and the result multiplied by 100 to form index B.

The illustration given in Table 9-11 poses no problems. It is only when one tries to combine series in some meaningful way to present a combined index that serious problems arise. These problems generally are related to *aggregation*, that is, the adding together of quantitative data.

We can illustrate some of these problems with a hypothetical retail price index. Our index purportedly will relate retail prices, in the aggregate, to some base period. First, we must define what we mean by "retail prices." Then we must decide where and when to collect prices and on what items. Since it is obviously impossible to collect all retail prices on all items, some sampling scheme is called for. After these problems have been solved, we must decide how to combine the various prices, that is, what "weights" to use. In the case of the price index we may decide to use quantity weights; that is, we may multiply the price of an item by

Table 9-11. Illustration of Simple Index Numbers of Sales

Year	Sales, thousands of dollars	Index A, 1949 = 100	Index B, 1949–1950 = 100
1949	120	100.0	88.9
1950	150	125.0	111.1
1951	160	133.3	118.5
1952	180	150.0	133.3
1953	200	166 7	148.1
1954	240	200 0	177.8
1955	210	175.0	155.6

the quantity of the item sold. In order to remove the influence on the index of varying quantities, we may select the quantity sold during a given period (usually the base period of the index) and use these constant weights throughout the life of the index. This procedure can lead to complications. For example, how meaningful is it to continue to price out buggy whips and kerosene lamps in this modern age? A price index based on quantities sold in the 1930s would exclude nearly all frozen foods, an extremely important part of the food budget in later years. Occasionally, the index numbers must be revised in the light of changing buying patterns.

An index number constructed by pricing out quantities of goods and services is called an *aggregative index*. That is, it is the result of dividing an aggregate (in dollars) at one date by an aggregate at another date. We can illustrate it by a cost-of-women's-clothing index for a South Sea Islander. We figure that four grass skirts and two sarongs per year are required for an island miss. The prices of these items, along with the required computations for 1939 and 1955, are shown in Table 9-12. The

aggregate cost of a year's supply of clothing is 180 shells in 1939 and 480 shells in 1955. Thus the cost of the same quantity of merchandise is 2.67 times as much in 1955 as in 1939. Note that the constant weights make it possible for us to attribute the total change in cost to price changes rather than to a combination of price and quantity.

Another common method of index-number construction is called the *average of relatives.* Here a number of separate index numbers are computed and then combined by weighted averages. For example, suppose we wish to have an index for a particular area of "business activity"—whatever that means. We decide that the indicators of business activity we shall use are bank deposits, retail sales, and freight carloadings.

Table 9-12. Computation of Hypothetical Cost-of-women's-clothing Index for a South Sea Islander (1939 = 100)

Item	Weights	Base year (1939)		1955	
		Price (shells)	Price × weight	Price (shells)	Price × weight
Grass skirt	4	20	80	60	240
Sarong	2	50	100	120	240
Total cost			180		480
Index			100		267

Obviously we cannot get any meaningful aggregate by adding these three quantities together. We can find a separate simple index number for each variable by dividing the amount for the given year by the amount for the base year. Then we must find weights for the averaging process. This is the real problem with this method, and the solution relies heavily upon knowledge of economics. In our illustration, let us assume that this problem has been solved and that we shall give retail sales a weight of 3, freight carloadings a weight of 2, and bank deposits a weight of 1. The computations for 5 years (with 1952 as the base year) are given in Table 9-13. The weighted average is found by multiplying each index number by its weight, summing, and dividing by 6 (the sum of the weights). If an increase in price causes a change in consumption, then one has difficulty in interpreting such an index number. In fact, many of the real problems of index numbers center around interpretation rather than construction.

There are many special-purpose index numbers, formed by a combination of special weights and special averaging procedures. Generally speaking, when index numbers become too complex, it is difficult, if not impossible, to interpret them.

The person who must construct index numbers should have access to a good reference on the subject. Usually, however, although the business manager should understand the significance of index numbers and some of their limitations, he need not have detailed knowledge of their construction. It is desirable for the business manager to become acquainted with the common index numbers in his field. In particular, he should acquaint himself with the index numbers in *Survey of Current Business*, a monthly publication of the U.S. Department of Commerce.

Table 9-13. Hypothetical Index of Business Activity (1952 = 100)

Year	Retail sales (weight = 3)	Bank deposits (weight = 1)	Freight carloading (weight = 2)	Weighted average
1952	100	100	100	100.0
1953	95	98	102	97.8
1954	102	100	104	102 3
1955	106	102	104	104.7
1956	110	105	108	108 5

EXERCISE 9-7. From the *Statistical Abstract of the United States* obtain average prices of selected foods (say, wheat flour, round steak, pork chops, eggs, fresh milk, lard, cheese, oranges, potatoes, coffee, and sugar) for the years 1949 to 1955. From the same publication find suitable weights for the year 1953 for these items of food. (*a*) Construct a food price index for these years, with 1953 as base year. (*b*) Compare your index with the National Industrial Conference Board's index of food prices for the same years. What is wrong, if anything, with your index? Is the NICB index an exact representation of price changes? (*c*) What would happen if 90 per cent of wage agreements in the United States were tied to a cost-of-living index which was biased upward, by 5 per cent, for all years starting with the current year?

10. MULTIPLE REGRESSION

10-1. THE MULTIPLE-REGRESSION EQUATION

In multiple regression, it is assumed that we have a "dependent" variable Y which is dependent upon p "independent" variables X_1, X_2, . . . , X_p. Another way of saying the same thing is to assert that each Y value in the population can be expressed in the following form:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i \qquad (10\text{-}1)$$

where α, β_1, β_2, . . . , β_p are parameters and the X_{1i}, X_{2i}, . . . , X_{pi} are known. The ϵ_i are observations on a random variable, independent of the X's. For example, we may postulate that the life of a particular type of small electric motor is dependent upon (1) the load under which it operates, (2) the temperature of the room, (3) humidity, and (4) random factors. That is, in Eq. (10-1), we might let X_1 = load, X_2 = temperature, X_3 = humidity, and ϵ_i = factors not otherwise accounted for. In most cases this last element is not truly random, because it includes some factors that could be accounted for if one had the time and patience to investigate them. However, in many cases it is not inappropriate to consider it to be a random variable following some postulated probability distribution.

It is important to note that the model expressed as Eq. (10-1) is a linear model. That is, Y is a sum of products of parameters and known functions of the X's. Our model would still be linear in the parameters if we included such terms as $\nu X_1/X_2$ and $\delta X_3/X_4$, but we could not include terms such as X_2^{ν}. It must be obvious that linear-regression methods will be of use only in those cases in which the linear model is a reasonably close approximation to reality, for there can be little meaning in "assuming" a statistical model if that model has no correspondence to reality. This point is discussed further in a later chapter.

Our first problem in regression analysis is to find estimators for α, β_1, β_2, . . . , β_p, so that we can write the estimated regression equation

$$\bar{Y}_X = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p \tag{10-2}$$

After finding this equation, we can insert various sets of values into the X's to estimate Y.

Reference to Sec. 7-3 shows that the model (10-1) is a general linear model, so under appropriate assumptions concerning ϵ_i the method of least squares will provide best linear unbiased estimates of the parameters α, β_1, . . . , β_p. The expression we wish to minimize is

$$Q = \Sigma_i(Y_i - \alpha - \beta_1X_{1i} - \cdots - \beta_pX_{pi})^2 \tag{10-3}$$

The student with a little training in calculus can easily verify that, by differentiating Q with respect to α, β_1, β_2, . . . , β_p, successively, and setting the resulting derivatives equal to 0, one obtains the following normal equations:

$$\begin{aligned}
&\text{I.} && na + b_1\Sigma X_1 + b_2\Sigma X_2 + \cdots + b_p\Sigma X_p = \Sigma Y \\
&\text{II.} && a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 + \cdots + b_p\Sigma X_1X_p = \Sigma X_1Y \\
&\text{III.} && a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 + \cdots + b_p\Sigma X_2X_p = \Sigma X_2Y \\
&&& \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&P+1. && a\Sigma X_p + b_1\Sigma X_1X_p + b_2\Sigma X_2X_p + \cdots + b_p\Sigma X_p^2 = \Sigma X_pY
\end{aligned} \tag{10-4}$$

Note that, since these equations are used to estimate the parameters, a has been substituted for α and b_j has been substituted for β_j.

The nonmathematical student can easily remember the form of these equations as follows:

1. The first normal equation is found by summing the equation

$$a + b_1X_1 + b_2X_2 + \cdots + b_pX_p = Y \tag{10-5}$$

2. The second normal equation is found by multiplying both sides of (10-5) by X_1 and summing.
3. The third normal equation is found by multiplying both sides of (10-5) by X_2 and summing.

The process is repeated for each X variable. When the equations are all written, there is one for each of the parameters to be estimated. The first equation (I, above) is called the α normal equation, since it is found by minimizing (10-3) with respect to α. Similarly, the second normal equation is the β_1 equation, the third the β_2 equation, and so forth.

The solution of the set of normal equations can become tedious if the number of independent variables (and consequently the number of equa-

tions) is large. Economical methods of solution are discussed in a later section.

10-2. DIGRESSION ON MATRICES

Although this is not a book in mathematics, some understanding of matrices is necessary in order to simplify the computational methods. Furthermore, increasing numbers of problems of managerial interest are being expressed in matrix terms. Among them are problems in linear programming, the theory of games, and production models. It seems not out of order, then, to devote a minimum of space to some elementary matrix concepts.

A matrix is a rectangular array of numbers (or symbols) as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \cdots & \cdots & \cdots & \cdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix} = A \tag{10-6}$$

Since there are r rows and c columns, this is called an $r \times c$ matrix with *elements* a_{ij}.

Two matrices are said to be equal if and only if their corresponding elements are equal.

Addition of two matrices (of the same dimension) is accomplished as follows:

$$\begin{bmatrix} a_{11}a_{12}a_{13} \\ a_{21}a_{22}a_{23} \end{bmatrix} + \begin{bmatrix} b_{11}b_{12}b_{13} \\ b_{21}b_{22}b_{23} \end{bmatrix} = \begin{bmatrix} (a_{11}+b_{11})(a_{12}+b_{12})(a_{13}+b_{13}) \\ (a_{21}+b_{21})(a_{22}+b_{22})(a_{23}+b_{23}) \end{bmatrix} \tag{10-7}$$

Addition can be denoted by $A + B = C$ where A, B, and C are matrices of the same dimension.

$$\begin{bmatrix} 5 & 6 & 7 \\ 8 & -1 & 2 \end{bmatrix} + \begin{bmatrix} 4 & -2 & 6 \\ 3 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 9 & 4 & 13 \\ 11 & 0 & 4 \end{bmatrix}$$

Subtraction is similarly defined, with the minus sign replacing the plus sign.

EXERCISE 10-1. Complete the following equation:

$$\begin{bmatrix} 5 & 13 \\ -11 & 21 \\ 4 & -40 \end{bmatrix} - \begin{bmatrix} 3 & 6 \\ 20 & 11 \\ -15 & -30 \end{bmatrix} =$$

In multiplying two matrices, A and B, the number of columns in the first, A, must be equal to the number of rows in the second, B. The first element in the product is the sum of the products of the elements

in the first row of A times the elements in the first column of B. The second element in the first row of the product is the sum of the products of the elements in the first row of A times the elements in the second column of B. That is,

$$\begin{matrix} A & & B & & = & & C \end{matrix}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21})(a_{11}b_{12} + a_{12}b_{22}) \\ (a_{21}b_{11} + a_{22}b_{21})(a_{21}b_{12} + a_{22}b_{22}) \\ (a_{31}b_{11} + a_{32}b_{21})(a_{31}b_{12} + a_{32}b_{22}) \end{bmatrix} \tag{10-8}$$

In numbers,

$$\begin{bmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 9 \end{bmatrix} \times \begin{bmatrix} -1 & 2 \\ 3 & -4 \end{bmatrix} = \begin{bmatrix} 11 & -12 \\ 15 & -16 \\ 19 & -20 \end{bmatrix}$$

Notice that $AB \neq BA$ except in special cases. In the above illustration, for example, BA is not even defined.

Exercise 10-2. Do the following multiplication operations:

(*a*) $$\begin{bmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \end{bmatrix} \times \begin{bmatrix} -1 & 3 & -4 \\ 2 & 9 & 0 \\ 4 & 6 & 2 \end{bmatrix} =$$

(*b*) $$[1 \quad 0 \quad 1] \times \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} =$$

(*c*) $$\begin{bmatrix} 2 & 2 \\ 3 & 1 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} =$$

Exercise 10-3. We wish to construct an index number of consumer stock prices for 4 corporations. The following data are gathered:

Year	Stock prices in dollars			
	Corp. A	Corp. B	Corp. C	Corp. D
1956	14	10	14	16
1957	10	8	15	18
1958	12	10	15	24
1959	14	9	17	30

The following weights are to be assigned, based on outstanding shares: A, 50; B, 100; C, 20; D, 30. (*a*) Multiply the 4×4 matrix of prices by the 4×1 matrix of weights to obtain the "price aggregates." (*b*) Divide each price aggregate by the 1956 aggregate to obtain the index number.

Multiplication of matrices has been defined so that the expression $AB = C$ has meaning. One may wonder, however, whether the above expression can be solved for B. That is, can one write $B = D$, where D is some function of A and C? The answer is yes, in some circumstances. Suppose, for example, A, B, and C are identified as follows:

$$\begin{matrix} A & & B & = & C \end{matrix}$$
$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{10-9}$$

Upon expansion by the laws of multiplication and by the rule for equality of two matrices, one obtains the identical expression

$$\begin{aligned} X_1 + 2X_2 &= 1 \\ 3X_1 + 4X_2 &= 2 \end{aligned}$$

We wish to solve for X_1 and X_2 (i.e., for the elements of matrix B). This solution requires that we find the "inverse" of A, which is written A^{-1}, where A^{-1} is a matrix such that $A^{-1}A = I$, and I is the matrix

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I \tag{10-10}$$

I is called the *identity* matrix, because multiplying any matrix by I (where such multiplication is defined) does not change the matrix so multiplied.

Exercise 10-4. Compute

$$\begin{bmatrix} 3 & 2 & 1 \\ 1 & 4 & 6 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} =$$

It may be seen in the above example that the inverse of A may be found by solving for the c_{ij} in the following:

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

By the rules of multiplication,

$$\begin{aligned} 2c_{21} + 4c_{22} &= 1 & \qquad 2c_{11} + 4c_{12} &= 0 \\ c_{21} + 3c_{22} &= 0 & \qquad c_{11} + 3c_{12} &= 1 \end{aligned}$$

These equations are easily solved to give:

$$\begin{aligned} c_{11} &= -2 & \qquad c_{12} &= 1 \\ c_{21} &= \tfrac{3}{2} & \qquad c_{22} &= -\tfrac{1}{2} \end{aligned}$$

Therefore,

$$A^{-1} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$$

To verify that this is, in fact, the inverse of A, we see that

$$A^{-1}A = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

EXERCISE 10-5. Compute AA^{-1} in the above illustration. This is a numerical verification of a general rule.

If we multiply $AB = C$ on the left by A^{-1}, we get $A^{-1}AB = A^{-1}C$, or $B = A^{-1}C$, since $A^{-1}AB = IB = B$. Hence we obtain a solution for B. In the above example,

$$\begin{matrix} B & = & A^{-1} & & C & = & D \end{matrix}$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$$

Since two matrices are equal if and only if their corresponding elements are equal, we have

$$x_1 = 0$$
$$x_2 = \tfrac{1}{2}$$

The student can easily verify that these values satisfy the equations

$$x_1 + 2x_2 = 1$$
$$3x_1 + 4x_2 = 2$$

which are the equations represented by $AB = C$.

EXERCISE 10-6. Do so.

This brief introduction to matrix algebra is the very minimum required for an understanding of the computational techniques which follow.

10-3. SIMPLIFICATION OF NORMAL EQUATIONS

Suppose that all observations have been expressed as deviations from their means. That is,

$$y_i = Y_i - \bar{Y}$$
$$x_{1i} = X_{1i} - \bar{X}_1$$
$$x_{2i} = X_{2i} - \bar{X}_2$$

and so forth. Then it is clear that, in the normal equations (10-4), ΣY, ΣX_1, ΣX_2, . . . will vanish, so that there will be one less normal

equation. With this simplification, the normal equations can be written

$$\begin{aligned}
&\text{I.} && b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 + \cdots + b_p\Sigma x_1x_p = \Sigma x_1y \\
&\text{II.} && b_1\Sigma x_1x_2 + b_2\Sigma x_2^2 + \cdots + b_p\Sigma x_2x_p = \Sigma x_2y \\
& && \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&P. && b_1\Sigma x_1x_p + b_2\Sigma x_2x_p + \cdots + b_p\Sigma x_p^2 = \Sigma x_py
\end{aligned} \tag{10-11}$$

If each of these equations is multiplied by n, the values of the b's will not be affected, and the equations can be written

$$\begin{aligned}
&\text{I.} && b_1G_{11} + b_2G_{12} + \cdots + b_pG_{1p} = G_{1y} \\
&\text{II.} && b_1G_{12} + b_2G_{22} + \cdots + b_pG_{2p} = G_{2y} \\
& && \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&P. && b_1G_{1p} + b_2G_{2p} + \cdots + b_pG_{pp} = G_{py}
\end{aligned} \tag{10-12}$$

The G's have the same meaning as given in Chap. 7. That is,

$$\begin{aligned}
G_{11} &= n\Sigma X_1^2 - (\Sigma X_1)^2 \\
G_{14} &= n\Sigma X_1X_4 - \Sigma X_1\Sigma X_4 \\
G_{3y} &= n\Sigma X_3Y - \Sigma X_3\Sigma Y
\end{aligned} \tag{10-13}$$

and so forth.

As in the case of simple regression, the b's computed from the normal equations (10-12) are the same as those computed from (10-4). However, the value a must be computed in order to arrive at the estimating equation (10-2). The student interested in algebra can show that

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - \cdots - b_p\bar{X}_p \tag{10-14}$$

Exercise 10-7. In the equation

$$y = b_1x_1 + b_2x_2 + \cdots + b_px_p$$

substitute $Y - \bar{Y}$ for y, $X_1 - \bar{X}_1$ for x_1, and so on, and collect all constant terms. This is the constant a.

An orderly procedure for finding the G's is provided by the computing form shown in Table 10-1. The figures in the top section of the table constitute the raw data or observations. The last column is the sum of the columns to the left. Its purpose is for checking the computations below. It is important that all computations be checked in a systematic manner even though some additional time is spent in the process. Since each step uses the previously computed figures, much time is wasted if errors are permitted to enter the computations.

The figures in the middle section of Table 10-1 are entered directly from the calculating machine. The last column is handled like any other column. A check on the first row is provided as follows:

$$\Sigma X_1^2 + \Sigma X_1X_2 + \Sigma X_1X_3 + \cdots + \Sigma X_1X_p + \Sigma X_1Y = \Sigma X_1S \tag{10-15}$$

The check is indicated by arrows on the table. If these figures do not check exactly, the student should find the error before proceeding. Checks on other rows are provided by starting at the top, going down to the diagonal term, then moving across to the right-hand side. As an example, consider the double arrows on the table:

$$\Sigma X_1X_3 + \Sigma X_2X_3 + \Sigma X_3^2 + \cdots + \Sigma X_3X_p + \Sigma X_3Y = \Sigma X_3S \quad (10\text{-}16)$$

The element G_{ij} in the lower section of the table is computed as $n\Sigma X_iX_j - \Sigma X_i\Sigma X_j$, where, obviously, $\Sigma X_iX_j = \Sigma X_i^2$ if $i = j$. Thus it may be seen that a G_{ij} is computed by multiplying the corresponding

Table 10-1. Computing Form for G Matrix

X_1	X_2	X_3	$\cdots$	X_p	Y	S
X_{11}	X_{21}	X_{31}		X_{p1}	Y_1	S_1
X_{12}	X_{22}	X_{32}	$\cdots$	X_{p2}	Y_2	S_2
$\cdots$	$\cdots$	$\cdots$		$\cdots$	$\cdots$	$\cdots$
X_{1n}	X_{2n}	X_{3n}		X_{pn}	Y_n	S_n
ΣX_1	ΣX_2	ΣX_3	$\cdots$	ΣX_p	ΣY	ΣS
$\Sigma X_1^2 \longrightarrow$	$\Sigma X_1X_2 \longrightarrow$	$\Sigma X_1X_3 \longrightarrow$	$\cdots \longrightarrow$	$\Sigma X_1X_p \longrightarrow$	$\Sigma X_1Y \rightarrow$	ΣX_1S
	ΣX_2^2	ΣX_2X_3	$\cdots$	ΣX_2X_p	ΣX_2Y	ΣX_2S
		$\Sigma X_3^2 \rightarrow\rightarrow$	$\cdots \rightarrow\rightarrow$	$\Sigma X_3X_p \rightarrow\rightarrow$	$\Sigma X_3Y \rightarrow\rightarrow$	ΣX_3S
				$\cdots\cdots\cdots\cdots$	ΣY^2	ΣYS
G_{11}	G_{12}	G_{13}	$\cdots$	G_{1p}	G_{1y}	G_{1s}
	G_{22}	G_{23}	$\cdots$	G_{2p}	G_{2y}	G_{2s}
		G_{33}	$\cdots$	G_{3p}	G_{3y}	G_{3s}
				$\cdots\cdots\cdots\cdots$	$\cdots\cdots\cdots$	$\cdots\cdots$
					G_{yy}	G_{ys}

entry in the middle section of the table by n and subtracting the product of ΣX_i and ΣX_j. The entire operation can be accomplished on the calculator without recording anything but the final result. The last column is handled the same as any other, and the checks are accomplished as above.

10-4. A SIMPLE COMPUTING FORM

One is likely to get the impression that the solution of a multiple-regression problem is a tedious proposition. This need not be the case. Many simplified solutions exist, and one of the most straightforward is given in this section.

First of all, observe that the normal equations (10-12) can be written in the following matrix form:

$$\begin{matrix} G & & B & = & Y_X \end{matrix}$$

$$\begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1p} \\ G_{21} & G_{22} & \cdots & G_{2p} \\ & & \cdots & \\ G_{p1} & G_{p2} & \cdots & G_{pp} \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_p \end{bmatrix} = \begin{bmatrix} G_{1y} \\ G_{2y} \\ \cdots \\ G_{py} \end{bmatrix} \tag{10-17}$$

The G matrix is symmetric. That is, $G_{ij} = G_{ji}$. This fact simplifies the computations somewhat. A simple solution for the b's is found by using the computing form of Table 10-2.

The method is commonly called the *square-root method.** As in all other computing methods, a regular pattern is employed to arrive at each stage of the computations. In this case the pattern can be learned

Table 10-2. Computing Form for Solving Normal Equations (the Square-root Method)

Line	X_1	X_2	X_3	X_4	Y	S	Check
1	G_{11}	G_{12}	G_{13}	G_{14}	G_{1y}	G_{1s}	G'_{1y}
2		G_{22}	G_{23}	G_{24}	G_{2y}	G_{2s}	G'_{2y}
3			G_{33}	G_{34}	G_{3y}	G_{3s}	G'_{3y}
4				G_{44}	G_{4y}	G_{4s}	G'_{4y}
5	B_{11}	B_{12}	B_{13}	B_{14}	B_{1y}	B_{1s}	
6		B_{22}	B_{23}	B_{24}	B_{2y}	B_{2s}	
7			B_{33}	B_{34}	B_{3y}	B_{3s}	
8				B_{44}	B_{4y}	B_{4s}	
9	b_1	b_2	b_3	b_4			

easily and applied to any size of matrix. Instructions for computing are as follows:

LINES 1 TO 4. Copy the G_{ij} from Table 10-1. The S column is for checking purposes. Each G_{is} is the sum of the elements in the corresponding row of the G matrix. That is,

$$G_{3s} = G_{31} + G_{32} + G_{33} + G_{34} + G_{3y}$$

Since the G matrix is symmetric, only the diagonal elements and those elements above them are shown. Since $G_{31} = G_{13}$, and $G_{32} = G_{23}$, we can write

$$G_{3s} = G_{13} + G_{23} + G_{33} + G_{34} + G_{3y} \tag{10-18}$$

and so forth.

* For a discussion of the method see Paul S. Dwyer, *Linear Computations*, pp. 113–117, John Wiley & Sons, Inc., New York, 1951.

LINE 5:

$$B_{11} = \sqrt{G_{11}}$$
$$B_{12} = \frac{G_{12}}{B_{11}}$$
$$B_{13} = \frac{G_{13}}{B_{11}}$$
$$B_{14} = \frac{G_{14}}{B_{11}}$$
$$B_{1y} = \frac{G_{1y}}{B_{11}}$$
$$B_{1s} = \frac{G_{1s}}{B_{11}}$$

In general,

$$B_{1j} = \frac{G_{1j}}{B_{11}} \tag{10-19}$$

A check is provided by adding B_{11} through B_{1y} to obtain B_{1s}.

LINE 6:

$$B_{22} = \sqrt{G_{22} - B_{12}^2}$$
$$B_{23} = \frac{G_{23} - B_{13}B_{12}}{B_{22}}$$
$$B_{24} = \frac{G_{24} - B_{14}B_{12}}{B_{22}}$$
$$B_{2y} = \frac{G_{2y} - B_{1y}B_{12}}{B_{22}}$$
$$B_{2s} = \frac{G_{2s} - B_{1s}B_{12}}{B_{22}}$$

In general,

$$B_{2j} = \frac{G_{2j} - B_{1j}B_{12}}{B_{22}} \tag{10-20}$$

We check again by verifying that the last element, B_{2s}, is the sum of the other elements on the same line.

LINE 7:

$$B_{33} = \sqrt{G_{33} - B_{13}^2 - B_{23}^2}$$
$$B_{34} = \frac{G_{34} - B_{14}B_{13} - B_{24}B_{23}}{B_{33}}$$
$$B_{3y} = \frac{G_{3y} - B_{1y}B_{13} - B_{2y}B_{23}}{B_{33}}$$
$$B_{3s} = \frac{G_{3s} - B_{1s}B_{13} - B_{2s}B_{23}}{B_{33}}$$

In general,

$$B_{3j} = \frac{G_{3j} - B_{1j}B_{13} - B_{2j}B_{23}}{B_{33}} \tag{10-21}$$

LINE 8:

$$B_{44} = \sqrt{G_{44} - B_{14}^2 - B_{24}^2 - B_{34}^2}$$

$$B_{4y} = \frac{G_{4y} - B_{1y}B_{14} - B_{2y}B_{24} - B_{3y}B_{34}}{B_{44}}$$

$$B_{4s} = \frac{G_{4s} - B_{1s}B_{14} - B_{2s}B_{24} - B_{3s}B_{34}}{B_{44}}$$

In general,

$$B_{4j} = \frac{G_{4j} - B_{1j}B_{14} - B_{2j}B_{24} - B_{3j}B_{34}}{B_{44}} \tag{10-22}$$

LINE 9:

$$b_4 = \frac{B_{4y}}{B_{44}}$$

$$b_3 = \frac{B_{3y} - B_{34}b_4}{B_{33}}$$

$$b_2 = \frac{B_{2y} - B_{24}b_4 - B_{23}b_3}{B_{22}}$$

$$b_1 = \frac{B_{1y} - B_{14}b_4 - B_{13}b_3 - B_{12}b_2}{B_{11}}$$

In general,

$$b_i = \frac{B_{iy} - B_{ip}b_4 - \cdots - B_{i,i+1}b_{i+1}}{B_{ii}} \tag{10-23}$$

A final check on the accuracy of the computations is provided by finding the G'_{ij} (check column):

$$\begin{aligned} G'_{1y} &= b_1G_{11} + b_2G_{12} + b_3G_{13} + b_4G_{14} \\ G'_{2y} &= b_1G_{12} + b_2G_{22} + b_3G_{23} + b_4G_{24} \\ G'_{3y} &= b_1G_{13} + b_2G_{23} + b_3G_{33} + b_4G_{34} \\ G'_{4y} &= b_1G_{14} + b_2G_{24} + b_3G_{34} + b_4G_{44} \end{aligned} \tag{10-24}$$

These figures should agree with the G_{iy} figures in column Y, except for a reasonable rounding error. An illustration is provided by the data shown in Tables 10-3 and 10-4.

A word is in order about the amount of accuracy required. In the computation of regression coefficients, the number of significant digits (see Sec. 4-1) is the important consideration. If one wishes to have the coefficients correct to five significant digits, all the computations in the table should be carried to at least six significant digits. This was done in Table 10-4. In order to make the check column agree with the Y column, the number of significant digits carried throughout the table must be one

greater than the maximum number of significant digits in the G table. Carrying the same number of significant digits as in the maximum G will ordinarily ensure that only the last place will differ (as was the case in Table 10-4).

From Table 10-4 we obtain the regression coefficients b_1, b_2, b_3, and b_4. We can compute the constant a from Eq. (10-14):

$$\begin{aligned} a &= \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - \cdots - b_p\bar{X}_p \\ &= \frac{1}{n}(\Sigma Y - b_1\Sigma X_1 - b_2\Sigma X_2 - \cdots - b_p\Sigma X_p) \\ &= \tfrac{1}{20}[2{,}854 - 0.956377(1{,}142) - 2.003311(822) \\ &\qquad + 0.965165(1{,}138) + 0.948983(1{,}042)] = 110.11 \end{aligned}$$

The final regression equation is then

$$\bar{Y}_X = 110.11 + 0.956377X_1 + 2.003311X_2 - 0.965165X_3 - 0.948983X_4 \tag{10-25}$$

The constants are carried out to more decimal places than are needed. As pointed out earlier, they are carried to this accuracy to facilitate checking.

Suppose now that we take (10-25) as a good estimate of the real relationship of Y to the four X variables. Then, if we know the four X values, we can estimate Y. Let us assume that Y is an index of performance on the job and that X_1, X_2, X_3, and X_4 are scores on aptitude tests. We may also assume that a predicted employment index of less than 100 is very poor and that an index of over 200 is outstanding. Mr. Murphy is seeking a job with us and achieves the following scores on the four placement tests: $X_1 = 40$, $X_2 = 25$, $X_3 = 80$, $X_4 = 60$. We wish to predict how successful Mr. Murphy will be. We insert the above scores into the estimating equation (10-25) and compute

$$\bar{Y}_X = 110 + 0.956(40) + 2.003(25) - 0.965(80) - 0.949(60) = 64$$

We conclude that Mr. Murphy is no ball of fire, and we probably should not hire him.

The size of the figures in the G matrix ordinarily can be reduced by careful coding of the original data. In the usual case, two or three significant digits in the original data are adequate to achieve the accuracy desired in the regression coefficients. In Chap. 7 it was observed that in simple linear regression (two variables) the subtraction of a constant from each of the variables has no effect on the value of b (the simple regression coefficient). The same principle holds for multiple regression. Coding by subtraction does not affect the b coefficients, but it does affect the value of a. Ordinarily, however, there is little interest in a. If one codes by subtraction, he gets a regression equation in coded units. He

Table 10-3. Computation of G Matrix (Hypothetical Data)

X_1	X_2	X_3	X_4	Y	S
47	33	51	27	130	288
90	32	94	1	168	385
50	45	40	64	158	357
13	14	54	58	37	176
58	43	24	61	189	375
30	86	80	74	160	430
4	66	43	67	144	324
57	46	58	33	176	370
95	19	27	38	170	349
94	31	38	32	184	379
36	60	68	90	107	361
98	0	82	83	57	320
12	48	36	83	111	290
82	4	87	99	11	283
94	58	16	82	217	467
22	42	72	63	80	279
78	68	89	13	223	471
35	86	3	17	299	440
86	12	98	8	124	328
61	29	78	49	109	326
1,142	822	1,138	1,042	2,854	6,998

Sums of squares and cross products

83,882	40,097	68,337	54,706	168,429	415,451
	45,426	42,212	42,301	138,980	309,016
		79,586	57,773	143,591	391,499
			70,728	128,921	354,429
				494,262	1,074,183

G matrix

373,476	−136,784	67,144	−95,844	109,312	317,304
	232,836	−91,196	−10,504	433,612	427,964
		296,676	−30,336	−376,032	−133,744
			328,796	−395,448	−203,336
				1,739,924	1,511,368

an convert this equation back to units of the original data by replacing ach coded variable by the variable minus the subtracted constant and olving. That is,

$$\hat{Y}_X - C_Y = a + b_1(X_1 - C_1) + b_2(X_1 - C_2) + \cdots + b_p(X_p - C_p) \quad (10\text{-}26)$$

here the C's are the subtracted constants. For example, suppose that 1 the above illustration 10 had been subtracted from X_1, 50 from X_2,

Table 10-4. Computation of Regression Coefficients (Hypothetical Data)

X_1	X_2	X_3	X_4	Y	S	Check
373,467	−136,784	67,144	−95,844	109,312	317,304	109,312
	232,836	−91,196	−10,504	433,612	427,964	433,613
		296,676	−30,336	−376,032	−133,744	−376,032
			328,796	−395,448	−203,336	−395,448
611.127	−223.823	109 869	−156.832	178.870	519.211	
	427 480	−155 808	−106.687	1,108 000	1,272.984	
		510.224	−58 2640	−437.159	14 8011	
			537 980	−510.534	27.447	
.956377	2.003311	−0.965165	−0.948983			

00 from X_3, 20 from X_4, and 25 from Y. If we insert these figures into 10-26), we get

$$\hat{Y}_X - 25 = 110.11 + 0.956377(X_1 - 10) + 2.003311(X_2 - 50) - 0.965165(X_3 - 100) - 0.948983(X_4 - 20)$$

Expanding and collecting terms, the equation becomes

$$\hat{Y}_X = 140.88 + 0.956377X_1 + 2.003311X_2 - 0.965165X_3 - 0.948983X_4$$

This is the regression equation in units of the original data. It is called he *decoded* regression equation.

Exercise 10-8. The strength of a molding is supposed to be a function of ressure used in the forming process, amount of chemical A used in the mix, and emperature at time of forming. A special treatment B is under consideration. Moldings will receive the treatment or they won't. There are no other levels f the treatment. Thirty moldings are made, half of which receive treatment B. The variables are designated as follows in the table on page 212:

X_1 = pressure
X_2 = amount of chemical A
X_3 = treatment B (= 0 if no treatment, = 1 if treated)
Y = strength of molding

All data are coded for easier computation.

X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
12	20	0	28	15	25	0	28
15	24	1	31	7	17	1	28
13	19	0	25	10	19	0	26
10	23	1	25	12	23	1	29
8	22	0	26	14	21	0	30
17	20	1	33	12	22	1	27
13	22	0	28	16	24	0	30
11	18	1	24	9	18	1	27
12	21	0	28	11	20	1	25
11	24	0	23	14	23	1	30
8	22	0	25	9	21	1	21
10	22	0	24	10	24	1	29
12	18	0	24	11	17	1	28
14	21	0	31	13	19	1	30
15	25	0	30	9	20	1	25

Find the multiple-regression equation for predicting Y. This exercise illustrates the use of a variable, X_3, which can take on only one of two possible values.

10-5. MEANING OF THE REGRESSION COEFFICIENTS

The regression coefficients $b_1, b_2, \ldots, b_p$ are sometimes called *partial regression coefficients.* They show how much influence each X variable has on Y *in units of the original data.* That is, in the above illustration, $b_1 = 0.956377$, so that, if all the other X variables are held constant, a 1-unit increase in X_1 will increase Y, on the average, by 0.95 unit. (There is spurious accuracy in carrying the b's to six decimal places. It has been done in this illustration to demonstrate the use of the check column and for consistency.) Similarly, with X_1, X_2, and X_4 held constant, a 1-unit increase in X_3 will, on the average, decrease Y by 0.97 unit.

Note that a linear, or straight-line, relationship has been assumed between Y and each X. The assumption of straight-line relationship may be reasonably sound in general over short ranges of X_1, but the true relationship may be far from linear over wider ranges. For example, if Y is production of machined parts and X_1 is speed of lathe, a certain degree of increase in X_1 (speed) is likely to increase Y (production), but a further increase in speed may actually be expected to decrease production (if we count only good units). Sometimes the unit of measurement of the X's can be so chosen as to make the relation linear. For example,

suppose Y is weight of pebbles which are approximately round and X is maximum diameter. We should expect the relationship between diameter and weight to be far from linear, because the volume of a sphere is $\frac{4}{3}\pi r^3$, where r is the radius. However, the relationship between weight and the cube of the radius should be approximately linear.

When there is reason to believe that the relationship between Y and an X variable, say, X_3, is curvilinear and there is no suitable way of transforming X_3 so that the relationship is linear, it may be desirable to compute a curvilinear multiple regression. Suppose the relationship between Y and X_3 is parabolic, that is, of the form $Y = A + BX + CX^2$ (see Fig. 9-4). Then one can square the X_3 variable, call it X_4 (say), and proceed in the usual manner to estimate the regression coefficients. The regression equation sought is

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_3^2$$

but, for convenience in notation, it is written

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

where $X_4 = X_3^2$. No new problems arise by this approach.

Similarly, if the regression of Y on any X can be expressed as any simple function of X, such as $1/X$, $\sqrt{X}$, arcsin X, or log X, this function can be used in place of the original X in the regression computation. We see then that the methods of solution presented here can be made to fit, at least approximately, a very wide range of practical problems.

One cannot compare regression coefficients directly unless the X's are in the same units. Ordinarily they will not be. One variable may be in dollars, another may be in pounds, and a third in feet. The b's can be converted to *standardized regression coefficients* as follows:

$$\begin{aligned} B_1 &= b_1\sqrt{\frac{G_{11}}{G_{yy}}} \\ B_2 &= b_2\sqrt{\frac{G_{22}}{G_{yy}}} \end{aligned} \tag{10-27}$$

and so forth.

These standardized regression coefficients can also be found by solving the following equations:

$$\begin{aligned} &B_1 + r_{12}B_2 + r_{13}B_3 + \cdots + r_{1p}B_p = r_{1y} \\ &r_{12}B_1 + B_2 + r_{23}B_3 + \cdots + r_{2p}B_p = r_{2y} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &r_{1p}B_1 + r_{2p}B_2 + r_{3p}B_3 + \cdots + B_p = r_{py} \end{aligned} \tag{10-28}$$

where the r's are the simple correlation coefficients between the variables indicated by the subscripts.

EXERCISE 10-9. Solve for the standardized regression coefficients of the data presented in Exercise 10-8.

10-6. TEST OF THE HYPOTHESIS THAT THERE IS NO REGRESSION

The hypothesis that there is no regression may be stated as follows:

$$H_0: \qquad \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0 \tag{10-29}$$

If this hypothesis is true, we cannot predict Y any more accurately by the use of regression than we can by simply taking the average Y value. Following the pattern of analysis established in Chap. 8, we can subdivide the total sum of squares in Y (Σy^2) into a component due to regression and a component representing residuals, or deviations, from regression.

The total sum of squares in Y is G_{yy}/n. The component due to linear regression is:

$$\text{SS due to regression} = \frac{1}{n}(b_1 G_{1y} + b_2 G_{2y} + \cdots + b_p G_{py}) \tag{10-30}$$

Note that this is just the sum of the products of the b's and the right-hand sides of the normal equations (10-12).

The sum of squares due to the residuals from regression is the difference between these two quantities. It is possible, then, to set up the analysis-of-variance summary as shown in Table 10-5.

Table 10-5. Analysis-of-variance Summary for Multiple-regression Problems

Due to	SS	df	MS
Linear regression	$(1/n)(b_1 G_{1y} + \cdots + b_p G_{py})$	p	A
Residuals from regression	$G_{yy}/n - (1/n)(b_1 G_{1y} + \cdots + b_p G_{py})$	$n - p - 1$	E
Total	G_{yy}/n	$n - 1$	

The mean squares (SS divided by df) are indicated by A and E. The hypothesis (10-29) is tested by the F ratio:

$$F = \frac{A}{E} \tag{10-31}$$

with p and $n - p - 1$ degrees of freedom. Again, the assumption which makes this a valid test is that the ϵ_i are normally and independently distributed with mean 0 and constant variance.

As a simplification in notation, n times the sum of squares for residual may be denoted by G_{ee}, that is,

$$G_{ee} = G_{yy} - b_1 G_{1y} - b_2 G_{2y} - \cdots - b_p G_{py} \tag{10-32}$$

As an illustration, consider the data given previously.

$$\frac{1}{n}(b_1G_{1y} + \cdots + b_pG_{py}) = \tfrac{1}{20}[0.956377(109,312) + 2.003311(433,612) + 0.965165(376,032) + 0.948983(395,448)] = 85,570$$

$$\frac{G_{yy}}{n} = \frac{1,739,924}{20} = 86,996$$

Table 10-6. Analysis-of-variance Summary for the Hypothetical Regression Data

Due to	SS	*df*	MS
Linear regression	85,570	4	21,392
Residuals from regression	1,426	15	95
Total	86,996	19	

$$F = \frac{21,392}{95} = 225 \qquad \text{with 4 and 15 } df$$

This F ratio is highly significant; hence we reject the hypothesis that there is no regression in the population from which the sample was drawn.

The investigator may wonder whether a particular X variable (or group of X variables) contributes significantly to the prediction of Y. Another way of saying the same thing is to ask whether these variables contribute significantly to the reduction in residual sum of squares, because, obviously, if the residual sum of squares is reduced, the prediction is "better" in this sense.

We shall assume that the variables of doubtful value have been given the highest subscripts, so that they appear last in the computation table 10-4. The numbering of the X variables is purely arbitrary, so that any particular variable or group of variables can be placed last. If the variables to be tested are placed last, one can use the same table for the computation of the new b's.

Consider the data of Table 10-3 and suppose we wish to test the hypothesis that $\beta_3 = \beta_4 = 0$. Then, ignoring the X_3 and X_4 columns and the last two rows of the B section of the computation table, we have

$$b_2' = \frac{1,108.000}{427.480} = 2.591934$$

$$b_1' = \frac{178.870 + 223.823(2.591934)}{611.127} = 1.241975$$

The sum of squares due to the regression equation $\bar{Y}_X = a + b_1'X_1 + b_2'X_2$ is

$$\text{SS due to regression on } X_1 \text{ and } X_2 = \frac{1}{n}(b_1'G_{1y} + b_2'G_{2y})$$

$$= \tfrac{1}{20}[1.241975(109,312) + 2.591934(433,612)] = 62,983$$

These computations permit us to construct a new analysis-of-variance summary table, shown as Table 10-7.

The test of the hypothesis $\beta_3 = \beta_4 = 0$ is accomplished by

$$F = \frac{11{,}294}{95} = 119 \qquad \text{with 2 and 15 } df$$

This F ratio is highly significant, so we reject the hypothesis that X_3 and X_4 contribute nothing to the prediction of Y.

If a particular b is to be tested for significance, that variable can be placed last and the test accomplished as above. Another procedure, requiring a different method of computation, will be given later.

Table 10-7. Analysis-of-variance Summary with Test of the Hypothesis $\beta_3 = \beta_4 = 0$

Due to	SS	df	MS
Regression on X_1 and X_2	62,983	2	
Additional reduction due to X_3 and X_4	22,587	2	11,294
Regression on X_1, X_2, X_3, X_4 (previous table)	85,570	4	
Residual	1,426	15	95
Total (previous table)	86,996	19	

In simple regression the variance about the line of regression was denoted by $s^2_{Y \cdot X}$. In multiple regression we have a similar variance about the equation of regression. It is denoted by

$$s^2_{Y \cdot 12 \cdots p} = \frac{G_{yy} - b_1 G_{1y} - b_2 G_{2y} - \cdots - b_p G_{py}}{n(n-p-1)} = \frac{G_{ee}}{n(n-p-1)} \qquad (10\text{-}33)$$

It may be seen that this is simply the mean-square residual in the analysis-of-variance summary table.

Exercise 10-10. Using the data of Exercise 10-8, test the hypothesis that the special treatment B does not contribute to strength of moldings.

10-7. THE COEFFICIENT OF MULTIPLE CORRELATION

In simple correlation we obtained a correlation coefficient r which expressed the degree of linear relationship between two variables. In multiple correlation we can obtain a *coefficient of multiple correlation*, which expresses the degree of linear relationship between a variable Y and a group of variables $X_1, X_2, \ldots, X_p$. The coefficient of multiple correlation is denoted by the symbol $R_{Y \cdot 12 \ldots p}$, the subscripts indicating the variables involved. This coefficient is always positive or 0 and may have any value between 0 and 1, inclusive. The coefficient of multiple

correlation may be defined as the simple correlation between the actual Y values and the Y values estimated from the multiple-regression equation. A more useful definition for our purposes, however, is that $R_{Y\cdot 12\ldots p}$ is the square root of the fraction of the total sum of squares in Y accounted for by the regression equation. That is,

$$R_{Y\cdot 12\ldots p} = \sqrt{\frac{\Sigma(\bar{Y}_X - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}} \tag{10-34}$$

$$R_{Y\ 12\ldots p} = \sqrt{\frac{b_1 G_{1y} + b_2 G_{2y} + \cdots + b_p G_{py}}{G_{yy}}} \tag{10-35}$$

The numerator of the radicand is n times the sum of squares of deviations of the fitted points from the mean of the Y's. The denominator is n times the total sum of squares in Y. With the illustrative data of the previous section, we have

$$R_{Y\cdot 1234} = \sqrt{\frac{85{,}570}{86{,}996}} = \sqrt{0.9836} = 0.992$$

The correlation is exceptionally high. Again, it may be seen that $R^2_{Y\cdot 12\ldots p}$ has more meaning than $R_{Y\cdot 12\ldots p}$. The square of the above coefficient indicates that 98 per cent of the variation in Y can be accounted for by linear regression on the variables X_1, X_2, X_3, and X_4.

The regression on the first two variables only (from the last illustration of the previous section) shows that

$$R_{Y\cdot 12} = \sqrt{\frac{62{,}983}{86{,}996}} = \sqrt{0.7239} = 0.851$$

The coefficient of correlation was increased from 0.85 to 0.99 by adding variables X_3 and X_4.

A test of the hypothesis that the population coefficient of multiple correlation, $\rho_{Y\cdot 12\ldots p}$, equals 0 is provided by the F test:

$$F = \frac{\text{variance explained by regression equation}}{\text{residual variance}}$$

The degrees of freedom are p and $n - p - 1$.

Other tests of hypotheses concerning $\rho_{Y\cdot 12\ldots p}$ can be made by using Fisher's z transformation in the manner described in Chap. 7. The only modification of those procedures is that in the multivariate case the standard error of z must take account of the additional variates. That is,

$$\sigma_z = \frac{1}{\sqrt{n - p - 2}} \tag{10-36}$$

EXERCISE 10-11. (*a*) Use only X_1, X_2, and Y of Exercise 10-8 and compute the coefficient of multiple correlation. (*b*) Test it for significance ($\alpha = 0.01$), using the F ratio and (*c*) Fisher's z transformation.

10-8. THE GAUSS MULTIPLIERS

Another method of computation which has certain advantages for statistical inference is based upon computation of the inverse of the matrix of G's. The elements of this inverse are denoted by c_{ij} and are called the *Gauss multipliers.*

Referring to the matrix form of the normal equations (10-12), we see that the b's may be found by the expression $B = G^{-1}Y_X$. That is,

$$\begin{matrix} B \\ \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_p \end{bmatrix} \end{matrix} = \begin{matrix} G^{-1} \\ \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix} \end{matrix} \times \begin{matrix} Y_X \\ \begin{bmatrix} G_{1y} \\ G_{2y} \\ \cdots \\ G_{py} \end{bmatrix} \end{matrix} \tag{10-37}$$

Then,

$$\begin{aligned} b_1 &= c_{11}G_{1y} + c_{12}G_{2y} + \cdots + c_{1p}G_{py} \\ b_2 &= c_{21}G_{1y} + c_{22}G_{2y} + \cdots + c_{2p}G_{py} \end{aligned} \tag{10-38}$$

by the rules of matrix multiplication.

We see then that the problem reduces itself to finding the inverse of the G matrix. The computation procedure is outlined below for the symbols of Table 10-8.

Table 10-8. Computation Table for Inverse of the G Matrix

Line	X_1	X_2	X_3	Y	Identity			S
1	G_{11}	G_{12}	G_{13}	G_{1y}	1	0	0	G_{1s}
2		G_{22}	G_{23}	G_{2y}	0	1	0	G_{2s}
3			G_{33}	G_{3y}	0	0	1	G_{3s}
4	B_{11}	B_{12}	B_{13}	B_{1y}	I_{11}	0	0	B_{1s}
5		B_{22}	B_{23}	B_{2y}	I_{21}	I_{22}	0	B_{2s}
6		.	B_{33}	B_{3y}	I_{31}	I_{32}	I_{33}	B_{3s}
7	b_1	b_2	b_3					
8	c_{11}	c_{12}	c_{13}					
9		c_{22}	c_{23}					
10			c_{33}					

LINES 1 TO 6. Precisely the same computation is employed as was described for Table 10-2, including the check column. There are three

added columns, however, representing the identity matrix (Sec. 10-2). For example,

$$I_{32} = \frac{0 - 0B_{13} - I_{22}B_{23}}{B_{33}}$$

LINE 7. The b's are computed exactly as before.

LINE 8:

$$c_{13} = \frac{I_{31}}{B_{33}}$$

$$c_{12} = \frac{I_{21} - B_{23}c_{13}}{B_{22}}$$

$$c_{11} = \frac{I_{11} - B_{13}c_{13} - B_{12}c_{12}}{B_{11}}$$

LINE 9:

$$c_{23} = \frac{I_{32}}{B_{33}}$$

$$c_{22} = \frac{I_{22} - B_{23}c_{23}}{B_{22}}$$

LINE 10:

$$c_{33} = \frac{I_{33}}{B_{33}}$$

The C matrix is symmetric, that is, $c_{ij} = c_{ji}$, so that the elements below the diagonal are not shown.

The Y column can be omitted because one can obtain the b's by formula (10-38). In order to use the c's for this purpose, one must carry their computation to as many significant digits as he requires the value of the b's. This computation is tedious and, as we shall see, the other use we make of the c's does not require such accuracy. For this reason the additional column for the Y's is recommended.

A check on the accuracy of the C matrix is made by multiplying the G matrix by its inverse, as follows:

$$\underset{G}{\begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix}} \times \underset{G^{-1}}{\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}} = \underset{I}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \tag{10-39}$$

That is,

$$G_{11}c_{11} + G_{12}c_{21} + G_{13}c_{31} = 1 \tag{10-40}$$

and so forth.

The principal reason for computing the c's is to utilize them in testing hypotheses (or computing confidence intervals) related to the b's.

The standard error of b_j is the standard error of estimate, $s_{Y.12\ldots p}$, times

the square root of the corresponding diagonal c value, times n. That is,

$$s_{b_j} = s_{Y \cdot 12 \ldots p} \sqrt{c_{jj} n} \tag{10-41}$$

If the error ϵ_i is normally distributed, the regression coefficient b_j is normally distributed with mean β_j and variance $s^2_{b_j}$; hence a valid test of the hypothesis $\beta_j = \beta_0$ is provided by:

$$t = \frac{b_j - \beta_0}{s_{b_j}} \tag{10-42}$$

with $n - p - 1$ degrees of freedom.

The standard error of the difference between two b's (say, b_1 and b_2) is

$$s_{(b_1 - b_2)} = s_{Y \cdot 12 \ldots p} \sqrt{(c_{11} + c_{22} - 2c_{12})n} \tag{10-43}$$

Again, for the t test, there are $n - p - 1$ degrees of freedom.

In simple regression, we used the standard error of an estimated Y value for a given value of X [formula (7-18)]. In multiple regression we have a similar formula, which is illustrated as follows for a three-variable problem:

$$s_{Y_e} = s_{Y \cdot 123} \sqrt{\frac{1}{n} + (c_{11}x_1^2 + c_{22}x_2^2 + c_{33}x_3^2 + 2c_{13}x_1x_3 + \cdots + 2c_{23}x_2x_3)n} \tag{10-44}$$

The x's are deviations from the means; that is, $x_1 = X_1 - \bar{X}_1$. The formula becomes quite complex for large numbers of independent variables. However, if the number of observations is large, or if the particular x's chosen are close to their means, the formula becomes, approximately,

$$s_{Y_e} = s_{Y \cdot 123} \sqrt{\frac{1}{n}} \tag{10-45}$$

This approximation is commonly used.

If one wishes to estimate a particular Y value, the standard error is found by replacing the quantity under the radical sign of (10-44) by 1 plus that quantity.

The computation of the Gauss multipliers (the inverse of the G matrix) is demonstrated in Table 10-9. The check column has been omitted to save space. The data are the first three X variables from Table 10-4. The following computations illustrate the use of the Gauss multipliers.

$$s^2_{Y \cdot 123} = \frac{1}{n(n-4)} (G_{yy} - b_1G_{1y} - b_2G_{2y} - b_3G_{3y})$$

$$= \frac{1}{320} [1{,}739{,}924 - 1.281637(109{,}312) - 2.279648(433{,}612) - (-0.856798)(-376{,}032)] = 903.6 \tag{10-33}$$

$$s_{b_2} = \sqrt{s^2_{Y\cdot 123} c_{22} n}$$
$$= \sqrt{903.6(0.0^5598258)(20)}$$
$$= 0.329 \qquad (10\text{-}41)$$

A test of the hypothesis that $\beta_2 = 2$ is provided by

$$t = \frac{b_2 - \beta_2}{s_{b_2}} = \frac{2.280 - 2.000}{0.329} = 0.85 \qquad (10\text{-}42)$$

with 16 degrees of freedom. We cannot reject the hypothesis on the basis of our data.

Now, suppose we have the following observations on an individual:

$$X_1 = 30 \qquad X_2 = 60 \qquad X_3 = 40$$

We wish to place 90 per cent confidence limits on an estimate of his Y value. First of all, we estimate his Y value as

$$\hat{Y}_X = a + b_1X_1 + b_2X_2 + b_3X_3$$
$$= 110.1 + 1.282(30) + 2.280(60) - 0.857(40) = 251.2$$

The standard error of the estimate of the individual Y value is provided by

$$s_{\hat{Y}} = s_{Y\cdot 123}\sqrt{1 + \frac{1}{n} + n(c_{11}x_1^2 + c_{22}x_2^2 + c_{33}x_3^2 + 2c_{12}x_1x_2 + 2c_{13}x_1x_3 + 2c_{23}x_2x_3)}$$

$$= 30.6\sqrt{1 + \tfrac{1}{20} + 20[0.0^5342(30 - 57.1)^2 + (0.0^5598)(60 - 41.1)^2 + (0.0^5384)(40 - 56.9)^2 + 2(0.0^5194)(30 - 57.1)(60 - 41.1) + 2(-0.0^6178)(30 - 58.1)(40 - 56.9) + 2(0.0^5110)(60 - 41.1)(40 - 56.9)]}$$

$$= 306\sqrt{1 + 0.05 + 0.054}$$
$$= 32.7 \qquad (10\text{-}46)$$

With 16 degrees of freedom $t_{(0.10)} = 1.75$. Therefore the required confidence interval is

$$P_c[\hat{Y}_X - t_{(0.1,16)}s_{\hat{Y}} < \mu_X < \hat{Y}_X + t_{(0.1,16)}s_{\hat{Y}}] = 0.90$$
$$251.1 - 1.75(32.7) < \mu_X < 251.1 + 1.75(32.7)$$
$$193.9 < \mu_X < 308.3$$

The limits appear to be quite wide. That is, if X_1, X_2, and X_3 are placement tests and Y is a performance score, we cannot tell within very narrow limits how well the applicant will perform on the job. This situation is fairly typical, however. The truth is that, although we can predict the average behavior of a group of unrelated individuals quite accurately, we cannot predict individual behavior within very narrow limits.

A point of interest in the computation is the small contribution of the third term under the radical in (10-46). This also is typical, and with

Table 10-9. Computation of Inverse of G Matrix (Hypothetical Data)

X_1	X_2	X_3	Y	Identity		
373,476	−136,784	67,144	109,312	1	0	0
	232,836	−91,196	433,612	0	1	0
		296,676	−376,032	0	0	1
611.127	−223.823	109 869	178 870	$.0^{2}163632$*	0	0
	427.480	−155.808	1108 000	$0^{3}856756$	$.0^{2}233929$	0
		510.224	−437.159	$-.0^{4}907276$	$.0^{3}714353$	$.0^{2}195992$
1.281637	2.279648	−.856798				
$.0^{5}341981$	$.0^{5}193939$	$-.0^{6}177819$				
	$.0^{5}598258$	$.0^{5}140009$				
		$.0^{5}384129$				

* The exponent on 0 indicates the number of 0's to the left of the first significant digit. That is, $0.0^{2}163632$ is read 0.00163632, and so forth.

somewhat larger n one can safely ignore this term as well as the second, thus replacing (10-46) by

$$s_{\hat{Y}} = s_{Y.123} \tag{10-47}$$

Exercise 10-12. (*a*) Use the computing method of Table 10-8 on the data of Exercise 10-9. (*b*) Test the hypothesis that $\beta_3 = 0$. (*c*) Place 90 per cent confidence limits on Y if $X_1 = 12$, $X_2 = 20$, and the new treatment B is used.

10-9. COMPUTATION FROM SIMPLE CORRELATION COEFFICIENTS

We have seen that we must carry computations to a constant number of significant digits (rather than a fixed number of decimal places) in order to assure accuracy within prescribed limits. This procedure is somewhat tedious and requires special handling if one employs electronic computing machinery. It is simpler computationally to retain a fixed number of decimal places (rather than significant digits).

If one divides all the G's in each row of the matrix by the diagonal element, that is, G_{ii}, he can usually carry all computations to a fixed number of decimal places. Another technique, frequently employed, is to perform all computations with simple correlation coefficients rather than the G's.

If one uses correlation coefficients r_{ij} in Table 10-2 rather than the G_{ij}, the solutions at the bottom of the table are B_1, B_2, B_3, and B_4 rather than

b_1, b_2, b_3, and b_4 (see Sec. 10-5). Then one can find the usual regression coefficients by the formula

$$b_j = B_j \sqrt{\frac{G_{yy}}{G_{jj}}} \tag{10-48}$$

Many other computing methods are available. Knowledge of them and acquaintance with their characteristics are a must for the practitioner but are ordinarily not required of the business manager who is the ultimate user of the computations. The methods employed in this chapter should provide both a basis for understanding regression and correlation problems and adequate methods of computation for all common regression problems.

EXERCISE 10-13. Redo Exercise 10-8 using simple correlation coefficients in the computing form.

10-10. PARTIAL CORRELATION

The regression coefficients b_j have been described as partial regression coefficients, that is, coefficients which show the influence of a variable X_j on Y when the values of the remaining variables are held constant. A correlation coefficient which shows the correlation between Y and an "independent" variable, say, X_2, when the other X's are held constant is called a *coefficient of partial correlation* and is denoted by $r_{Y2.13\ldots p}$.

It will be convenient to attach subscripts to the b's indicating the dependent and independent variables. That is,

$$\begin{aligned} b_1 &= b_{Y1.23\ldots p} \\ b_2 &= b_{Y2.13\ldots p} \end{aligned} \tag{10-49}$$

and so on.

The first subscript indicates the dependent variable; the second, the particular independent variable upon which the regression is computed; and the subscripts to the right of the decimal point indicate the variables held constant. It is possible to consider some variable other than Y as the dependent variable. Suppose X_1 is considered dependent. Then solution of the revised normal equations will yield

$$\begin{gathered} b_{1Y.23\ldots p} \\ b_{12.Y3\ldots p} \\ b_{13.Y2\ldots p} \end{gathered}$$

and so on.

If this computation is made, one can find the partial correlation coefficient $r_{Y1.23\ldots p}$ by the following formula:

$$r_{Y1.23\ldots p} = \sqrt{b_{Y1.23\ldots p} b_{1Y.23\ldots p}} \tag{10-50}$$

This method of computing the partial correlation coefficients is not very practical, since it requires another solution of the normal equations. More useful methods are available, but are not presented here, since in practice very little use is made of the partial correlation coefficients.

10-11. APPLICATION OF LINEAR REGRESSION TO ECONOMIC FORECASTING

Many attempts have been made to predict such economic series as sales, stock market prices, and general business conditions by means of multiple-regression methods. The theory is quite simple. If we want to predict stock prices next month, we find some X variables which are available this month (or last) which are related to next month's stock prices. The difficulties are numerous. First, it is difficult (and many times impossible) to find X variables for which data are available in time to do any good in the prediction of Y. Second, predictive power of the X variables tends to be low, because of the many factors, such as political changes, world tensions, and tax changes, that cannot be predicted and that influence the Y variable more than the X variables which appear in the equation.

Evidence of the fallibility of the method is that all statisticians do not become independently wealthy playing the stock market. On the encouraging side is some evidence indicating fair success in predicting series for a firm or industry whose growth is determined by known factors outside the industry. For example, demand for children's clothing for next year can be determined quite accurately from the known population of children this year, just as need for classroom space for high school pupils can be predicted accurately many years in advance.

In recent years much work has been done in the analysis of econometric models in an attempt to explain mathematically the workings of the total economy or a segment of it. Instead of the simple linear-regression model in which a Y variable is dependent on several X variables, the econometric model encompasses a set of variables, some of which are determined outside the system (independent) and some of which are determined within the economic system as functions of the other variables. The construction and analysis of econometric models constitute a huge field in themselves, and only passing reference can be made to them here.*

* For an introductory text in the field, consult Lawrence R. Klein, *A Textbook of Econometrics,* Row, Peterson & Company, Evanston, Ill., 1953.

11. SAMPLING

In one sense, most of statistics concerns itself with sampling. When we compute a sample mean, for example, we try to make some generalizations about the population mean, either by placing confidence limits or by testing hypotheses about the population mean. In this chapter we are concerned primarily with *survey sampling*. In this type of sampling the investigator examines some portion of a finite population and then makes generalizations about the entire population. The characteristic which distinguishes methods discussed here from those presented in the first part of this book is that here we are dealing almost exclusively with finite populations.

We have defined the variance of a finite population* of size N as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N} \tag{11-1}$$

where μ is the true mean of the population [see Eq. (4-5)]. Now, suppose we have two populations—P_1, which is finite, and P_2, which is infinite—both having the same variance σ^2. If we draw a sample of size n from each population and compute the mean, the variance of the mean drawn from P_1 will be smaller on the average than the variance of the mean drawn from P_2.

We can illustrate the principle by a simple numerical example. Let a population consist of the 3 values 1, 2, and 3 only, that is, $N = 3$. We shall draw samples of $2(n = 2)$. We define P_1 to be the above population after specifying that we *cannot* replace the first sample value before

* For infinite populations one can define the variance in terms of an integral as follows: $\sigma^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f(y)\, dy$, where $f(y)$ is the probability density function of the variable y.

drawing the second. We define P_2 to be the above population provided that we *can* replace each value after it is drawn. This has the effect of making P_2 inexhaustible, or infinite. We consider in Table 11-1 all the possible samples that can be drawn under both assumptions.

Actually we need not have computed the variance of the column of means. We could have observed that the first group of means varies from 1.5 to 2.5, a range of 1.0, and that the second group varies from 1.0 to 3.0, a range of 2.0. It is interesting as well as instructive to note that the means of both groups have the same *average*. This tells us that

Table 11-1. Comparison of Sample Variances from Finite and Infinite Populations

Finite population		Infinite population	
Sample values	Mean	Sample values	Mean
1, 2	1.5	1, 1	1.0
1, 3	2 0	1, 2	1.5
2, 1	1.5	1, 3	2.0
2, 3	2.5	2, 1	1.5
3, 1	2.0	2, 2	2.0
3, 2	2.5	2, 3	2.5
		3, 1	2.0
		3, 2	2.5
		3, 3	3.0
Variance of the mean = $\frac{1}{6}$		Variance of the mean = $\frac{1}{3}$	
Average of the mean = 2.0		Average of the mean = 2.0	

on the average we shall get the same value whether we sample with replacement or without replacement (assuming that all values are equally likely to be drawn).

The factor by which the variance of the mean is reduced in sampling from finite populations is called the *finite-population correction* (f.p.c.). The correction becomes insignificant if the ratio n/N (sample size to population size) is very small. More is said about this later.

Exercise 11-1. Verify the computation of the variances in Table 11-1.

11-1. RANDOM SAMPLING

Whenever the word "sample" is used, it implies a sample drawn in a random manner, or at least in a manner which is "reasonably" random.

If the population is numbered in some manner, good use can be made of a table of random numbers to identify the units to be included in the

sample. For example, if one wishes to draw a sample of automobile licenses numbered from 1 to 18,439, he may use five columns in a table of random digits (see Table 1-1) to identify the licenses to be drawn. If a number larger than 18,439 is drawn, it is ignored. Also, all duplicate drawings are ignored, so that drawing is without replacement.

Sometimes we do not draw at random even though the elements in the population are already numbered for us. For example, suppose we need a sample of 1,000 policyholders from a population of 300,000. Drawing and identifying the sample would be a major task. We are more likely to take a systematic sample, that is, to take, say, every three-hundredth unit. To do this properly, we draw a random number from 1 to 300 to identify the first element and then take every three-hundredth unit thereafter. If the units to be sampled appear in the form of cards in a file drawer, one can locate every kth card(approximately) by measuring rather than counting. He may push two properly spaced pins through a piece of cardboard and use this as a measuring device.

A word of caution is in order here. Although systematic sampling works well if the sample units are reasonably random or "scrambled," it generally should not be used if there is a systematic pattern in the population. For example, if one wishes to sample tourist traffic during the days of the summer season and chooses every seventh day, he will sample only 1 day of the week.

11-2. THE SAMPLING FRAME

Frequently a population is in such form that we cannot sample at random from it. Consider, for example, the impossibility of drawing a sample at random from the inhabitants of New York City. Sometimes it is necessary for us to make compromises and select something other than a purely random sample. In any case, it is necessary to arrange the population elements (at least conceptually) in such a manner that they can be identified. This process is referred to as *constructing the frame.* In some cases we may construct the frame by delineating small areas on a map, so that each sample unit is identified by whether or not it lies in an area.

For a sample of the residents of a city it is customary to identify areas by city blocks and then to draw a random sample of blocks. Once a block is chosen, every person in the block may be included in the survey. This is referred to as *cluster sampling.* Note that we do not abandon entirely the concept of randomness, but here we apply it to groups of individuals rather than to the individuals themselves. Statistical techniques have been devised, and are presented later, for cases such as this.

11-3. SIMPLE RANDOM SAMPLING

We shall use the term *simple random sampling* to denote a sampling scheme which selects units at random from a *finite population with equal probability and without replacement.* The true mean of the finite population is defined by

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{11-2}$$

and the sample mean is

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{11-3}$$

As demonstrated by the numerical illustration of Table 11-1, $\bar{Y}$ is an unbiased estimate of μ. That is, the average value of $\bar{Y}$ over all possible samples is μ, the mean of the population.

We have previously used

$$\sigma^2 = \frac{\Sigma(Y_i - \mu)^2}{N} \tag{11-1}$$

as a definition of the finite-population variance. It is more convenient in finite populations to define a parameter

$$S^2 = \sum_{i=1}^{N} \frac{(Y_i - \mu)^2}{N - 1} \tag{11-4}$$

The difference is that in (11-4) we divide by $N - 1$ and in the formula for σ^2 (11-1) we divide by N. In both cases N is the number of units in the population.

The variance of the mean of a sample of size n from a finite population of size N is

$$V(\bar{Y}) = \frac{N - n}{N} \frac{S^2}{n} \tag{11-5}$$

Note the similarity between this expression and the familiar variance of the mean of an infinite population. We divide the variance by the sample size, as usual, and then multiply by $(N - n)/N$, which is called the *finite-population correction.* Inspection of the formula shows that, as n approaches N in size, the variance of the mean approaches 0. This is to be expected, since if one samples the whole population his results are not subject to variation from trial to trial. Also, it may be observed that if n is very small with respect to N the f.p.c. (finite-population correction) approaches 1, so that the variance of the mean becomes simply S^2/n.

EXERCISE 11-2. Using the numerical data of Table 11-1, verify that Eq. (11-5) yields the variance of the mean for the finite population given and that σ^2/n yields the variance of the mean of the infinite population.

Although (11-5) is the formula for the variance of the mean, it is of little value to us in an actual sampling situation, because ordinarily we do not know S^2. However, we can estimate S^2 by

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1} \tag{11-6}$$

Therefore an estimate of $V(\bar{Y})$ is

$$v(\bar{Y}) = \frac{N - n}{N} \frac{s^2}{n} \tag{11-7}$$

This estimate of the variance is an unbiased estimate. That is, its average value over all possible samples is equal to (11-5). It will be noted that (11-7) is just the square of (4-17).

Sometimes the population total T rather than the population mean is the quantity we wish to estimate. For example, we may wish to estimate the total consumption of milk in an area, rather than the average consumption per family. It is clear that, if we know the population mean and the total number of sample units in the population, we can compute the population total by

$$T = N\mu \tag{11-8}$$

If we have estimated μ by $\bar{Y}$, then an estimate of the population total is provided by

$$\hat{T} = N\bar{Y} \tag{11-9}$$

The variance of $\hat{T}$ is simply N^2 times the variance of $\bar{Y}$. That is,

$$V(\hat{T}) = N(N - n) \frac{S^2}{n} \tag{11-10}$$

An unbiased estimate of this variance is

$$v(\hat{T}) = N(N - n) \frac{s^2}{n} \tag{11-11}$$

EXERCISE 11-3. A sample of 100 accounts receivable from a total of 500 accounts shows an average of \$20 per account past due more than 30 days. The sample variance [Eq. (11-6)] is 500. (*a*) What is the standard error of the mean? (*b*) What is your estimate of the total amount of accounts past due 30 days? (*c*) What is the standard error of this estimate? (*d*) Place 90 per cent confidence limits on this estimate. (*e*) What assumptions did you make? Do you think they are justified?

11-4. RATIO ESTIMATION IN SIMPLE RANDOM SAMPLING

In this section we discuss a method of estimation which, in many situations, will increase the precision of sample estimates. The method requires that we use an auxiliary variable X, about which we have rather detailed information, to adjust, or *calibrate*, our estimates of μ or T.

Assume that the variable of interest is Y and that it is highly correlated with the variable X, for which we know the population total T_X. Then an estimate of the total of the Y population from a sample of n observations is

$$\hat{T}_r = \frac{\bar{Y}}{\bar{X}} T_X \tag{11-12}$$

This estimate, $\hat{T}_r$, is "better" in the sense of having a smaller variance than the simple estimate of the previous chapter if the correlation between X and Y is high. Just how high the correlation must be is shown later.

As an illustration, let us consider a problem in estimation of inventories. Suppose a perpetual inventory is kept on $N = 10{,}000$ items. The necessary information, including current balance and price per unit for inventory-evaluation purposes, is recorded on magnetic tape. For convenience, we may assume that the lifo (last-in–first-out) method is used. The tapes are run weekly, recording receipts, withdrawals, and new balances. At the same time, orders for items for which the balance has fallen below the reorder point are printed automatically. An evaluation of the inventory can be obtained at any time by running the tape so as to multiply balances by prices and accumulate the dollar value of inventory.

In an audit at the end of the year we wish to check the inventory. We decide actually to count the inventory of 500 items and to price it out. Then we shall estimate the total inventory from this sample.

Let Y_i = actual count of the ith item times price per unit
X_i = balance according to the tape of the ith item times price per unit
T_X = machine run of the inventory evaluation

Then our revised estimate of the end-of-year inventory, in dollars, is

$$\hat{T}_r = \frac{\bar{Y}}{\bar{X}} T_X$$

This estimate will be more precise than the simple estimate if Y_i and X_i are sufficiently correlated. We should certainly expect them to be correlated if our perpetual inventory system is working satisfactorily.

It is entirely possible that most errors in inventory are made on one side of the ledger, either in failure to record receipts or in failure to record

withdrawals. Suppose that most errors are due to failure to reduce inventory when a sale is made. Then ΣY_i will tend, on the average, to be smaller than ΣX_i, so that the ratio $\bar{Y}/\bar{X}$ will have the effect of reducing the computed inventory $\hat{T}_r$. In this sense, the ratio is a calibrating factor.

The estimate $\hat{T}_r$ will be unbiased if, on the average, it is equal to T. If this is true, the ratio estimate will yield the population total, on the average. Unfortunately this is not always true for the ratio estimate, because the average value of a ratio is not, in general, equal to the ratio of the average values. In most applied sampling situations, the bias is of low order, so that the error due to bias is more than offset by a reduction in the variance of the estimator (assuming, of course, that X and Y are sufficiently correlated).

The bias tends to be of low order if the correlation between $\bar{Y}/\bar{X}$ and $\bar{X}$ is of low order. In terms of our inventory illustration, this means that the percentage by which inventory is overstated is not related to the dollar value of the inventory item. It seems reasonable to suppose that this correlation is not high and, therefore, that the bias is of low order.

The exact formula for the variance of a ratio estimate involves an infinite series and is therefore not very useful in the applications. We can use the first few terms of the series and arrive at a useful approximation, however, as follows:

$$V(\hat{T}_r) = \frac{N(N-n)}{n}(S_Y^2 + R^2 S_X^2 - 2R\rho S_Y S_X) \tag{11-13}$$

where $S_Y^2 = \sum^N \frac{(Y_i - \mu_Y)^2}{N-1}$ (and S_X^2 is similarly defined for X) (11-14)

$$R = \frac{\mu_Y}{\mu_X} \tag{11-15}$$

ρ = population correlation between X and Y

$$= \frac{\sum^N (Y_i - \mu_Y)(X_i - \mu_X)}{\sqrt{\sum^N (Y_i - \mu_Y)^2 \sum^N (X_i - \mu_X)^2}} = \frac{\text{cov } XY}{S_X S_Y} \tag{11-16}$$

Note that the f.p.c. is the same as in the simple estimate of the total.

Formula (11-13) is a population formula. That is, it expresses the variance in terms of the parameters of the population. The population parameters involved are R, ρ, S_X, and S_Y. Since we seldom, if ever, know these parameters, the formula is of use to us principally in theoretical work for comparing various methods of estimation. When we consider estimates of the variance, we are in the embarrassing position of having available no unbiased estimate of (11-13).

An estimated variance which is subject to bias no greater than the order of $1/n$ is as follows:

$$v(\hat{T}_r) = \frac{N(N-n)}{n(n-1)}\left(\sum^{n} Y_i^2 + \hat{R}^2 \sum^{n} X_i^2 - 2\hat{R} \sum^{n} Y_i X_i\right) \quad (11\text{-}17)$$

where $\hat{R} = \bar{Y}/\bar{X}$. Note that the sums of squares and cross products in (11-17) are uncorrected. That is, we need not take account of the means in computing them.

If one wishes a ratio estimate of the average, rather than the total, he computes:

$$\bar{Y}_r = \frac{1}{N}\hat{T}_r$$

To find the variance, one divides (11-13) or (11-17), as the case may be, by N^2.

Exercise 11-4. We wish to estimate the number of television sets in a community. We decide to let our sample unit be the city block. We send out a survey crew that estimates the number of dwelling units in every block of the community. The estimate is made by driving around each block and estimating the number of dwelling units from the number of houses and apartment buildings. This estimate is the X variable. There are 800 city blocks in the community, and the estimated total of dwelling units is 7,500. Next, we draw a random sample of 60 blocks and survey them completely to find out how many television sets are on each block. This is the Y variable. The results of the survey follow:

Y	X	Y	X	Y	X
3	5	6	8	4	6
5	8	6	10	6	11
5	16	7	11	5	4
8	10	6	10	12	19
12	25	6	9	4	6
3	4	11	16	6	7
5	9	2	3	4	8
7	15	9	20	9	21
6	16	2	4	4	2
5	10	5	5	1	1
4	5	7	8	4	5
4	4	6	9	6	9
3	5	5	7	5	8
7	10	5	12	6	11
7	12	8	11	8	16
7	10	4	10	8	16
4	9	3	4	3	3
9	16	4	3	7	23
3	8	6	20	2	4
2	3	5	10	7	15

(*a*) Ignoring the X variable, estimate the total number of television sets in the community and compute the standard error. (*b*) Using a ratio estimate, find an estimate of the total and its standard error.

11-5. CONDITIONS UNDER WHICH THE VARIANCE OF A RATIO ESTIMATE IS LESS THAN THE VARIANCE OF A SIMPLE ESTIMATE

For convenience, we consider estimates of the total, rather than the mean. Referring to formulas (11-10) and (11-13), we specify that (11-10) is to be larger than (11-13):

$$V(\hat{T}) > V(\hat{T}_r)$$

that is,

$$N(N-n)\frac{S^2}{n} > \frac{N(N-n)}{n}(S_Y^2 + R^2S_X^2 - 2R\rho S_Y S_X)$$

It is clear that this relationship will hold whenever $R^2S_X^2 - 2R\rho S_Y S_X < 0$. Solving the inequality for ρ, we have

$$\rho > \frac{R^2S_X^2}{2RS_Y S_X} = \frac{1}{2}R\frac{S_X}{S_Y} = \frac{1}{2}\frac{S_X/\mu_X}{S_Y/\mu_Y} \tag{11-18}$$

Therefore, one will gain in terms of reduced variance whenever the population coefficient of correlation between X and Y is greater than $\frac{1}{2}$ the ratio of the coefficients of variation (ratio of X to Y). If the coefficients of variation are nearly equal, as is frequently the case, then one will gain by ratio estimation whenever the correlation is greater than $\frac{1}{2}$.

Suppose we have a population consisting only of the following 3 values: 1, 10, 100. Suppose then that we draw a random sample of 2 values and estimate the population total from the 2 values by multiplying their sum by $\frac{3}{2}$. The total possible samples and their corresponding estimates of the population total follow:

Sample values	*Estimate of the total*
1, 10	16.5
1, 100	151.5
10, 100	165.0

The average of the above estimates is 111, which is precisely the population total. We could have anticipated this, since $\bar{Y}$ is an unbiased estimate of μ. The variance of the above estimates is

$$\frac{(16.5-111)^2 + (151.5-111)^2 + (165.0-111)^2}{3} = 4{,}496$$

Now, assume that an auxiliary variable is associated with each of the above population values as follows:

Population value, Y	*Auxiliary variable, X*
1	1
10	8
100	110
111	119

The X variable follows in general the same pattern as Y. Now let us draw a sample of 2 and estimate the total by $\hat{T}_r = (\bar{Y}/\bar{X})T_X$, where $T_X = 119$. Our samples and the possible estimates are as follows:

Sample		Estimate
Y	X	
1, 10	1, 8	$(\frac{1}{9}\frac{1}{1})119 = 145\ 4$
1, 100	1, 110	$(\frac{1}{1}\frac{0}{1}\frac{1}{1})119 = 108.3$
10, 100	8, 110	$(\frac{1}{1}\frac{1}{1}\frac{0}{8})119 = 110.9$

The average of these estimates is 121.5, so the ratio estimate is no longer an unbiased estimate of the total, 111. However, the variance of the estimates is now 285.9, compared with 4,496 for the simple estimate above. We give up the feature of unbiasedness, but we gain in precision of estimates.

EXERCISE 11-5. Assume the following population values of Y and X:

Y	X
10	30
40	20
80	40

(*a*) Draw all possible samples of 2 and compare the variance of the ratio estimate with the variance of the simple estimate. (*b*) Compute the bias of the ratio estimate.

11-6. MULTISTAGE SAMPLING

Frequently it is virtually impossible to identify the units in the population before the sample is drawn. Suppose, for example, that we wish to sample the residents of a city to obtain income or opinion data. There

is no feasible way that we can prepare a list of such persons. It is relatively easy to identify them, however, by their place of residence. Therefore, we may mark off the city into blocks or subdivisions. Then, within the sampled blocks, we can draw a random sample of households (after proper definition of the term "household"). We can either include the entire household as a unit or perhaps draw another random sample within each sampled household. In the latter case we say that we have drawn a *three-stage sample.* The sample unit for the first stage is the city block. This is called the primary unit. The secondary unit (the household) is drawn at the second stage. The tertiary (or ultimate) unit is the individual drawn at the third stage. The process can be extended, with resultant increase in complexity of estimators and variances. The problem of constructing the sample frame is simplified. The primary units are defined by areas marked off on a map of the city. After a sample of these is taken, a list is made of the households in each of the sampled blocks. This is a much less prodigious task than listing every household in the city. After the next stage we need only select the individuals at random within the households.

We consider first a situation in which we have two stages. We let N equal the number of primary units in the population and assume that each of these primary units contains M_i secondary units. Then $\Sigma M_i = M$, the total sample units in the population, where the summation extends over all N primary units. We draw at random a sample of n primary units, and within each of these we draw m_i secondary units, so that our total sample size is $\sum_{i}^{n} m_i = m$. Several situations are possible within this general framework.

Sampling without replacement and with equal probabilities. We find an estimate of the population total as follows:

$$\hat{T} = \frac{N}{n} \sum_{i} \frac{M_i}{m_i} \sum_{j} Y_{ij} \tag{11-19}$$

where Y_{ij} is the jth observation in the ith primary unit. Note that $(1/m_i) \sum_{j} Y_{ij}$ is, in fact, the sample mean of the ith primary unit and that this quantity multiplied by M_i is an estimate of the ith primary total. Then the estimated primary totals summed over the sampled primaries and divided by n form an estimate of the average primary total. This quantity multiplied by N yields an estimate of the population total.

The variance of this estimate becomes excessively high if the primary units vary greatly in size. The reason for this becomes obvious upon careful inspection of the following variance formula:

$$V(\hat{T}) = \frac{N}{n}\sum_i M_i(M_i - m_i)\frac{S_i^2}{m_i} + \frac{N(N-n)}{n}\sum_i \frac{(T_i - T^*)^2}{N-1} \tag{11-20}$$

where $$S_i^2 = \sum_{j=1}^{M_i} \frac{(Y_{ij} - \mu_i)^2}{M_i - 1} \tag{11-21}$$

$$T^* = \frac{\sum_i T_i}{N} \tag{11-22}$$

The second term of (11-20) dominates the expression. Note that it is a constant times the sum of the squares of deviations of true primary totals from the average primary total. Hence, if primary totals vary greatly in size, the variance is excessively large. When one employs this method of sampling, he should try to make the primary units as nearly equal in size as possible.

One may obtain an unbiased estimate of the variance by the following formula:

$$v(\hat{T}) = \frac{N}{n}\sum_i^n M_i(M_i - m_i)\frac{s_i^2}{m_i} + \frac{N(N-n)}{n}\sum_i^n \frac{(M_i\bar{Y}_i - Y^*)^2}{n-1} \tag{11-23}$$

where $$Y^* = \sum_i^n \frac{M_i\bar{Y}_i}{n}$$

Note that Y^* is an estimated average primary total.

Stratified random sampling. If one decides to take every primary unit into the sample, then $n = N$, and one has *stratified sampling*. In effect, the population is subdivided into N subpopulations, and a sample is drawn at random within each of these. The result is greatly increased precision of estimates. Let us see what change is required in the formulas. For the estimated total [referring to (11-19)] we have $n = N$; hence

$$\hat{T} = \sum_i \frac{M_i}{m_i}\sum_j Y_{ij} \tag{11-24}$$

For the population variance of a total, we have

$$V(\hat{T}) = \sum_i M_i(M_i - m_i)\frac{S_i^2}{m_i} \tag{11-25}$$

Note that the second term disappears because $N - n = 0$. Since the second term tends to be large in ordinary two-stage sampling, we have

accomplished a major reduction in variance by stratification. The estimate of the variance is

$$v(\hat{T}) = \sum_i M_i(M_i - m_i) \frac{s_i^2}{m_i} \tag{11-26}$$

Actually, the comparison given above between stratified sampling and two-stage sampling is not quite fair. If one plans to stratify, he constructs his primary units (strata) in such a way that the sample units in each stratum are as nearly alike as possible. This generally means making the strata as different from each other as possible. In two-stage sampling, however, one tries to make the primary units as alike as possible. Therefore, one would not ordinarily think of drawing a two-stage sample from a stratified population. More is said about stratification in a later section.

Sampling with replacement and unequal probabilities. It has been observed that the lion's share of the variance is caused by variation among primary units that differ in size. Sometimes it is impossible to set up the primaries in such a manner that their sizes are equal. In this case we can accomplish a reduction in variance by varying the probabilities with which the primaries are drawn. Drawing primaries with probability proportional to size, that is, $p_i = M_i/M$, where p_i is the probability of drawing the ith primary unit, accomplishes approximately the same result as equalizing the size of the primaries.

Let us investigate a method for drawing with probability proportional to size. First, we cumulate the M_i. Then we draw a random number from 1 to M. The primary unit to be included in the sample is the unit whose cumulative M_i contains the random number chosen.

Consider the example in Table 11-2. The first 5 random numbers in the range 0,001 to 1,480 are, let's say, 722, 146, 501, 974, and 53. Number 722 identifies primary unit number 7, number 146 identifies primary unit 2, 501 identifies unit 6, number 974 identifies unit 11, and 53 identifies unit 2 again.

Now the question arises, What shall we do when we select the same primary unit a second time? We should like to draw without replacement (as we do in simple random sampling). However, the formulas for the computation of the variance become so complex under this system of drawing that it is not feasible to proceed in this manner. If we want to be able to compute the variance for our estimator, we must draw *with replacement.* That is, we should include primary unit 2 twice in our sampling procedure.

However, it is known that sampling without replacement yields a smaller variance than sampling with replacement. Therefore, we may

sample without replacement and compute our variance by the with-replacement formula, realizing as we do so that our variance is being overstated. This procedure places us on the conservative side in any confidence-interval statements about the population mean or total. It yields essentially correct results if the sampling proportion n/N is small.

Table 11-2. Illustration of Drawing of Primary Units with Probability Proportional to Size of Unit

Primary unit	M_i	Cumulative M_i
1	30	30
2**	200	230
3	10	240
4	40	280
5	60	340
6*	300	640
7*	100	740
8	80	820
9	70	890
10	60	950
11*	40	990
12	10	1,000
13	20	1,020
14	60	1,080
15	400	1,480

It is time now to consider some formulas. We shall assume that primaries are drawn *with replacement* and that secondaries are drawn *without replacement* within the primaries. The population total is estimated by

$$\hat{T} = \frac{M}{n} \sum^{n} t_i \bar{Y}_i \tag{11-27}$$

where t_i is the number of times the ith primary is drawn. Note that $\Sigma t_i = n$. The variance of this estimate of the population total is

$$V(\hat{T}) = \frac{1}{n} \sum^{N} (nM_i - M_i + M)(M_i - m_i) \frac{S_i^2}{m_i} + \frac{M}{n} \sum^{N} M_i(\mu_i - \mu)^2 \tag{11-28}$$

We can estimate this variance in an unbiased manner by

$$v(\hat{T}) = \frac{M}{n} \sum^{n} t_i(M_i - m_i) \frac{s_i^2}{m_i} + \frac{1}{n(n-1)} \sum^{n} t_i(M\bar{Y}_i - \hat{T})^2 \tag{11-29}$$

Where $\hat{T}$ is defined by (11-27). The formulas appear to be rather messy, but the computations are simple, if a bit tedious. The tedium in computation is more than compensated for by the resulting reduction in variance.

When the same primary unit is drawn twice, the question arises whether to draw 2 independent samples within that primary or to take a single sample and give it a weight of 2. Actually, there is not much difference in the variance, and simplicity usually dictates a single sample. The above formulas are based on a single sample.

EXERCISE 11-6. A population consists of 8 primary units with the following characteristics:

Primary unit	M_i	S_i^2	μ_i
1	60	480	50
2	90	690	60
3	100	1,000	50
4	80	200	40
5	40	350	80
6	100	250	60
7	20	1,200	70
8	150	500	60

If we decide to draw 4 primaries with probability proportional to size and to take a 10 per cent sample from each primary, what is the true variance of an estimate of the total?

EXERCISE 11-7. With reference to Exercise 11-6, draw the 4 primaries with probability proportional to size.

EXERCISE 11-8. Suppose that the 4 primaries drawn and their sample means and variances are as follows:

Primary	m_i	$\bar{Y}_i$	s_i^2
2	9	58	400
3	10	53	800
6	10	59	500
8	15	61	650

Estimate the population total and its standard error.

11-7. STRATIFIED RANDOM SAMPLING

We have seen above that stratified sampling can be considered as a special case of two-stage sampling in which all the primary units have

been included in the sample. However, stratified sampling is so important in practice that it warrants additional treatment.

First a word about formation of strata. The criterion by which we decide into which stratum we shall put a sample unit is called the *stratifying factor.* Common stratifying factors are age, sex, economic status, educational level, geographic area, and the like. A stratifying factor is considered to be an "effective" factor if it divides the population into mutually exclusive groups, or strata, such that the sample units in each stratum are homogeneous (or nearly so) with respect to the variable under study. This generally means that the sample units in different strata are as unlike as possible. The reason that this organization of strata reduces the variance is that the strata are considered separate populations in the estimation procedure. Since a separate estimate is made for each stratum and the estimates are then added together, the only variance which enters the computation is the within-stratum variance S_i^2. This is obvious from inspection of Eq. (11-25). Since the strata are organized so that the S_i^2 are as small as possible, the stratified estimate has a small variance.

A question arises concerning how many units to draw from each stratum in order to make up the total sample. This is called the problem of allocation. A frequently used solution is *proportional allocation.* That is, more sample units are drawn from a large stratum than from a small stratum. However, it seems fairly obvious that, if a particular stratum has a large variance, one should take a larger sample from it than from another stratum having a small variance. In fact, if a stratum has 0 variance (all the units being exactly alike), a single sample unit from this stratum gives as much information as all the units put together.

The rule on *optimum allocation,* that is, allocation such that for a fixed sample size m the variance will be a minimum, can be stated as follows: Let m be the total sample size; then optimum allocation is achieved by choosing the m_i such that

$$m_i = m \frac{M_i S_i}{\Sigma M_i S_i} \tag{11-30}$$

This says, in effect, that we should allocate in proportion to the product of the stratum size and the standard deviation.

Sometimes cost becomes a factor in allocation. That is, it may cost more to obtain information about a sample unit in one stratum than in another. For example, it is more costly, because of travel expense, to interview people in rural areas than in urban areas. Our problem then is to determine an allocation so that for a fixed cost we can get the maximum amount of information. Here our problem is twofold. First, we must determine the total sample size that will be permitted by a particular

cost. Second, we must allocate this sample size to the strata. If C denotes the total cost which we allocate to the collection of data and c_i the cost per unit of obtaining a sample unit from the ith stratum, then the total sample size is given by

$$m = \frac{C\Sigma M_i S_i/\sqrt{c_i}}{\Sigma M_i S_i \sqrt{c_i}} \tag{11-31}$$

Allocation of m to the separate strata is given by

$$m_i = m \frac{M_i S_i/\sqrt{c_i}}{\Sigma M_i S_i/\sqrt{c_i}} \tag{11-32}$$

That is, allocation is directly proportional to stratum size and standard deviation and inversely proportional to the square root of cost per unit.

An illustration is given in Table 11-3. Note the change in allocation caused by introduction of the cost factor. The total sample size of 190 is determined as follows:

$$m = \frac{C\Sigma M_i S_i/\sqrt{c_i}}{\Sigma M_i S_i \sqrt{c_i}} = \frac{600(16{,}036)}{50{,}669} = 190$$

Table 11-3. Illustration of Optimum Allocation with and without Consideration of Cost

Stratum	Allocation for $m = 200$				Allocation for $C = \$600$			
	M_i	S_i	M_iS_i	m_i	c_i	$1/\sqrt{c_i}$	$M_iS_i/\sqrt{c_i}$	m_i
1	50	10	500	4	\$1.00	1.000	500	6
2	100	20	2,000	15	0.50	1.414	2,828	33
3	250	10	2,500	19	2.00	0.707	1,768	21
4	400	50	20,000	154	5.00	0.447	8,940	106
5	200	5	1,000	8	0.25	2.000	2,000	24
Total	1,000		26,000	200			16,036	190

EXERCISE 11-9. Suppose that the following costs were listed in Table 11-3:

Stratum	*Cost, c_i*
1	\$2.00
2	3.00
3	1.00
4	2.00
5	3.00

Find the optimum allocation of a sample where a total cost of \$1,000 is permitted.

EXERCISE 11-10. Suppose that the primary units in Exercise 11-6 are strata. (*a*) Allocate a sample of size 200 by proportional allocation (proportional to size) and compute the variance of the mean. (*b*) Allocate a sample of size 200 by optimum allocation and compute the variance of the mean.

11-8. THREE-STAGE SAMPLING

The multistage sampling process need not stop at two stages. We can have three, four, or more stages. Here the estimates of totals and their variances are presented for the three-stage case, but the student can see how the formulas can be extended for further stages. The estimate of the total is

$$\hat{T} = \frac{N}{n} \sum_i \frac{M_i}{m_i} \sum_j G_{ij} \bar{Y}_{ij} \tag{11-33}$$

The variance of the estimated total is

$$\begin{aligned} V(\hat{T}) = {} & \frac{N}{n} \sum_i \frac{M_i}{m_i} \sum_j \frac{G_{ij}}{g_{ij}} (G_{ij} - g_{ij}) S_{ij}^2 \\ & + \frac{N}{n} \sum_i \frac{M_i}{m_i} (M_i - m_i) \sum_j \frac{(T_{ij} - T_i^*)^2}{M_i - 1} \\ & + \frac{N}{n} (N - n) \sum_i \frac{(T_i - T^*)^2}{N - 1} \end{aligned} \tag{11-34}$$

where G_{ij} = number of tertiary units within the ijth secondary
g_{ij} = number of tertiary units included in the sample

$$T_i^* = \frac{1}{M_i} \sum_j T_{ij}$$

$$T^* = \frac{1}{N} \sum_i T_i$$

The first term in (11-34) is the within-secondaries component, the second term is the within-primaries component, and the third term is the between-primaries component. Generally, the first term is so small compared with the others that it may be ignored.

An unbiased estimate of the variance is given by

$$v(\hat{T}) = \frac{N}{n}\sum_i \frac{M_i}{m_i}\sum_j \frac{G_{ij}}{g_{ij}}(G_{ij} - g_{ij})s_{ij}^2 + \frac{N}{n}\sum_i \frac{M_i}{m_i}(M_i - m_i)\sum_j \frac{(G_{ij}\bar{Y}_{ij} - Y_i^*)^2}{m_i - 1} + \frac{N}{n}(N - n)\sum_i \frac{(M_i\bar{Y}_i - Y^*)^2}{n - 1} \quad (11\text{-}35)$$

where $$Y_i^* = \sum_j \frac{G_{ij}\bar{Y}_{ij}}{m_i}$$

11-9. TESTS OF HYPOTHESES IN SAMPLING FROM FINITE POPULATIONS

We have seen how we can construct estimators for means and totals in the common finite sampling situations. We have seen also how we can estimate the variance of the means or totals. Nothing has been said about the *distribution* of the means or totals, however. In fact, little is known about their distribution. It will be recalled (Sec. 4-6) that, if Y is a normally distributed random variable, then $\bar{Y}$ is normally distributed. However, if the population of Y's is finite, then the variable Y cannot be normally distributed, and since there can be only a finite number of $\bar{Y}$'s, this variable cannot be normally distributed either.

It has been demonstrated, however, that in repeated sampling from finite populations the sample means tend toward normality when certain conditions are imposed on sample size. Perhaps it would be more accurate to say that these empirical distributions are not violently nonnormal. This empirical evidence is employed to justify the use of the t distribution in testing hypotheses about, or in placing confidence intervals around, means or totals. In using the t distribution it should be remembered that the probabilities are approximate rather than exact.

12. STATISTICAL QUALITY CONTROL

The question, what is the quality of our product? represents an ever-present problem of management. One way of finding the answer might be to take a sample and find out. Assuming the sample to be drawn properly, one might be able to evaluate the quality of his product in this manner, but the results would be out of date before they could be placed on the plant superintendent's desk. The control of quality is a continuing process, and statistical methods must take account of the dynamic nature of the production process.

From management's viewpoint it is unsatisfactory to rely entirely on a check of final assemblies to determine quality. It is much more meaningful to control the quality of subassemblies and of the parts making up those subassemblies, so that the check on the final assembly is a check on the assembly operation itself rather than on a combination of parts. This check on intermediate stages of production can be carried all the way back to a check on the quality of incoming materials. The principle involved might be called the principle of assignment of cause. This principle asserts, in effect, that one checks quality at the points where variations from established standards will direct attention directly to the cause of the imperfections.

The more complex the mechanism being produced, the more checking of quality at intermediate points is necessary. To illustrate, suppose we are producing a complex mechanism, consisting of 1,000 parts. Suppose also that the probability that each part is defective is 0.001. Then the probability that the mechanism itself will be defective is far greater than 0.001. It is, in fact, 1 minus the probability that there will be no defective parts. From our knowledge of binomial probabilities we can compute this probability to be

$$1 - (0.999)^{1,000} = 0.63 \qquad \text{(approximately)}$$

Therefore, there are about 2 chances out of 3 that, unless subassemblies are checked, the mechanism will not work when finally assembled.

12-1. INSPECTION

Since any continuous check on quality involves drawing a sample of the product at various stages in the production process, we are faced with all the problems relating to sampling. The number of samples drawn depends upon such factors as (1) variability of the production, (2) cost of inspection, (3) size of the lot sampled, and (4) resources available for inspection. There is usually a problem associated with defining the lot to be sampled. The more homogeneous this lot, the more meaningful is a sample drawn from it. Therefore, the lot may be chosen as (1) one hour's production from a machine, (2) one operator's production for a day, or (3) one day's production under the supervision of a single foreman. In other words, there is something common to each lot sampled. The thing that is in common is dependent upon what one wishes to check. In (1) the machine is being checked, in (2) the operator is being checked, and in (3) the foreman is being checked.

Some attention also must be given to the matter of selecting a random sample. Experiments have shown that one cannot in general obtain a random sample by relying upon an inspector to choose a sample arbitrarily. Sometimes a table of random numbers is helpful, but this is not true if the lots are large. At other times a systematic sample can be employed—say, every twentieth item on a conveyor belt. In such cases one must make sure that the operator does not know which items will be checked.

There is no problem in obtaining a sample of water from a tank when the contents are constantly stirred, but when the product to be sampled is divided into individual pieces, each with an identity of its own, all sorts of complications can arise. Since it is so costly and time-consuming to draw a completely random sample, the criterion of "reasonable randomness" is often used. To illustrate, if items are placed on a conveyor belt in an unorganized, or "scrambled," manner, then a systematic sample of those items might be considered reasonably random. Again, if bolts are sacked by machinery, there is probably little error introduced by taking a sample bolt out of the top of the sack. However, one could be seriously in error if he sampled apples by taking one from the top layer in a basket!

We shall assume that we have solved satisfactorily the problem of defining a "lot" of production and that we can select 5 units from each lot in a "reasonably random" manner. We then are faced with deciding what characteristic of the product to measure. If we are producing

twine, it may be that the only variable of interest is the tensile strength. If we are producing brass bushings, we may be interested in the interior diameter, the exterior diameter, the length, the thickness of the metal, or perhaps the eccentricity, that is, the difference between maximum and minimum diameters. If one of these characteristics is of utmost importance, we may limit ourselves to recording that measurement. In other cases we may record all measurements, in which case we employ the techniques of "multivariate" quality control. Another technique is simply to examine each sampled bushing and to classify it as either "accepted" or "rejected" on the basis of our specifications. For example, to check interior diameters, a GO–NOT GO gauge may be used. The diameter must be large enough to permit the entrance of the GO end of the gauge but not large enough to permit the entrance of the NOT GO end.

If this sort of inspection is resorted to, it is important that the reason for rejection be recorded. Otherwise it is difficult for management to take corrective action.

12-2. CONTROL CHARTS FOR MEASURED VARIABLES

A convenient device for maintaining a continuing control at critical stages of production is the control chart. Such charts may be used to control the average or variance of some measurable characteristic of the product or simply to control the fraction of each lot that is defective.

Suppose we are producing wire and are concerned with its diameter. We take 5 samples from each lot and compute the mean and variance of the 5 diameters. We must compare the means and variances against some standard if they are to be useful to us.

Standards are sometimes specified by an outside agency such as the American Society for Testing Materials or by a government purchasing agency such as the Army or Navy. At other times the standard is developed internally by measuring and recording samples of the product. In any case, some standard is assumed to be necessary.

Suppose that the standard for comparison is to be determined from the production process itself. We could take samples of 5 from many successive lots and develop standards from an average of the sample means and variances. We should be fairly certain, however, that the lots making up the standard are all acceptable lots. It might be easier to work with fewer lots and to select more than 5 samples from each. The lots selected should have been produced under as nearly "normal" conditions as possible. Exactly what constitutes normal conditions is purely a matter of judgment, but certainly management would avoid periods

during which new employees were being trained, or there were power failures, or new sources of materials were being tried.

Returning to our wire illustration, suppose we select 10 lots and take 50 samples from each lot, obtaining the data of Table 12-1. The samples have an average mean of 62.0 and an average variance of 23.7. We accept these tentatively as our standards. Note that, if the samples had not all been of equal size, we should have weighted the means $\bar{Y}_i$ by sample size and the variances s_i^2 by degrees of freedom in the averaging process.

With a process average of 62.0 and a process variance of 23.7, we can certainly place confidence limits on the variation to be expected in

Table 12-1. Samples of Wire Diameters, in Thousandths of an Inch, for Determining Quality-control Standards

Lot number	Number of samples	Average diameter, $\bar{Y}_i$	Lot variance, s_i^2
1	50	62.0	15.6
2	50	63.1	18.0
3	50	61.7	17.6
4	50	62.8	38.3
5	50	62.0	24.2
6	50	61.5	28.5
7	50	63.1	21.4
8	50	60.6	33.1
9	50	62.2	20.5
10	50	61.3	19.4
Average		62 0	23.7

samples of size 5. However, since the process mean and variance are themselves estimated, it is customary in some applications to place three standard-error limits, rather than probability limits, on the mean. The standard error of the mean for a sample of size 5 is

$$s_{\bar{Y}} = \sqrt{\frac{23.7}{5}} = 2.18$$

If we want three standard-error limits, we compute

$$62.0 \pm 3(2.18) = 62.0 \pm 6.54$$

so that the limits are 68.54 and 55.46. Any sample mean which falls within these limits will cause us little concern, but if a sample mean falls outside these limits, we say that the process is out of control and take immediate steps to remedy the situation.

Also, we may be concerned with the variation in diameters, particularly if the wire is to be used in a tying operation, where extra-thick wire would jam up the mechanism and very thin wire would break. If we accept 23.7 as the population variance, it is a simple matter to place an upper limit (with some fixed probability) on the variance. We are not concerned with a lower limit because, presumably, the more uniform the wire, the happier we are.*

Since
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

we can write
$$s^2 = \frac{\chi_\alpha^2 \sigma^2}{n-1} \tag{12-1}$$

where χ_α^2 is the χ^2 value at the α level of probability with $n-1$ degrees of freedom. If we let $\alpha = 0.01$, $n = 5$, and $\sigma^2 = 23.7$, we have

$$s^2 = \frac{13.3(23.7)}{4} = 78.8$$

That is, we shall not examine our production process so long as the sample variances are less than 78.8.

It is convenient to prepare a chart of the means and another of the variances to serve as a running record of the quality of the particular item. These charts show the limits beyond which one does not expect the mean (or variance) to go as long as the process is in control. For this reason they are called control charts. The control chart for means and the control chart for variances are shown in Fig. 12-1 for our wire illustration. The data, accumulated by inspection of lots of production, are shown in Table 12-2.

Both mean and variance remain in control until the seventeenth lot, when the mean goes out of control by exceeding the upper control limit. The suspicion that something is wrong with the production process is verified by the fact that the previous means (of lots 12 to 16) were all above the supposed process average. The probability of getting 6 means in a row, all above the process average, is $(\frac{1}{2})^6 = \frac{1}{64}$, if we can assume that the mean is normally distributed.

When the mean went out of control, the process was examined to discover the source of the faulty product. After adjustment of the difficulty, the mean returned within the control limits. The variance went out of control with lot 23. At this point the process was presumably examined again and the difficulty found and corrected.

* This is not universally true, however. If one is producing lightweight aggregate for concrete, he may wish to have small aggregates, as well as large, in the mix. Here a small variance might be as bad as a large one, or worse.

We note that considerable computation is required each time a sample of 5 units is drawn. This computation makes on-the-spot plotting of the points on the control chart difficult. Since the effectiveness of the control chart is enhanced if corrective action can be taken immediately, we may wonder whether there is some way of avoiding the computation. Actually there is. We can use the median in place of the mean and the range in place of the variance, so that both figures can be obtained by glancing at the 5 sample observations.

By adopting this process, however, we lose a certain amount of efficiency. That is, departures from the "expected" or "process average"

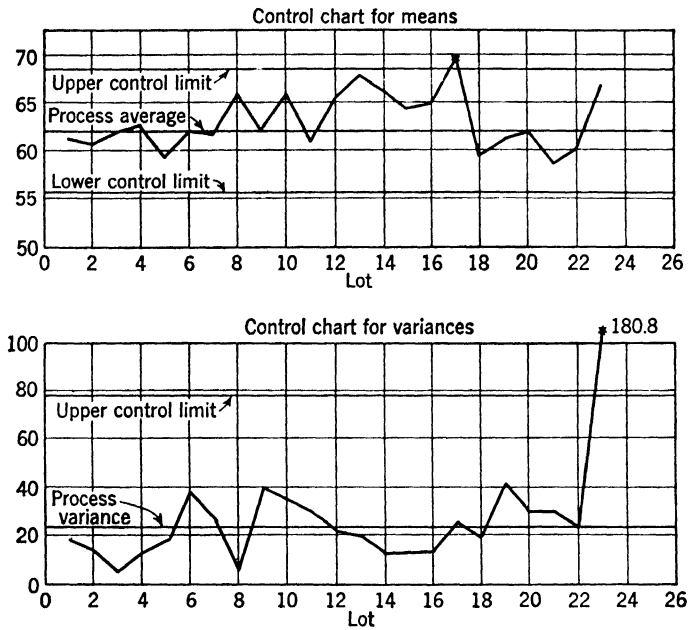

FIG. 12-1. Control charts for means and variances of wire diameters

must be greater to reveal a real change in quality when we use the median and the range in place of the mean and variance. Nevertheless, their ease in computation may more than offset this disadvantage.

The standard error of the median is approximately 1.2533 times the standard error of the mean. That is, the standard error of the median in our illustration is $1.253(2.18) = 2.73$. Assuming normality of the observations, the average value of the median should be the same as the average value of the mean. Therefore there should be no change in the process average. Upper and lower control limits are therefore

$$62.0 \pm 3(2.73) = 62.0 \pm 8.19$$

Table 12-2. Inspection Record of Wire Samples with Means and Variances

Lot	Diameter, in thousandths of an inch					Mean, $\bar{Y}_i$	Variance, s_i^2
	I	II	III	IV	V		
1	58	67	63	61	57	61.2	16.2
2	64	61	64	60	55	60 8	13 7
3	60	65	61	60	63	61.8	4.7
4	65	64	63	56	63	62.2	12.7
5	57	66	58	60	55	59.2	17.7
6	68	59	59	55	69	62 0	38.0
7	66	58	62	55	67	61.6	26.3
8	69	63	65	67	66	66.0	5.0
9	68	53	68	61	60	62.0	39.5
10	73	68	57	65	67	66.0	34.0
11	65	57	54	67	61	60.8	29.2
12	70	64	60	62	70	65.2	21.2
13	65	64	74	65	71	67.8	19.7
14	65	66	63	72	64	66.0	12.5
15	65	60	61	68	67	64.2	12.7
16	64	63	71	63	63	64.8	12.2
17	70	63	67	76	72	69.6	24.3
18	56	66	60	60	55	59.4	18.8
19	57	57	71	56	64	61.0	41.5
20	68	62	55	58	66	61.8	29 2
21	56	53	67	56	60	58.4	29.3
22	65	59	64	53	59	60.0	23.0
23	78	60	55	84	56	66.6	180.8

The average range is a function of both the population standard deviation and the size of the sample. The average range for a sample of size 5 is substantially smaller than the average range for a sample of size 50.* It is assumed that the variable (wire diameter) is normally distributed.

For our illustration with $n = 5$, the upper 1 percentage point is given by 4.60 times the standard deviation. Since our process standard deviation is 4.87, the upper control limit on the range is

$$4.60(4.87) = 22.4$$

* Factors for converting standard deviations to ranges for various-sized samples and various probability levels are given by E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1, table 22, Cambridge University Press, New York, 1956.

If we read the medians and ranges from Table 12-2, we obtain the data of Table 12-3. The control charts for the median and range are given in Fig. 12-2. It may be seen that these charts show essentially the same picture as Fig. 12-1. The control limits on the median are a little wider, however, so that the median of lot 17 falls just short of the upper control limit. Its value is 70, and the upper control limit is 70.19. This is the lot whose mean went beyond the control limit.

Table 12-3. Medians and Ranges Obtained from the Data of Table 12-2

Lot	Median	Range	Lot	Median	Range
1	61	10	13	65	10
2	61	9	14	65	9
3	61	5	15	65	6
4	63	9	16	63	8
5	58	11	17	70	13
6	59	14	18	60	11
7	62	12	19	57	15
8	66	6	20	62	13
9	61	15	21	56	14
10	67	16	22	59	12
11	61	13	23	60	29
12	64	10			

EXERCISE 12-1. Suppose we observe the following samples of wire from lots 24 to 30 (a continuation of Table 12-2):

Lot	Diameter, in thousandths of an inch				
	I	II	III	IV	V
24	71	64	55	61	70
25	62	67	64	75	72
26	77	53	56	66	54
27	60	87	79	50	74
28	65	66	59	55	66
29	69	53	57	67	70
30	70	64	63	68	63

(*a*) Which of the above lots are out of control on the charts of means and variances? (*b*) Which are out of control on the charts of medians and ranges?

EXERCISE 12-2. (*a*) Set two standard-deviation limits on the chart of means and an upper 95 per cent limit on the chart for variances, for the data of Table 12-2. (*b*) What are the advantages and disadvantages of these limits?

EXERCISE 12-3. (*a*) Under the three-sigma sampling plan of Fig. 12-1, what is the probability that an increase of 5 units in process average will not be detected on the first sample following the change? (*b*) What is the probability that it will not be detected on the next 5 samples? Use the multiplication law for probabilities.

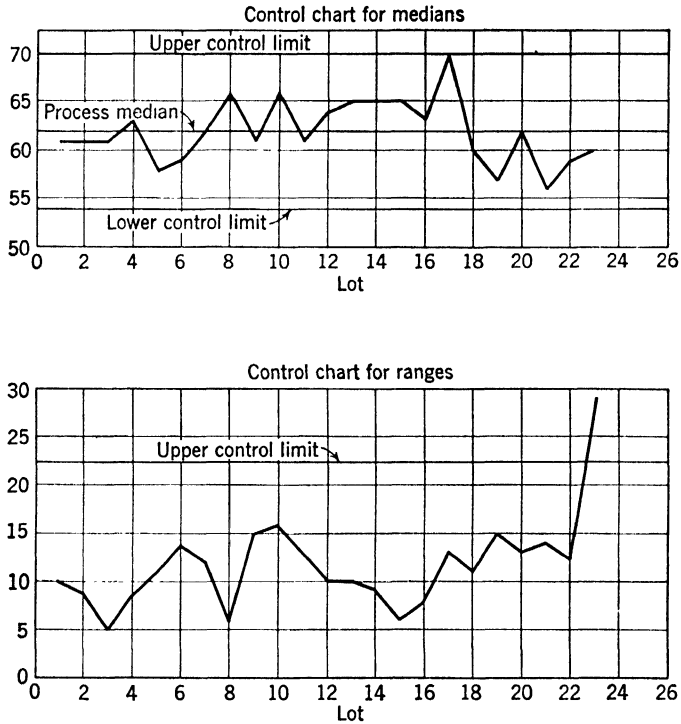

FIG. 12-2. Control charts for medians and ranges of wire diameters

12-3. CONTROL CHARTS FOR FRACTION DEFECTIVE

As was pointed out earlier, inspection may consist simply of judging whether a unit produced is to be accepted or rejected. In this case, if we assume that a proportion p of the population is defective, then, in samples of n from this population, the probability that y of these will be defective is precisely

$$ {}_nC_y p^y q^{n-y} \tag{12-2} $$

by the familiar binomial law. Here we are assuming that the population is infinite (or at least very large), so that the probability of a defective unit remains constant from one draw to the next. This is the typical quality-control situation.

For example, if $p = 0.2$ (determined by extensive sampling during

"normal" production periods) and if 10 units at a time are examined from large lots (so that the finite-population correction can be ignored), then the probability that there are y defective units out of the 10 examined is

$$_{10}C_y(0.2)^y(0.8)^{10-y} = \frac{10!}{y!(10-y)!}(0.2)^y(0.8)^{10-y}$$

Letting $y = 0, 1, 2, \ldots, 10$, successively, we obtain the probabilities of Table 12-4.

Suppose we decide to set our control limits between 5 and 6. That is, if we find 5 or fewer defectives in a sample of 10, we shall assume that our level of quality (20 per cent defective) has not changed. If we find 6 or more defectives in a sample of 10, we shall assume a change in quality and take corrective action. Table 12-4 shows that the probability of

Table 12-4. Probabilities of Obtaining y Defectives in Lots of 10, with $p = 0.2$

y	$P(y)$	Cumulative probability
0	0.1074	1.0000
1	0.2684	0.8926
2	0.3020	0.6242
3	0.2013	0.3222
4	0.0881	0.1209
5	0.0264	0.0328
6	0.0055	0.0064
7	0.0008	0.0009
8	0.0001	0.0001
9	0.0000	0.0000
10	0.0000	0.0000

obtaining 6 or more defectives is 0.0064, or less than 1 in 100. This represents the risk that we shall examine the production process unnecessarily. If we set the control limit between 4 and 5, we run a risk of 0.0328 of examining the process needlessly. More is said about such risks in a later section.

If inspection is a simple process, we are likely to examine a sample larger than 10, thereby obtaining a greater assurance of quality. Tables of the binomial probability distribution for various values of p and n are available to simplify the job of computation.* If n is quite large, one may obtain adequate results by employing the normal approximation

* For example, *Tables of the Cumulative Binomial Probability Distribution*, Harvard University Press, Cambridge, Mass., 1955.

to the binomial. If n is large enough for the process average p (see Sec. 6-2), the variable y (number of defectives) is approximately normally distributed with mean np and standard error $\sqrt{npq}$. Usually in quality control the value of p is small enough so that it is not practical to increase n to such a size that the normal approximation is appropriate.

Instead of recording the number of defective items, one may record the number of defects in a sample of n units. It is possible under this system for a defective unit to have several defects. If the sample size is adequate, the variable representing number of defects will approach the normal form, and the techniques of the previous section will apply.

Sometimes the probability of finding a defective unit is quite small, so that in a sample of size n there is a substantial probability that no defects will be found. In these cases one may employ the Poisson distribution, rather than the binomial or the normal, in placing confidence limits. If m is the process average number of defects in a sample, then the probability of y defects in a sample is

$$P(y) = \frac{m^y e^{-m}}{y!} \tag{12-3}$$

As was pointed out in Chap. 2, tables of this function are available to simplify the placing of control limits. A limited table is given as Appendix Table V.

EXERCISE 12-4. Suppose it is established that about 15 per cent of a product is defective. This standard is checked by examining samples of size 400 from lots of 10,000. Find an upper control limit such that the production process will be examined needlessly only 1 per cent of the time.

EXERCISE 12-5. A sampling plan calls for recording the number of defects (not defectives) in a sample of 10 assemblies. The number of defects per sample has averaged 4 over a substantial period of time. Find an upper control limit such that the production process will be examined needlessly approximately 5 per cent of the time.

12-4. ESTABLISHMENT AND MODIFICATION OF STANDARDS

The problem of establishing standards has been oversimplified in the previous two sections. In the first place, there may be a difference between the statistical standards and the engineering standards. Examination of lots of production will reveal the standards which are currently in existence, but they may not conform to engineering tolerances or advertised standards. In this case management may be faced with some tough problems. For example, what is the satisfaction of knowing that a process for turning shafts is "in control" statistically if 25 per

cent of the shafts produced are outside the engineering tolerances on diameters?

Management must consider these questions: Can the production process be improved so that the control limits lie within the tolerance limits? Are the tolerance limits realistic? The first question leads to an examination of the production process, the second to an evaluation of the product. It may be that the solution lies in a compromise.

If one can assume that product specifications (including tolerances) have been set realistically, the supervisor of quality control will be happy if the control limits are inside the tolerance limits, that is, the limits beyond which the product is unsatisfactory. If tolerance limits are much wider than control limits, however, management may begin to wonder whether it can step up production—by increasing machine speed, by substituting materials, or by some other device—and still keep the product within the tolerances.

We see then that statistical standards are meaningless unless they are within prescribed tolerances and that engineering tolerances cannot be set without regard to the capabilities of the production process. Therefore, the relationship between the two concepts is subject to review constantly.

Even if there are neither established engineering tolerances nor precise product specifications, one may still want to review control limits periodically. A previously uncontrolled process will typically show substantial improvement during the initial phase of control. This has nothing to do with statistics. It simply reflects the fact that employees at all levels are made aware that production is being checked. Frequently, then, it is desirable to set new control limits based on the "normal" for the controlled process. Clearly this revision cannot be carried on indefinitely, because of the almost inevitable conflict between quality and output. Adherence to unrealistic levels of quality may hamper output seriously.

In Sec. 12-2 the original standards were set by careful scrutiny of a few lots of product. They need not be set in this manner. One may simply start examining 5 units per lot and recording the data. After 30 or more lots are examined, he may then establish control limits on the basis of this experience. Ordinarily it is not safe to establish standards on the basis of fewer than 30 lots.

12-5. NATURE OF THE RISKS IN QUALITY CONTROL

If we denote the process average by μ_0, we may consider the control chart on means to be a continuing test of the hypothesis $\mu = \mu_0$ against the alternatives $\mu > \mu_0$ and $\mu < \mu_0$. Suppose that our measured process

average is, in fact, μ_0. Also, suppose that the sample means are normally distributed and that our estimate of the standard error of the mean $\sigma_{\bar{x}}$ is precise. What is the effect of placing three standard-error limits on the means? Reference to the normal-area table (Table I, Appendix) shows that 99.73 per cent of the area under the normal curve is contained within three standard-error limits. Therefore, the probability that a sample mean will lie outside these limits is 0.0027, or 23 chances out of 10,000. This is the risk that management incurs of examining the production process unnecessarily. This is α, in the language of Chap. 3.

If the examination of the production process is costly, then management will want α to be quite small. On the other hand, if the process can be examined without stopping the process or otherwise incurring undue expense, it may be desirable to set α at a higher figure, say, approximately 0.05 (that is, setting *two* standard-error control limits).

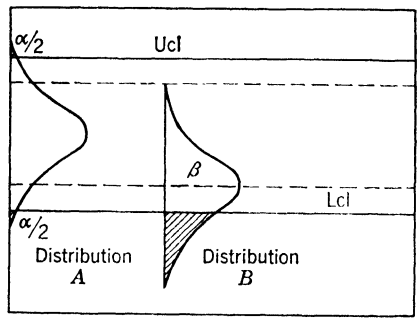

FIG. 12-3. Type II error (β) after a drop in process average

Now, let us consider the other type of risk which we have previously related to Type II error, β (see Chap. 3). This is the risk that we shall *not* examine the production process when a real change in quality has occurred. According to the argument given in Chap. 5, one cannot decrease α without increasing β. Let us examine the argument again, with particular attention to the control chart on means. Two normal distributions are shown in Fig. 12-3. The first distribution, A, is the distribution of sample means at the time the control limits were set. Distribution B is the distribution of sample means after there has been a decline in process average. The shaded portion of each curve represents the probability that a sample mean will lie outside the control limits and thus cause a rejection of the hypothesis $\mu = \mu_0$ and an examination of the production process. The unshaded portions represent the probability that the hypothesis $\mu = \mu_0$ will be accepted, with a consequent failure to examine production processes.

If we are disturbed by the large unshaded area under distribution B (that is, the probability of accepting a false hypothesis), we may wonder

whether Type II error can be reduced. Obviously it can. All we need to do is move the upper and lower control limits closer together—say, to the positions indicated by the dotted lines. If we do this, we shall reduce the β error in distribution B to about $\frac{1}{2}$, but notice what damage we do to the α error of distribution A. We increase seriously the probability that the production process will be examined unnecessarily. There must be an economic balancing of these two types of error.

Sometimes two sets of control limits are used. *Inner control limits* are used as a sort of warning sign. When a sample value lies outside the inner control limits but *inside the outer control limits*, a superficial examination of the production process is made, but the process is not halted. If a sample value lies outside the *outer control limits*, the process is halted until the cause of the deficiency is located and corrected.

So far, our discussion of risk has referred exclusively to the point of view of the producer of the goods. In some situations the consumer may actually dictate inspection procedures. This is particularly true if the principal consumer is the United States government.

To introduce the concept of producer's and consumer's risks, let us suppose that the producer strives for a product which is only 5 per cent defective. In general, we shall denote the *producer's standard* by p_1. It represents an upper limit on the proportion of defectives as far as the producer is concerned. Since this is an announced or advertised standard, we must assume that the consumer would not purchase the goods if he were dissatisfied with this standard. He may, in fact, be willing to accept goods of a lower standard, that is, goods with a higher fraction of defective items, say, 10 per cent. We let p_2 denote this *consumer's standard,* which is the maximum fraction defective which he is willing to accept.

Now, let us assume that an inspection plan is instituted for the acceptance of the product by the consumer. Decisions are made concerning (1) the size of a standard lot, (2) the number to be inspected (sampled) in each lot, and (3) the number of defectives in the sample required to cause rejection of the lot (with consequent return of the goods to the producer). Clearly there is a risk to the producer that a lot which is better than his standard ($p_1 = 0.05$) will be returned to him. This risk we can denote by r_p. Similarly, the consumer faces a risk r_c of accepting a lot which is worse than his standard ($p_2 = 0.10$). Naturally, the consumer will want to adopt an inspection plan to make r_c small, and the producer will want to make r_p small. The only way that both objectives can be accomplished is to increase sample size. But an unrealistic increase in sample size will increase costs, and the increased cost must ultimately be passed on to the consumer. A compromise is indicated.

Let us consider a particular sampling plan and see what the producer's

and consumer's risks are. We shall assume that p_1 (producer's standard) is 0.05 and that p_2 (consumer's standard) is 0.10, as above. We assume further that the lots inspected are large enough that the finite-population correction [see Eq. (4-17)] may be ignored. If we sample 10 items out of each lot and reject the lot when we find one or more defectives, what are the characteristics of the sampling plan?

Suppose, first of all, that production is actually consistent with the producer's standard; that is, 5 per cent of the production is defective. What is the producer's risk, that is, what are the chances that a lot of standard quality will be returned? Clearly it is the probability of getting 1 or more defectives out of 10 trials where the probability is 0.05 of getting a defective in a single trial. That is,

$$r_p = \sum_{y=1}^{10} {}_{10}C_y(0.05)^y(0.95)^{10-y} = 1 - {}_{10}C_0(0.05)^0(0.95)^{10}$$

$$= 1 - 0.60 = 0.40$$

That is, the producer will have 4 lots out of 10 returned even though his process is operating at the advertised standard. This is not very satisfactory from his viewpoint.

What is the consumer's risk? Clearly it is 0 under the conditions stated, because the process fraction defective is, in fact, 0.05. Since the lots are assumed to be large, he will not get any lots that are worse than 0.10, which is his standard. We have, then, $r_c = 0$.

Now, suppose the product quality deteriorates until 20 per cent of the items are defective. It is clear that $r_p = 0$, because the product is all below the standard set by the producer and the probability is 0 that products above his standard will be returned to him. What about the consumer's risk? It is the probability that no defective items will show up in a sample of 10. That is,

$${}_{10}C_0(0.2)^0(0.8)^{10} = 0.11$$

From the above discussion it is clear that the computation of producer's and consumer's risks is straightforward *if the true process fraction defective is known.* Unfortunately this is the very thing that one is forced to estimate. The probability that a given lot is good or bad is unknown, and must remain unknown even after sampling. How then can we make use of producer's and consumer's risks in establishing sampling plans? The answer is by using "maximum" producer's risk and "maximum" consumer's risk.

It will be helpful to employ the concept of an *operating-characteristic curve* (OC curve). If we let p equal the true proportion defective in the lot being inspected, then the OC curve shows the probability of accepting

Table 12-5. Data for OC Curve for Samples of Size 10 under the Rule of Rejecting if Any Are Defective

p	*Probability of acceptance*	p	*Probability of acceptance*
0.01	0.904	0.11	0.312
0.02	0.817	0.12	0.278
0.03	0.737	0.13	0.248
0.04	0.665	0.14	0 221
0.05	0.599	0.15	0 197
0.06	0.539	0 16	0.175
0.07	0.484	0.17	0.155
0.08	0 434	0.18	0.137
0.09	0.389	0.19	0.122
0.10	0.349	0.20	0.107

a lot for varying values of p. Under our rule of accepting the lot if all sampled items are "good," the probability of acceptance can be computed easily as the probability of finding a good item, $1 - p$, raised to the tenth power. This probability for various values of p is given in Table 12-5. These figures, when plotted on Fig. 12-4, give an approximation to a

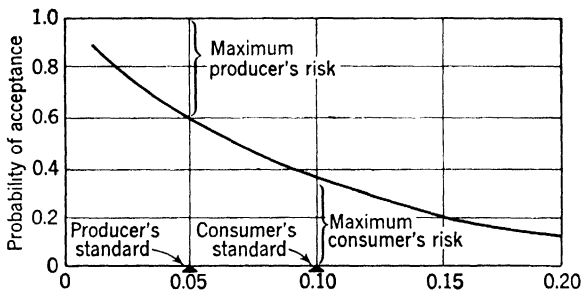

FIG. 12-4. Producer's and consumer's risks for the OC curve of sample size 10 and the decision rule to reject if any are defective

portion of the OC curve. The *maximum producer's risk*, that is, the risk that a lot of quality $p = 0.05$ or better will be returned, is 1 minus the probability of acceptance for $p = 0.05$. This risk is 0.401. The *maximum consumer's risk*, that is, the probability that a lot of quality $p = 0.10$ or worse will be accepted, is the probability of acceptance at $p = 0.10$. This risk is 0.349. These risks are shown on the chart.

We could reduce the maximum producer's risk by adopting the rule to reject if 2 or more items are defective. The new rule would inflate the maximum consumer's risk, however. The comparison of risks is as follows:

Rule	Maximum producer's risk	Maximum consumer's risk
Reject if one or more defective	0.401	0.349
Reject if two or more defective	0.086	0.736

This set of risks is completely unsatisfactory from the consumer's standpoint.

To reduce both risks, the sample size must be increased. Ideally, the producer's and consumer's risks would both be 0. Unfortunately, however, this goal is almost impossible to attain in practice. Sometimes the producer's and consumer's standards are identical, in which case the maximum consumer's risk is equal to 1 minus the maximum producer's risk. Here the ideal OC curve would show the rejection of every lot for which p is greater than the standard and the acceptance of every lot for which p is less than the standard.

By using a table of binomial probabilities, one can easily work out the OC curves for all the simple sampling plans. Such curves have been constructed for most of the common sampling plans.

EXERCISE 12-6. If $p_1 = 0.10$, $p_2 = 0.20$, and $n = 8$, work out the maximum consumer's risk for the following rules. (*a*) Reject if 1 or more are defective. (*b*) Reject if 2 or more are defective. (*c*) Reject if 3 or more are defective.

12-6. DOUBLE, MULTIPLE, AND SEQUENTIAL SAMPLING

Sometimes a lot can be judged good or bad on the basis of less than the full sample size. For example, suppose we have decided that a sample should contain 50 items and that if more than 3 items are defective the lot should be rejected. We might find 4 defective items in the first 10 examined, in which case it would be pointless to examine the remaining 40.

This philosophy has led to *double sampling*, a technique by which a small sample is examined first and then a decision is made to accept, reject, or take a larger sample. If a decision to accept or reject cannot be made after checking the first sample, it must be made after examining the second. For example, a first sample may consist of 25 items. If there are 3 or fewer defectives, the lot may be accepted. If there are 7 or more, the lot may be rejected. If there are 4, 5, or 6 defectives, an additional sample of 50 may be drawn. If out of the combined sample of 75 there are 6 or fewer defectives, the lot may be accepted. If there are 7 or more, the lot may be rejected. The possibility of deciding on the basis of the small first sample reduces the amount of inspection necessary and therefore reduces cost.

If double sampling is good, perhaps triple sampling is better, and

quadruple sampling better yet. This trend of thought leads to *multiple sampling*. An example of multiple sampling is given in Table 12-6. Note that a decision *must* be made on the last sample. The logical extension of multiple sampling is *sequential sampling*, a procedure by which the cumulative number of defects is compared against an "accept" and a "reject" standard after each sample drawing. A table devised for this plan would show the "accept" and "reject" limits for cumulative sample sizes of 1, 2, 3, 4, and so on. Sequential sampling is suited to types of inspection which are very costly—for example, those which result in the destruction of the product. Such sampling plans, when properly designed, assure one of reaching a decision based on a minimum of inspection. If inspection is cheap, the additional cost of comparing the cumulative sample results against a standard after every unit inspection may outweigh the cost of inspecting larger samples.

Table 12-6. Example of Multiple-sampling Procedure

Sample	Sample size	Cumulative sample size	Accept if defectives equal to or less than	Reject if defectives equal to or more than
1	40	40	1	6
2	40	80	4	8
3	40	120	7	11
4	40	160	10	14
5	40	200	13	17
6	40	240	16	21
7	40	280	20	21

12-7. SAMPLING TABLES

We have seen how single-sampling plans can be constructed. All we need is a table of binomial probabilities for various sample sizes and a little common sense. Double-, multiple-, and sequential-sampling plans are somewhat more complex to construct. However, sampling plans have been constructed to meet the needs of most inspection functions. An excellent catalogue of single-, double-, and sequential-sampling plans, with matching OC curves, is given in *Sampling Inspection*, by the Statistical Research Group, Columbia University (McGraw-Hill Book Company, Inc., New York, 1948). Another widely used set of tables is the Dodge and Romig Tables.* The importance of military contracts to industry has led to the widespread use of another set of sampling

* H. F. Dodge and H. G. Romig, *Sampling Inspection Tables: Single and Double Sampling*, John Wiley & Sons, Inc., New York, 1944.

tables, commonly designated MIL-STD-105A.* The Military Standard tables include tables for single-, double-, and multiple-sampling plans for various "acceptable quality levels" (the consumer's standard) and for both "normal" and "tightened" inspection. Normal inspection is instituted when the process average is within the limits prescribed by the acceptable quality level. Tightened inspection goes into effect when the process average lies outside these limits. The conditions under which normal inspection can be reinstituted are given also. In commercial practice another category of inspection, called "reduced inspection," may be introduced when quality has been particularly good for an adequate time. In any case, decisions concerning the kind of inspection to be employed are based on the process average attained in a prescribed prior period.

Since MIL-STD-105A is readily available, it is somewhat pointless to reproduce the tables here. However, a portion of a double-sampling

Table 12-7. Portion of Double-sampling Table

Sample-size code	Sample	Sample size	Cum. sample size	Acceptable quality levels (normal inspection)									
				1.0		1.5		2.5		4.0		6.5	
				Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re
I	First	35	35	0	3	1	3	1	5	2	7	3	12
	Second	70	105	2	3	2	3	4	5	6	7	11	12
J	First	50	50	1	4	1	6	2	7	3	10	5	15
	Second	100	150	3	4	5	6	6	7	9	10	14	15
K	First	75	75	1	6	2	8	4	9	5	12	7	20
	Second	150	225	5	6	7	8	8	9	11	12	19	20
L	First	100	100	2	6	3	8	5	12	7	17	10	31
	Second	200	300	5	6	7	8	11	12	16	17	30	31
M	First	150	150	3	8	5	14	7	19	11	29	15	47
	Second	300	450	7	8	13	14	18	19	28	29	46	47
N	First	200	200	4	10	6	17	9	25	12	36	18	67
	Second	400	600	9	10	16	17	24	25	35	36	66	67
O	First	300	300	6	17	8	26	12	36	18	55	26	88
	Second	600	900	16	17	25	26	35	36	54	55	87	88
P	First	500	500	9	25	12	37	18	65	27	89	43	131
	Second	1,000	1,500	24	25	36	37	64	65	88	89	130	131
Q	First	1,000	1,000	15	47	26	65	34	113	50	160	79	243
	Second	2,000	3,000	46	47	64	65	112	113	159	160	242	243
				1.5		2.5		4.0		6.5		10.0	
				Acceptable quality levels (tightened inspection)									

SOURCE: MIL-STD-105A.

* *Military Standard* 105*A: Sampling Procedures and Tables for Inspection by Attributes*, Washington, D.C., Government Printing Office, 1950. These tables are available for general distribution.

Table 12-8. Sample-size Code Letters

Lot size	Inspection levels		
	I	II	III
2–8	A	A	C
9–15	A	B	D
16–25	B	C	E
26–40	B	D	F
41–65	C	E	G
66–110	D	F	H
111–180	E	G	I
181–300	F	H	J
301–500	G	I	K
501–800	H	J	L
801–1,300	I	K	L
1,301–3,200	J	L	M
3,201–8,000	L	M	N
8,001–22,000	M	N	O
22,001–110,000	N	O	P
110,001–550,000	O	P	Q
550,001 and over	P	Q	Q

SOURCE: MIL-STD-105A, Table III.

table is presented to show the general form of the table construction (Table 12-7). The sample size is based upon the size of the lot and the inspection level desired. Sample-size codes are given in Table 12-8. Ordinarily inspection level II is used. Level III represents a more rigid inspection practice and level I a less rigid practice.

As an example of the use of the tables, suppose that a concern is producing an item in lots of 1,000 and that the product is assumed to be satisfactory if fewer than 4 per cent are defective. The process has been operating within prescribed limits, so that normal inspection can be used. What sampling plan should be adopted? Table 12-8 shows that for inspection level II (the customary level) and a lot size of 1,000 we shall need sample-size code K. Referring to Table 12-7, we see that for size class K and acceptable quality level 4.0 we shall take a first sample of size 75. If 5 or fewer are defective, we accept the lot. If 12 or more are defective, we reject. If between 6 and 11, inclusive, are defective, we take a second sample of 150. If 11 or fewer are defective out of the combined sample of 225, we accept. If 12 or more are defective, we reject.

EXERCISE 12-7. Items are produced in lots of 2,400, and a lot is acceptable if no more than 2.5 per cent are defective. (*a*) In normal circumstances, what double-inspection plan would be used? (*b*) Under tightened inspection, what plan would be used?

13. OPERATING-LEVEL PROBLEMS

13-1. INTRODUCTION

Many problems in management can be characterized as "arrival-and-departure" problems. For example, the customers of a restaurant arrive and wait to be served. After completing their meal (after their "service" is completed), they depart. The same can be said for customers at a filling station, a public pay phone, a department in a department store, and a cashier's window at a bank. Similarly, a machine in a factory may be said to "arrive" in the repair department's area of responsibility when it goes out of order and to "depart" when it is put back in service, whether the machine is actually moved or not.

In the above examples the persons or things requiring service arrive at random intervals of time which follow some probability distribution (whether known or not) and depart after a servicing time, which is another random variable following another probability distribution. Clearly, if the arrival rate is greater than the departure rate, a waiting line will form, so that sometimes these problems are called *queuing* problems.

Other operating-level problems are characterized by a single probability distribution rather than two. For example, a store may stock a given number of Christmas trees. Sales of the trees may be assumed to follow a probability distribution. Each tree stocked costs a given amount of money and will become worthless after Christmas. The question is, how many to stock? A similar problem concerns size of lot to produce. Suppose that 100 units are needed to fill an order but that spoilage rates are known to follow a probability distribution. How many units should be started?

Management's decision on operating level is good or bad according to whether it meets or fails to meet some objective of management. Some possible objectives might be:

1. Always have servicing capacity available so that no unit must wait.
2. Have enough servicing capacity so that waiting units will exceed a predetermined number only a small (but predictable) percentage of the time.
3. Have the servicing capacity available which will maximize profit.

There could be many others. The important point is that the statistician cannot "solve" a servicing problem for management unless he has been given an objective toward which to work. An objective such as "Provide the maximum service to the customer at the minimum cost" simply will not do. The maximum service certainly will not be at minimum cost. The problem of objectives becomes particularly knotty when qualitative criteria are permitted to enter, such as "No customer shall be dissatisfied." For convenience, we shall stick to cost or profit objectives in this brief introduction to the field.

A further complication arises when there are conflicts in objectives within the firm itself. The over-all objective of a firm may be to produce a profit for its stockholders. Parenthetically, it should be remarked that a company's objectives are never that simple; actually they contain many secondary objectives, such as maintaining steady employment and providing service to the public. For convenience, however, let us limit our discussion to the profit motive. How does this objective become translated into a meaningful objective for the manager of the department which services the producing machinery of the firm? Since he is not producing salable units, he has no profit objective. He probably has a tendency to think of his objective as maximum service, by virtue of the nature of his duties. In line with the firm's general objective, however, the specific objective prescribed for him may be "Provide machine servicing at minimum total cost." Here minimum total cost would certainly include the cost of "down time" of the machines, as well as direct servicing costs. This subobjective, then, may be considered to be in line with the objectives of the total firm.

This process of a subunit of the organization working toward its own objective is called *suboptimization*. A firm is fortunate indeed if all suboptimization activities are in accord with the major objectives of the organization, even if such major objectives have been defined. There is reason for believing that most company decisions are of the suboptimizing variety. The task is simply too great to fit every decision into an over-all framework of optimization. Most of the examples which follow illustrate attempts at suboptimization where the objective is clearly stated.

13-2. MODELS

Frequent reference to models has been made throughout this text. It is time for us to take a careful look at the process of model building and to see how models can serve us in decision making.

In the first place, there are many kinds of models. Architects make models of buildings, engineers make models of machines and dams, and designers make models of the things they design. These are physical models. We are interested here in mathematical models. Mathematical models are similar to physical models in that both are representations of something. Furthermore, neither kind of model *is* the thing it represents. This is extremely important. We sometimes become so intoxicated with the beauty of our models (mathematical, that is) that we forget they have been constructed out of our imagination. They are useful only if they are reasonably good representations of the system they model.

The concept of a mathematical model certainly should not be new to one who has worked his way through 12 chapters of a statistics book, but it is not always obvious that a model is being used as the basis for a decision. Let us review a few steps along the way.

Consider the physical system consisting of a device for flipping coins. Our mathematical model of this physical system is the binomial distribution with $p = q = \frac{1}{2}$. The behavior of our mathematical model will approximate the behavior of the physical system only if the coin being tossed is, in fact, a "fair" coin.

Consider another case. We believe that the average number of errors in setting type per typewritten page is 2. We want to know what proportion of our pages will have no errors. We use the Poisson table (Appendix Table V) with parameter 2 and find that our answer is 0.135 (to three places of accuracy). This is an exact solution to the problem posed by our mathematical model, but it is not a solution to our physical problem *unless the Poisson distribution with parameter* 2 *is an exact representation of that system.*

Similarly, each time we assumed the normality of our populations, as well as the more obvious times we assumed regression or analysis of variance models, we were basing our solution on a model. Our "solutions" were not solutions to the actual problem unless the "assumptions" were met, or approximately so.

The point is that a mathematical model needs to be verified. In verifying a probability model, one never expects to achieve experimental results exactly in accord with the model, but the results should be reasonably close, as judged, generally, by the χ^2 test for fit (Chap. 6).

The question, why use models? perhaps needs to be answered. We have seen above that models provide the basis for the statistical decisions

we make. Another use, which we discuss in this chapter, is to examine the operation of the model, rather than the physical system it represents, in order to arrive at decisions about the system. For example, if we construct a "good" model of an arrival-and-departure problem, we can tell by examination of the model whether or not a decision to provide service capacity at a particular level is a "good" decision. This decision, based upon the operation of the model, avoids the cost inherent in trying out various levels in the real system in order to arrive at the proper level. Needless to say, the process only works when the model is a good one.

13-3. PROBLEMS WITH A SINGLE PROBABILITY DISTRIBUTION

Suppose Mr. Shivers sells ice-cream cups in the park on Sundays. He buys a stock of ice cream on Sunday morning, and any part which is not sold that day melts and is thrown away. Suppose he buys the cups for 4 cents each and sells them for 10 cents. Suppose, further, that he (being a statistician at heart) has found that demand for ice-cream cups is normally distributed with mean 300 and standard deviation 40. How many ice-cream cups should he buy each Sunday morning in order to maximize his profit in the long run?

Let $P(y)$ be the probability that the yth cup will be sold. That is, $P(200)$ is the probability that 200 or more cups will be sold. It will be seen that, when y is small, $P(y)$ is near 1, and as y becomes large, $P(y)$ approaches 0. The probability that the yth cup will not be sold is $1 - P(y)$. By well-established economic theory, profits will be maximized when the expected gain from the next unit purchased just equals the expected cost. In economic terms this is the point at which "marginal profit" is equal to 0. That is, an additional unit will reduce average profit, as will 1 unit less.

Now, the expected profit on the yth cup is the profit per cup times the probability that the cup will sell. That is, if one receives 10 cents each time heads is obtained in tossing a fair coin, he expects, on the average, to make 5 cents per toss, that is, the gain per coin times the probability of gain. Similarly, the expected loss is the loss per item that doesn't sell times the probability that it won't sell. If we let L equal the loss per item that doesn't sell and G equal the gain per item sold, then equalizing expected gain and expected loss on the next item stocked results in

$$P(y)G = [1 - P(y)]L \tag{13-1}$$

Solving this expression for $P(y)$, we obtain

$$P(y) = \frac{L}{G + L}$$

That is, Mr. Shivers should stock a number of ice-cream cups, Y, such that the probability of that many or more being sold is equal to the ratio of loss per unit on unsold items to the sum of gain per unit and loss per unit. The result is independent of the type of probability distribution used. In Mr. Shivers' case we are working with the normal distribution. We have

$$P(y) = \frac{L}{G + L} = \frac{0.04}{0.06 + 0.04} = 0.4$$

We seek the value of Y (say, y) such that $P(Y > y) = 0.4$. From Appendix Table I we find

$$P(Z > 0.25) = 0.4$$

Since $Z = (Y - \mu)/\sigma$, we have $Y = \sigma Z + \mu$, so

$$P[Y > 40(0.25) + 300] = 0.4$$

or

$$P(Y > 310) = 0.4$$

Therefore Mr. Shivers should stock 310 cups of ice cream in order to maximize his profits in the long run.

Exercise 13-1. Find out what Mr. Shivers' expected profit is if he stocks (*a*) 310 cups, (*b*) 320 cups, (*c*) 300 cups.

Exercise 13-2. Suppose Mr. Shivers can return unsold cups for 2 cents each. How many should he stock?

Exercise 13-3. Suppose that if Mr. Shivers sells ice cream at 15 cents per cup, his demand is normal with mean 150 and standard deviation 30. (*a*) Under the same cost conditions as in the illustration in the text, should he charge 15 or 10 cents? (*b*) If he charges 15 cents, how many cups should he stock?

Next, we consider a lot-size problem. Suppose that the Ace Specialty Company has received an order for 200 gadgets. It costs \$65 to set up the machines to produce gadgets, and it costs \$1.50 in variable costs per unit started. Suppose that the spoilage rate is estimated at 4 per cent and that spoiled units have no salvage value. Also, assume that overproduction has no market value. How many units should be started?

It may be assumed that, since the probability of a spoiled unit is small and the total lot size is fairly large, the number of spoiled units will follow closely the Poisson distribution. Before attempting a solution to the problem, let us investigate the nature of the costs involved. Suppose 210 units are started and 198 are finished. Certainly, then, the cost is \$65 + 210(\$1.50) for the first batch, plus \$65 + 2(\$1.50) for the second batch, since the first lot did not supply 200 good items. The total cost is then \$448. On the other hand, suppose the 210 units started yielded 205 good units. Then the total cost is \$65 + 210(\$1.50) = \$380. But now we have 5 good units more than we need to fill the order.

Let n be the number of units started and m be the spoiled units. Additional setup costs will be incurred only when $n - m < 200$, and the probability of this cost is $P(n - m < 200)$. Hence the expected additional cost is $\$65P(n - m < 200)$. The total expected cost is

$$C = \$65 + \$1.50n + \$65P(n - m < 200) + \$1.50[P(n - m = 199) + 2P(n - m = 198) + \cdots] \quad (13\text{-}2)$$

The first two terms are the costs of original setup and variable cost, the third term is the expected cost for second setup, and the fourth term is the expected variable cost of the additional units which must be started. It is \$1.50 times the probability that 1 more unit will be needed plus \$3 times the probability that 2 more units will be needed, and so forth.

In order to evaluate the probabilities, note that

$$P(n - m < 200) = P(m > n - 200)$$

Suppose we let $n = 210$ and evaluate (13-2). We can make the evaluation easily by use of Table 13-1.

Table 13-1. Evaluation of Expected Cost When 210 Units Are Started

Number of spoiled units, m	$P(m)$	Cost of additional units, $(m - 10)\$1.50$	Expected cost of additional units, $P(m)(m - 10)\$1.50$
11	0.072	\$ 1.50	\$0.11
12	0 048	3.00	0.14
13	0.030	4.50	0.14
14	0.017	6.00	0.10
15	0.009	7.50	0.07
16	0.005	9.00	0.04
17	0.002	10.50	0.01
Total	0.183		\$0.61

We have indicated above that it is reasonable to assume that m is distributed by the Poisson law with parameter np. With $n = 210$ we have $np = 8.4$, and we require a table of the Poisson distribution showing this value for the parameter. However, since p is estimated, there will be little additional error in using the Poisson table (Appendix Table V) with parameter 8. This table is the source of the probabilities listed in the second column of Table 13-1. The sum of this column is

$$P(n - m < 200) = 0.183$$

The third column is the cost of producing enough additional units to make up the deficit. The fourth column is the expected cost of the

additional units. The total of this column is the total expected cost of additional units.

We are now ready to make the evaluation of (13-2):

$$\begin{aligned} C &= \$65 + \$1.50(210) + \$65(0.183) + \$0.61 \\ &= \$65 + \$315 + \$11.90 + \$0.61 = \$392.51 \end{aligned}$$

We wonder whether we can do better. We note that the expected cost of second setup appears rather large compared with the cost of adding another unit, so it may occur to us to try 211 units. Using the same Poisson table, we are now interested in the probabilities of 11 or more spoiled units. That is, we can eliminate the first row of Table 13-1 and find $P(m > 11) = 0.111$. The expected cost of additional units is

$$0.048(\$1.50) + 0.03(\$3) + 0.017(\$4.50) + 0.009(\$6) + 0.005(\$7.50) + 0.002(\$9) = \$0.35$$

The total cost is

$$\begin{aligned} C &= \$65 + \$1.50(211) + \$65(0.111) + \$0.35 \\ &= \$65 + \$316.50 + \$7.22 + \$0.35 = \$389.07 \end{aligned}$$

Since the expected cost here is less than it is under the other plan, we should favor this plan over starting only 210 units.

EXERCISE 13-4. Continue the above process until you find the value of n which minimizes expected cost.

It will be noted that the last term of (13-2) contributes little to total expected cost. If we ignore this term, we shall add units until the reduction in expected cost of second setup is equal to the cost of starting another unit ($1.50). The computation is facilitated by use of a table of the cumulative Poisson distribution. Note that we have ignored the probability of spoilage on the second run. A more exact solution would take this into account.

It must be remembered that a solution which yields a minimum expected cost does not guarantee minimum cost in a particular trial. It is only when one considers a long sequence of minimum-cost decisions that he has any assurance of actually achieving minimum cost. This may not be a very consoling thought to the young manager who recommends a minimum-expected-cost solution and finds himself 1 unit short of the required number on the first try.

EXERCISE 13-5. In the previous illustration, assume that the desired quantity is 2,000, rather than 200. How many units should be started? Note that now the number of units spoiled should follow closely the normal distribution with mean np (where $p = 0.04$, the average spoilage rate, and n is the number of units started) and variance npq. To accomplish the solution, it may be helpful to

find the expected cost for several values of n, as a means of locating the approximate value of n which will yield the minimum cost. A solution within 5 units of the optimum number is certainly adequate in this case. The solution for $n = 2{,}050$ is given here as an example:

$$np = 2{,}050(0.04) = 82$$
$$\sqrt{npq} = \sqrt{2{,}050(0.04)(0.96)} = 8.87$$

The probability that fewer than 2,000 good units will be produced is

$$P(n - m < 2{,}000) = P(m > 50)$$
$$= P\left(Z > \frac{50 - 82}{8.87}\right) = P(Z > -3.6)$$

which is close to 1.00. Therefore, by starting only 2,050 units one is almost certain to incur the additional setup cost plus the cost of starting approximately 32 new items (the difference between 82 and 50). One can certainly do better than this and probably would move to some value higher than 2,080 for the next trial.

13-4. A NOTE ON EXPECTATION

In Chap. 8 we discussed expected values at some length, and throughout this book we have used the technique for averaging consequences which follow probability distributions. In computing an expected cost or an expected profit, there is an assumption which is sometimes overlooked. The assumption is that the "utility" of money remains constant, regardless of the amount at stake. That is, the first dollar of cost (or profit) is no more or less important than the millionth dollar. This assumption is unrealistic over wide ranges, but it may be reasonable over the relatively narrow range of costs or projects associated with typical alternative business decisions.

Consider this example. You have a probability of 0.001 of winning \$1,000 or a probability of 0.5 of winning \$2. Each risk costs you \$1. Which will you prefer? Each has an expected value of \$1, but it is safe to say that opinions will vary greatly concerning the one which is to be preferred. Perhaps the point will be clear if we contrast a 0.999 probability of winning \$1,000 (which costs you \$999) with a 0.001 probability of winning \$1,000 (which costs you \$1). There is evidence that people in general prefer the latter type of risk. How else can one account for the fact that they will pay \$1 for a lottery ticket which represents a very small probability, say, 0.0001, of winning \$1,000?

These considerations enter expected costs when there is one possible outcome which is extremely costly but which has a small probability of occurrence. If one follows the minimum-expected-cost rule blindly, he takes no special precautions to avoid this eventuality. In practice,

however, it has been found that management will choose policies which are not consistent with minimum cost in order to avoid the possibility of large losses. It may be helpful in such cases to evaluate the cost of such "disaster avoidance."

13-5. QUEUING PROBLEMS

Many of management's scale-of-operations decisions involve a choice of the amount of facilities required to provide service, either to customers or to other departments of the company. The number of salesclerks, the number of repairmen, and the number of machine operators assigned to automatic machines fall into this category. A characteristic of such problems is that units to be serviced (customers or machines) arrive at random and require a random length of time for service. When arrivals occur faster than departures, a waiting line, or queue, forms, hence the name given to this class of problems.

There is a determinable cost associated with maintenance of each unit of servicing facilities and another cost (not always determinable) associated with requiring the unit to wait in the waiting line. The management problem is to choose that level of servicing facilities which will minimize total cost.

The reason that waiting cost is not always determinable is that one cannot always predict the behavior of the waiting unit. For example, if Mrs. Jones must wait at the check-out counter of a chain grocery store, she may (1) not let it bother her, (2) fail to purchase groceries at this market on this particular day, or (3) stop trading at the store altogether. There are still other actions, such as influencing the trade of others. It is clear that assignment of cost of waiting in such cases is somewhat arbitrary.

The problem of waiting lines, and associated costs, becomes more complex when priorities are established for the servicing of units in the queue. Also, peculiar arrival-and-departure distributions can further complicate solutions. In fact, a comprehensive treatment of the queuing problem is beyond the scope of this elementary text,* and the discussion here is limited to some well-known results, which may serve to prepare the student for better understanding of more complex cases when they arise.

We shall assume that the population of units to be served is infinite, so that the probability of arrival of a unit in a fixed interval of time remains constant, regardless of the length of the waiting line. It is easy to see

* The student with a background in calculus who wishes to pursue the study of queues further may be interested in Philip M. Morse, *Queues, Inventories and Maintenance*, John Wiley & Sons, Inc., New York, 1958.

that with a finite population and a long waiting line the probability of arrival of a unit in a fixed interval of time is reduced, because there are fewer units remaining in the "calling population." Also, we shall assume that there are no priorities in the waiting line. Finally, we shall assume that arrivals and departures are random and that the probability distributions of arrivals and departures remain constant over time.

We let A equal the average rate of arrival per unit of time and D equal the average rate of departure per unit of time. It seems clear that, if A is greater than D, an infinite waiting line will develop. We must assume, therefore, that $A < D$. It can be shown* that under these conditions the probability of n units in the waiting line at time t is

$$P(n) = \left(\frac{A}{D}\right)^n \left(1 - \frac{A}{D}\right) \tag{13-3}$$

For example, if the average arrival rate is 3 units per hour and the average service (or departure) rate is 4 units per hour, then

$P(0) = (\frac{3}{4})^0(1 - \frac{3}{4}) = \frac{1}{4}$ (the probability of a waiting line of zero)
$P(1) = (\frac{3}{4})^1(1 - \frac{3}{4}) = \frac{3}{16}$
$P(2) = (\frac{3}{4})^2(1 - \frac{3}{4}) = \frac{9}{64}$

and so forth.

We can compute the average length of waiting line by the rules we have learned about expected values. That is, the average length is

$$E(n) = 0P(0) + 1P(1) + 2P(2) + 3P(3) + \cdots \tag{13-4}$$

We can actually insert numbers and evaluate a finite number of terms of this infinite series in order to arrive at an approximation to the average value. We may observe, however, that

$$E(n) = \sum_{n=0}^{\infty} n \left(\frac{A}{D}\right)^n \left(1 - \frac{A}{D}\right) \tag{13-5}$$

and this infinite series has the value

$$E(n) = \frac{A/D}{1 - A/D} \tag{13-6}$$

In our numerical illustration the average length of waiting line with $A = 3$, $D = 4$ is

$$E(n) = \frac{\frac{3}{4}}{1 - \frac{3}{4}} = 3$$

* For example, see C. West Churchman, Russell L. Ackoff, and E. Leonard Arnoff, *Introduction to Operations Research*, pp. 393ff., John Wiley & Sons, Inc., New York, 1957.

Now we are ready to consider a management problem. Suppose a manufacturing company has a large number of identical machines which operate automatically but which occasionally require the attention of an operator. The machines in a textile plant might be considered of this type. We assume that the number of machines is large enough so that the length of the queue of machines requiring operator attention will not affect materially the probability of another machine's joining the queue. Suppose also that the average number of machines requiring attention per hour is 30 and that an operator can, on the average, clear the difficulty on a machine in 5 min, so that the servicing rate is 12 units per hour. Note that the unit of time chosen is arbitrary, since the unit will cancel out in the ratio of arrival rate to departure rate A/D. Since $A = 30$ and $D = 12$, it is obvious that one operator cannot serve all the machines. The question is, how many operators should be employed?

We need to introduce costs in order to arrive at a solution. Suppose that the direct cost assignable to an operator is \$3 per hour and that the value of the production of a machine is \$10 per hour. It is clear that it will cost a lot of money to have many machines in the waiting line (i.e., standing idle).

We know that D must be greater than A in order to serve the machines effectively and we also know that D can be increased only in units of 12 per hour. Suppose as our first trial we consider 3 operators, so that $D = 36$. Then the average waiting line is

$$E(n) = \frac{\frac{5}{6}}{1 - \frac{5}{6}} = 5$$

It costs us \$50 per hour to have 5 machines idle plus \$9 per hour to hire 3 operators, an expected total relevant cost of \$59 per hour. We think we might do better by employing 4 operators.

Then
$$E(n) = \frac{\frac{30}{48}}{1 - \frac{30}{48}} = \frac{5}{3}$$

Now, our expected total relevant cost per hour is

$$\tfrac{5}{3}(\$10) + 4(\$3) = \$28.67$$

which is a considerably better plan. We try 5 operators and obtain the following results:

$$E(n) = \frac{\frac{30}{60}}{1 - \frac{30}{60}} = 1$$

Now the expected total relevant cost is \$25. A sixth operator will yield

$$E(n) = \frac{\frac{30}{72}}{1 - \frac{30}{72}} = \frac{5}{7}$$

and now the expected cost is \$25.14. It may be seen that adding more operators will add more cost, so our optimum solution is to employ 5 operators.*

EXERCISE 13-6. Suppose that in the above illustration the cost of lost production is \$6 per hour. Compute the optimum number of operators to be assigned.

13-6. THE MONTE CARLO METHOD

Those who, upon reading the title to this section, expect to learn how to make a killing in the gambling casinos are doomed to disappointment. The Monte Carlo method refers to a process by which "experience data" are produced synthetically by some form of random-number generator.

This procedure should not seem unfamiliar, for we have referred to it on many previous occasions. Note particularly that in Sec. 5-4 we were concerned with the empirical production of probability distributions for t, χ^2, and F by drawing of random numbers from an approximately normal population. Intuitively, we can see that the process, if continued long enough, will produce a close approximation to the true distribution of t, χ^2, or F. In such cases there is no point in constructing empirical distributions of these variables (except for classroom-demonstration purposes), because the probability distributions have been derived mathematically. There are many situations in statistics, however, where the mathematical form of the probability distribution is so complex that it cannot be derived mathematically, or at least it may be so complex that such derivation is not worthwhile. In such cases, one may construct an empirical probability distribution by the Monte Carlo method.

For example, suppose x is normally distributed with mean 50 and standard deviation 10, y is a χ^2 variable with 4 degrees of freedom, and z is a Poisson variable with parameter 5. We want to find the probability distribution of

$$f(x,y,z) = \frac{xz}{y}$$

We could make a random drawing from each of x, y, and z by a technique to be presented later, do the computation xz/y, and record the resulting value. Continuing this process for, say, 1,000 times would yield an approximation to the true distribution of xz/y. The procedure sounds tedious, but if one has access to an electronic computer, the empirical distribution can be generated quite rapidly.

* The student of the calculus will observe that he may obtain a solution quickly by setting the partial derivatives of $C = \$10\ [30/(D - 30)] + (D/12)(\$3)$ equal to 0 and solving for D. This will yield a value of D which is not divisible by 12, but should permit one to obtain an optimum solution by trying at most two values.

A more important application of the Monte Carlo method in management is the simulation of actual experience to judge the effectiveness of alternative solutions to a management problem. The simulation, being accomplished by a computer, can be done prior to installation of the proposed plan of action. If the model upon which the simulation is based is correct, or nearly so, costly errors in management decision can be avoided by the Monte Carlo method.

We shall illustrate the use of the Monte Carlo method with an inventory problem. First of all, we consider a simple generalization of the problem. We assume that it costs a fixed amount C_o to place an order. We also assume that it costs an amount C_I to hold an item in inventory for a unit of time. A third consideration is the shortage cost C_s, which represents the loss in revenue resulting from not having an item in stock when it is demanded by a customer. We shall assume that, if an item is not in stock when demanded, the customer will obtain it from another source. Another common assumption about shortage cost is that demand for the item continues during the period of shortage but that the shortage cost is a fixed quantity per day per item short. We wish to balance the three costs C_o, C_I, and C_s by selecting a quantity Q to be ordered and a reorder point R, which is the level to which we permit inventory to fall before we reorder. We shall assume that demand follows a prescribed probability distribution which is constant over time. That is, the probability of a demand D_1 is the same today as it is any other day. It should be clear from what follows that this assumption can be relaxed to permit a seasonal change in demand, which can be approximated by the methods of Chap. 9. We shall assume that delivery time follows another probability distribution. For convenience we shall take the day as a unit of time and assume that all decisions to reorder, and so forth, are made at the end of the day. We shall also assume that goods will be available for sale on the day they are received. The following costs will be assumed:

C_o = \$50 per order
C_I = \$1 per day
C_s = \$5 per item short

The probability distribution for demand is as presented in Table 13-2, and the probability distribution for delivery time is given in Table 13-3. It may be assumed that these probability distributions have been developed empirically by observations collected over a long period of time.

We now investigate a method by which we can extract a random observation from each of these probability distributions. For this purpose we use the cumulative-probability columns. We draw a two-digit random number from a table of random numbers (Table 1-1) and

Table 13-2. Probability Distribution for Demand (Hypothetical Data for Monte Carlo Illustration)

Units demanded	Probability	Cumulative probability
0	0.01	0.01
1	0.03	0.04
2	0.10	0.14
3	0.12	0.26
4	0.20	0.46
5	0.20	0.66
6	0.15	0.81
7	0.10	0.91
8	0.05	0.96
9	0.03	0.99
10	0.01	1.00

Table 13-3. Probability Distribution for Delivery Time (Hypothetical Data for Monte Carlo Illustration)

Days	Probability	Cumulative probability
2	0.05	0.05
3	0.15	0.20
4	0.30	0.50
5	0.20	0.70
6	0.15	0.85
7	0.10	0.95
8	0.05	1.00

identify it with a number in the cumulative-probability column. The corresponding variable is the random drawing. If, for example, the random number is 12, it is identified with 0.14 in the cumulative-probability column of Table 13-2, because 12 is greater than 4 but equal to or less than 14. Therefore, we assume that for the particular day in question 2 units were demanded (since 2 is the variable which corresponds to the cumulative probability of 0.14). If the random number were 81, then 6 units would be demanded; if 63, then 5 units would be demanded, and so forth. It may be seen that the variable "3 units demanded" has 12 chances out of 100 of being drawn, the variable "4 units demanded" has 20 chances out of 100 of being drawn, and so forth. Our method of drawing reproduces the probability distribution, then, when a large number of drawings is considered. Using random drawings from Tables 13-2 and 13-3 simulates actual experience with an inventory plan.

Table 13-4. Simulated Experience with Inventory Plan, Generated by Monte Carlo Method

Day	Random numbers: For demand	Random numbers: For delivery time	Demand	Receipts	Balance	Inv. cost, C_I	Order cost, C_o	Shortage cost, C_s
0	...	.....	...	...	30			
1	57	.....	5	...	25	25		
2	01	. .	0	...	25	25		
3	14		2	...	23	23		
4	45	76 (6)	4	...	19	19	50	
5	63		5	...	14	14		
6	42	.	4	...	10	10		
7	76	.....	6	...	4	4		
8	19	....	3	..	1	1		
9	03	. . .	1		0	0		
10	30	....	4	30	26	26		
11	36	.. ..	4	.	22	22		
12	97	41 (4)	9	. .	13	13	50	
13	42	. .	4	..	9	9		
14	04	.....	1	...	8	8		
15	20	.....	3	. .	5	5		
16	19	.....	3	30	32	32		
17	68	...	6	...	26	26		
18	58	.....	5	...	21	21		
19	20	30 (4)	3	...	18	18	50	
20	14	. . .	2	...	16	16		
21	38	... .	4	.	12	12		
22	13	.. ..	2	..	10	10		
23	95	. ..	8	30	32	32		
24	81	.	6	...	26	26		
25	90	80 (6)	7	..	19	19	50	
26	05	. . .	2	...	17	17		
27	15	.....	3	...	14	14		
28	74	.....	6	...	8	8		
29	81	.....	6	...	2	2		
30	34	....	4	. .	0	0	...	10
31	46	....	4	30	26	26		
32	88	85 (6)	7	.	19	19	50	
33	65	.....	5	..	14	14		
34	14	.....	2	...	12	12		
35	42	.....	4	...	8	8		
36	83	.....	7	...	1	1		
37	30	.....	4	...	0	0	...	15
38	65	.....	5	30	25	25		
39	02	.....	1	...	24	24		
40	66	01 (2)	5	...	19	19	50	
41	62	.....	5	...	14	14		

Table 13-4. Simulated Experience with Inventory Plan, Generated by Monte Carlo Method (*Continued*)

Day	Random numbers		Demand	Receipts	Balance	Inv. cost, C_I	Order cost, C_o	Shortage cost, C_s
	For demand	For delivery time						
42	34	.....	4	30	40	40		
43	31	.....	4		36	36		
44	71	.....	6	. .	30	30		
45	71	.....	6	. . .	24	24		
46	21	.....	3	. . .	21	21		
47	16	13 (3)	3	. . .	18	18	50	
48	39	. . . .	4	. .	14	14		
49	33	.....	4	. . .	10	10		
50	63	.....	5	30	35	35		
51	72	.....	6	.	29	29		
52	53	.....	5		24	24		
53	46	04 (2)	4	.	20	20	50	
54	77	. . .	6	.	14	14		
55	26	.....	3	30	41	41		
56	49	.....	5	.	36	36		
57	90	.....	7	. .	29	29		
58	59	.....	5	. .	24	24		
59	23	.....	3	. . .	21	21		
60	71	24 (4)	6	. . .	15	15	50	
61	21	. . . .	3	. . .	12	12		
62	79	....	6	. .	6	6		
63	48	.....	5	. .	1	1		
64	80	.....	6	30	25	25		
65	58	45 (4)	5		20	20	50	
66	68	.	6	.	14	14		
67	74	.....	6		8	8		
68	02	.....	1	. .	7	7		
69	66	.....	6	30	32	32		
70	73	.....	6	. .	26	26		
71	13	.....	2	. . .	24	24		
72	26	.....	3	. .	21	21		
73	83	59 (5)	7	. . .	14	14	50	
74	33	.....	4	. . .	10	10		
75	23	.....	3	. . .	7	7		
Total						1,327	550	25

For our illustration, we shall evaluate the inventory plan which calls for a reorder point of 20 and an order quantity of 30. The random drawings which simulate experience under this plan, along with costs incident to this simulated experience, are shown in Table 13-4.

It is necessary to draw a random number for each day's demand, but a random number for delivery time needs to be drawn only when an order is placed. We must assume some value for beginning inventory, and for convenience we have chosen 30. The random number 57 specifies a demand for 5 units (Table 13-2), reducing the inventory to 25 units, and so forth. At the end of the fourth day, inventory has dropped to 19 (equal to or below the reorder point of 20); hence we incur an order cost of \$50 and happen to draw the random number 74 to specify the delivery time. Reference to Table 13-3 shows that this indicates a delivery time of 6 days, so on the sixth day after the order (i.e., the tenth day) we record the receipt of 30 units, which, by assumption, are available for sale on the day received. There is no shortage cost, since demand for both the ninth day and tenth day can be met. The rest of the table is filled in similar fashion.

The first shortage cost is encountered on the thirtieth day, when 4 units were demanded but only 2 were available. Since the shortage cost is \$5 per unit, we incur a shortage cost of \$10 at this time.

Seventy-five days of simulated experience is shown in Table 13-4. In practice, one would probably want considerably more. However, it is clear from the totals of Table 13-4 that our proposed inventory plan is not the best possible. We note that our inventory cost is \$1,327, our order cost \$550, and our shortage cost only \$25. It is apparent that we can reduce inventory cost either by ordering smaller amounts more frequently or by reducing the reorder point to some value less than 20. The first alternative will increase order cost, which is already relatively high. The second alternative will increase shortage cost, which is quite low, so this is probably the alternative we should test out.

EXERCISE 13-7. Using the same random numbers, evaluate the inventory plan if the reorder point is dropped to 18 or less.

In a problem such as this one, which requires the determination of two parameters (in this case the quantity to be ordered and the reorder point), it is customary to prepare a two-way table of costs under various combinations of the two parameters. For this problem such a table might look like Table 13-5.

One would hope that such a table would locate the neighborhood of the optimum, so that by further slight refinement one could choose the lowest-cost plan.

EXERCISE 13-8. Using different random numbers, evaluate the inventory plan with reorder point 18 and reorder quantity 24. Use 100 days of simulated experience.

Sometimes it is possible to come close to the optimum solution by using the averages of the probability distributions. To see how we can

Table 13-5. Total Inventory Costs Generated by the Monte Carlo Method under Various Combinations of Reorder Points and Order Quantities

Reorder quantity	Reorder point							
	14	16	18	20	22	24	26	28
20								
22								
24								
26								
28								
30								
32								
34								
36								

use this technique on inventory problems it is necessary to investigate some of the facets of the general inventory problem. We consider, for simplicity, an inventory model in which demand is constant and fixed. That is, daily demand does not follow a probability distribution but is the same for all days. Similarly, delivery time is fixed. Then an

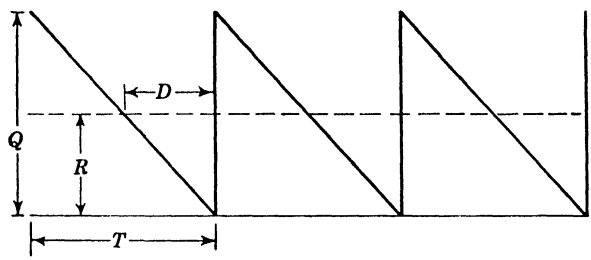

FIG. 13-1. Inventory cycle under constant demand and fixed delivery time

"inventory cycle" consists of an initial inventory quantity equal to the amount of an order, a constantly declining inventory until a reorder point is reached, placement of an order, further decline in inventory to zero, and then the delivery of the new order, whereupon the cycle begins again. The cycle is portrayed in Fig. 13-1.

We use the notation previously developed. That is,

R = reorder level
Q = quantity to be ordered
C_o = order cost per order
C_I = inventory cost per unit of time
C_s = shortage cost per item short

We add the following notations:

D = delivery time

d = Q/T demand per unit of time

T = length of cycle in time units

It might be more realistic to portray the level of inventory as a stair-step function than as a downward sloping line, since we assume that a discrete number of units is taken out during each time unit. However, the unit of time is arbitrary, and if it is taken arbitrarily small and a large number of units are demanded during the cycle, the inventory level can be made to approximate the straight lines of Fig. 13-1.

It will be noted that the model portrayed by Fig. 13-1 does not permit any shortage. This is due to our assumption that shortage cost is incurred the moment the unfulfilled demand arises. It is not, then, a function of time, and since we can predict delivery time exactly, there is no reason for incurring shortage cost in this fixed model.

Consider a total time of K units. Then total demand is Kd, and the number of cycles during the total time period is Kd/Q. Also, we have

$$T = \frac{K}{Kd/Q} = \frac{Q}{d}$$

The average number of inventory-time units during a cycle is

$$\frac{QT}{2} = \frac{Q^2}{2d}$$

There are Kd/Q cycles; hence the average number of inventory-time units over the entire time K is $KQ/2$. Total cost is the average units in inventory times inventory cost per unit plus the number of orders (cycles) times cost per order. That is

$$C = \frac{KQ}{2} C_I + \frac{Kd}{Q} C_o \tag{13-7}$$

We wish this cost to be a minimum. By the differential calculus it is easy to show that minimum cost will be achieved when

$$Q = \sqrt{\frac{2dC_o}{C_I}} \tag{13-8}$$

Referring to Table 13-2 we find that the average daily demand is 4.76 units. Inserting this average for d in Eq. (13-8), along with the

assumed values for C_o and C_I, yields

$$Q = \sqrt{\frac{2(4.76)(50)}{1}} = 22 \text{ units} \qquad \text{(approximately)}$$

This might serve as a trial value for Q in the Monte Carlo solution.

Examination of Table 13-3 shows that average delivery time is 4.75 days. Therefore we should place an order approximately 5 days before inventory is exhausted and an average of 5 days' sales is approximately 24 units. However, goods are available to meet demand on the day they arrive so 19 units might be used as a starting value for R in the Monte Carlo solution.

It should be emphasized that these starting values do not constitute an optimum solution, except by chance. This should be obvious when one recalls that the fixed model does not permit shortage. In our illustration, shortage cost is relatively minor with respect to inventory cost, so we would expect R to be considerably less than 19. Limited computation with an electronic computer has shown that this is the case. The optimum values of Q and R appear to be in the neighborhood of $R = 9$ and $Q = 25$ so that actually there is zero inventory a good share of the time.

EXERCISE 13-9. Redo Ex. 13-8 assuming that shortage cost is \$5 *per unit per day* and that when goods are received they must be used first to fill back orders. This assumption is likely to fit a factory where shortage of components must be met before regular production is resumed.

A question may arise in the reader's mind concerning how long to continue the Monte Carlo process before one is reasonably certain that one plan is better than another. There is no simple answer, but the following procedure might be considered to test whether, for a given number of trials, one plan is better than another.

a. Perform each Monte Carlo evaluation by an independent set of numbers from the table of random numbers.
b. Suppose each evaluation consists of the same amount of simulated experience, e.g., 500 days, 10,000 operations, etc.
c. Divide this experience into k subsets representing equal amounts of experience. For example, in the inventory problem one might choose subsets of 50 days' experience. One must be careful to choose a long enough period that occasional random drawings (such as the choice of a delivery time) will not unduly influence the results.
d. Use either the $\bar{D}$ test (Sec. 5-5) or the sign test (Sec. 6-5) to test the hypothesis that costs resulting from operations of the two plans are equal.

It is hoped that this section will have impressed the reader with the

usefulness of the Monte Carlo method in evaluating management decisions. In this section we have shown how it can be applied to an inventory problem, but it should be obvious that the method can be effectively used in a wide range of situations. For example, if one knows the probability distributions for failure of the components of a mechanism (such as a digital computer), he can use the Monte Carlo method to obtain a probability distribution for failure of the entire mechanism. Or, if one knows the arrival-and-departure distributions in a queuing problem, he can use the method to assign the proper level of service facilities. It is particularly useful here if several classes of service are provided to various classes of arrivals, so that the theoretical distributions become complicated mathematically.

The wide availability of the electronic computer has added materially to the importance of the Monte Carlo technique. Most computers have automatic routines for producing random numbers, so that the computer can produce and summarize very quickly vast amounts of simulated experience.

APPENDIX

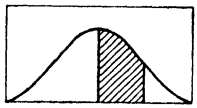

Table I. Table of Areas of the Normal Curve

$(Y - \mu)/\sigma = Z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0 0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3233
0 9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1 9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4758	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2 1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4989	.4988	.4989	.4989	.4989	.4990

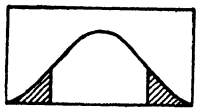

Table II. Distribution of t

Degrees of freedom	Probability												
	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63 657	636 619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6 965	9 925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12 924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4 604	8 610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4 032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0 896	1.119	1.415	1.895	2.365	2 998	3 499	5 408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5 041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3 250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4 587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3 055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3 012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2 947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2 921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3 965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2 552	2.878	3.922
19	0.127	0.257	0.391	0 533	0.688	0.861	1.066	1 328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2 819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3 745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0 256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2 771	3 690
28	0.127	0.256	0.389	0.530	0.683	0 855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1 699	2.045	2 462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0 681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0 848	1.046	1.296	1.671	2 000	2.390	2.660	3 460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

SOURCE: Table II is reprinted from Table III of Ronald A. Fisher and Frank Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 4 ed., 1953, published by Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

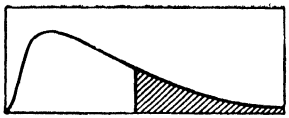

Table III. Table of Chi Square

Degrees of freedom	Probability that chi-square value will be exceeded									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^3393	0.0158	2.71	3 84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4 61	5.99	7.38	9 21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19 68	21.92	24.72	26.76
12	3.07	3 57	4 40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4 11	5.01	5 89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5 14	5.81	6 91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7 63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9 59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8 03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8 64	9 54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9 26	10.20	11 69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11 52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11 81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26 51	29.05	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.22
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.32
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67 33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2
Z_α	−2.58	−2.33	−1.96	−1.64	−1.28	+1.28	+1.64	+1.96	+2.33	+2.58

NOTE: For $\nu > 100$ (i.e., for more than 100 degrees of freedom) take

$$\chi^2 = \nu\left[1 - \frac{2}{9\nu} + Z_\alpha\sqrt{\frac{2}{9\nu}}\right]^3 \quad \text{or} \quad \chi^2 = \tfrac{1}{2}[Z_\alpha + \sqrt{(2\nu - 1)}]^2$$

according to the degree of accuracy required. Z_α is the standardized normal deviate corresponding to the α level of significance, and is shown in the bottom line of the table.

SOURCE: By permission of Prof. E. S. Pearson, from Catherine M. Thompson, "Tables of the Percentage Points of the Incomplete Beta Function and of the χ^2 Distribution," *Biometrika*, vol. 32, pp. 168–181, 188–189, 1941.

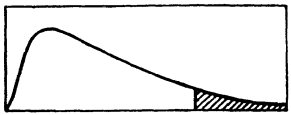

Table IV. F Distribution

df for denominator	Degrees of freedom for the numerator																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	Upper 10 per cent points																		
1	39.86	49.50	53.59	55 83	57.24	58 20	58.91	59.44	59.86	60.20	60.70	61 22	61 74	62 00	62 26	62 53	62.79	63.06	63.33
2	8.53	9.00	9.16	9 24	9 29	9.33	9 35	9.37	9 38	9 39	9 41	9 42	9 44	9 45	9 46	9 47	9.47	9.48	9 49
3	5.54	5.46	5 39	5.34	5.31	5 28	5.27	5.25	5 24	5 23	5 22	5 20	5 18	5 18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4 11	4.05	4 01	3 98	3.95	3.94	3.92	3.90	3 87	3 84	3 83	3.82	3.80	3 79	3 78	3.76
5	4.06	3.78	3 62	3.52	3.45	3.40	3 37	3.34	3.32	3.30	3 27	3 24	3 21	3.19	3.17	3.16	3.14	3.12	3.10
6	3 78	3.46	3 29	3.18	3.11	3.05	3 01	2 98	2 96	2 94	2.90	2.87	2 84	2.82	2 80	2.78	2 76	2.74	2 72
7	3 59	3.26	3.07	2.96	2.88	2.83	2 78	2.76	2.72	2.70	2.67	2 63	2 59	2 58	2.56	2.54	2.51	2.49	2.47
8	3.46	3 11	2 92	2.81	2.73	2.67	2 62	2 59	2 56	2 54	2.50	2 46	2 42	2 40	2 38	2.36	2.34	2.32	2.29
9	3 36	3.01	2 81	2.69	2.61	2.55	2 51	2 47	2.44	2 42	2.38	2 34	2 30	2 28	2.25	2.23	2.21	2.18	2.16
10	3 28	2.92	2.73	2.61	2.52	2 46	2.41	2 38	2 35	2 32	2.28	2 24	2 20	2 18	2.16	2 13	2.11	2.08	2.06
11	3.23	2.86	2 66	2 54	2 45	2.39	2 34	2 30	2.27	2 25	2.21	2 17	2 12	2.10	2 08	2.05	2 03	2 00	1.97
12	3 18	2 81	2 61	2.48	2.39	2.33	2 28	2 24	2.21	2 19	2.15	2.10	2 06	2.04	2.01	1.99	1.96	1.93	1.90
13	3 14	2 76	2 56	2.43	2 35	2 28	2 23	2 20	2 16	2.14	2.10	2 05	2 01	1 98	1.96	1.93	1 90	1.88	1.85
14	3.10	2.73	2 52	2.39	2.31	2.24	2.19	2.15	2 12	2.10	2.05	2.01	1 96	1 94	1.91	1 89	1.86	1 83	1.80
15	3 07	2 70	2 49	2.36	2 27	2.21	2 16	2 12	2 09	2 06	2.02	1 97	1 92	1 90	1.87	1.85	1.82	1 79	1.76
16	3.05	2 67	2.46	2.33	2 24	2.18	2 13	2 09	2 06	2 03	1.99	1.94	1 89	1 87	1.84	1.81	1.78	1 75	1.72
17	3 03	2 64	2.44	2.31	2.22	2 15	2 10	2 06	2 03	2 00	1 96	1 91	1.86	1 84	1.81	1.78	1.75	1.72	1.69
18	3 01	2 62	2.42	2.29	2.20	2.13	2 08	2 04	2 00	1 98	1.93	1.89	1 84	1 81	1.78	1.75	1.72	1.69	1.66
19	2.99	2 61	2.40	2.27	2 18	2 11	2 06	2 02	1.98	1.96	1 91	1.86	1.81	1 79	1 76	1.73	1.70	1 67	1.63
20	2.97	2.59	2 38	2 25	2.16	2 09	2 04	2.00	1.96	1.94	1.89	1.84	1 79	1 77	1.74	1.71	1.68	1.64	1 61
21	2.96	2.57	2 36	2 23	2.14	2 08	2.02	1.98	1 95	1.92	1 87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2 95	2.56	2.35	2 22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1 60	1.57
23	2 94	2.55	2 34	2.21	2.11	2 05	1.99	1.95	1.92	1.89	1.85	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2 10	2 04	1 98	1.94	1 91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1 53

25	2.92	2.53	2.32	2 18	2 09	2 02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1 69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2 17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1 68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1 67	1.64	1.60	1.57	1.53	1 49
28	2.89	2.50	2.29	2 16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1 65	1.62	1.58	1.55	1.51	1.47
30	2 88	2.49	2.28	2.14	2 05	1.98	1 93	1.88	1.85	1.82	1.77	1 72	1 67	1.64	1.61	1.57	1.54	1 50	1.46
40	2.84	2.44	2.23	2 09	2.00	1.93	1 87	1 83	1 79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1 47	1 42	1.38
60	2.79	2.39	2.18	2 04	1 95	1 87	1 82	1 77	1 74	1.71	1.66	1 60	1 54	1.51	1 48	1.44	1 40	1.35	1.29
120	2 75	2 35	2 13	1 99	1 90	1 82	1 77	1.72	1 68	1 65	1 60	1 54	1.48	1 45	1 41	1 37	1.32	1.26	1.19
∞	2.71	2.30	2 08	1 94	1 85	1 77	1 72	1 67	1.63	1 60	1.55	1 49	1 42	1 38	1 34	1 30	1 24	1 17	1.00

Upper 5 per cent points

1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19 4	19.4	19.4	19.4	19 4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9 12	9.01	8.94	8 89	8.85	8 81	8.79	8.74	8.70	8 66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6 04	6.00	5.96	5.91	5.86	5.80	5 77	5 75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4 62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4 15	4.10	4.06	4.00	3 94	3 87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3 44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3 02	2.98	2.91	2.85	2 77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3 01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2 91	2 85	2.80	2.75	2.69	2.62	2 54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2 92	2 83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2 25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2 76	2.70	2.65	2.60	2.53	2.46	2 39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2 79	2 71	2.64	2.59	2.54	2 48	2.40	2.33	2.29	2.25	2.20	2.16	2 11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2 66	2.59	2.54	2.49	2.42	2 35	2.28	2 24	2.19	2.15	2.11	2.06	2 01
17	4.45	3.59	3.20	2.96	2 81	2 70	2 61	2.55	2.49	2.45	2.38	2.31	2 23	2 19	2 15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2 93	2 77	2 66	2 58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2 02	1.97	1.92
19	4.38	3.52	3 13	2 90	2 74	2 63	2 54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1 93	1 88

Table IV. F Distribution (*Continued*)

df for denominator	Degrees of freedom for the numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	Upper 5 per cent points																		
20	4.35	3.49	3.10	2.87	2 71	2.60	2 51	2 45	2.39	2 35	2 28	2.20	2 12	2 08	2 04	1.99	1 95	1 90	1.84
21	4.32	3.47	3.07	2.84	2.68	2 57	2 49	2.42	2 37	2 32	2 25	2 18	2 10	2 05	2 01	1 96	1 92	1 87	1.81
22	4.30	3.44	3 05	2.82	2.66	2 55	2 46	2 40	2 34	2 30	2 23	2 15	2 07	2.03	1 98	1 94	1 89	1 84	1.78
23	4.28	3.42	3.03	2.80	2.64	2 53	2.44	2 37	2 32	2.27	2 20	2 13	2 05	2 01	1 96	1 91	1.86	1 81	1.76
24	4.26	3.40	3.01	2.78	2.62	2 51	2 42	2 36	2 30	2 25	2 18	2 11	2 03	1 98	1 94	1 89	1 84	1.79	1.73
25	4.24	3.39	2.99	2 76	2.60	2 49	2 40	2 34	2.28	2 24	2 16	2.09	2 01	1.96	1.92	1.87	1 82	1.77	1.71
26	4.23	3.37	2.98	2 74	2 59	2 47	2 39	2.32	2 27	2 22	2 15	2 07	1 99	1 95	1 90	1 85	1 80	1 75	1.69
27	4.21	3.35	2 96	2 73	2 57	2.46	2 37	2 31	2 25	2 20	2.13	2.06	1.97	1 93	1 88	1.84	1.79	1.73	1.67
28	4.20	3 34	2.95	2 71	2.56	2 45	2 36	2 29	2 24	2 19	2 12	2 04	1 96	1 91	1 87	1 82	1 77	1 71	1.65
29	4 18	3.33	2.93	2.70	2.55	2 43	2.35	2 28	2 22	2 18	2.10	2 03	1 94	1 90	1.85	1 81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2 33	2.27	2.21	2.16	2 09	2 01	1 93	1 89	1 84	1.79	1.74	1.68	1.62
40	4.08	3 23	2 84	2.61	2 45	2 34	2 25	2.18	2.12	2.08	2 00	1 92	1 84	1 79	1 74	1 69	1 64	1 58	1.51
60	4.00	3 15	2 76	2 53	2 37	2 25	2 17	2.10	2 04	1 99	1 92	1 84	1 75	1 70	1 65	1 59	1.53	1.47	1.39
120	3.92	3 07	2 68	2 45	2 29	2 18	2 09	2 02	1 96	1 91	1 83	1 75	1 66	1 61	1 55	1 50	1 43	1.35	1 25
∞	3.84	3.00	2.60	2 37	2.21	2.10	2.01	1 94	1 88	1 83	1.75	1 67	1 57	1 52	1.46	1.39	1 32	1.22	1.00
	Upper 1 per cent points																		
1	4052	5000	5403	5625	5764	5859	5928	5982	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99 2	99 3	99 3	99 4	99 4	99 4	99 4	99 4	99 4	99 4	99.5	99 5	99 5	99 5	99 5	99.5
3	34.1	30 8	29.5	28 7	28.2	27 9	27 7	27 5	27 3	27.2	27 1	26 9	26 7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16 7	16 0	15 5	15 2	15 0	14 8	14 7	14 5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13 5
5	16.3	13.3	12.1	11.4	11 0	10 7	10 5	10 3	10 2	10 1	9 89	9.72	9.55	9.47	9.38	9 29	9.20	9 11	9.02
6	13.7	10.9	9 78	9.15	8 75	8 47	8 26	8 10	7 98	7 87	7 72	7 56	7.40	7 31	7 23	7.14	7 06	6 97	6.88

7	12.2	9.55	8.45	7.85	7.46	7.19	6 99	6.84	6.72	6.62	6.47	6.31	6.16	6 07	5 99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6 18	6 03	5.91	5.81	5.67	5.52	5 36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8 02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5 11	4.96	4 81	4.73	4 65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4 94	4 85	4.71	4.56	4 41	4.33	4.25	4 17	4 08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5 07	4 89	4.74	4 63	4.54	4.40	4 25	4 10	4 02	3.94	3 86	3 78	3.69	3.60
12	9.33	6.93	5.95	5.41	5 06	4 82	4 64	4 50	4 39	4 30	4.16	4 01	3 86	3 78	3.70	3 62	3 54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4 62	4 44	4.30	4 19	4.10	3 96	3 82	3 66	3.59	3.51	3 43	3 34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4 14	4.03	3 94	3.80	3.66	3 51	3.43	3.35	3 27	3 18	3.09	3.00
15	8.68	6.36	5.42	4 89	4.56	4 32	4 14	4 00	3 89	3 80	3.67	3 52	3.37	3.29	3 21	3.13	3 05	2 96	2.87
16	8.53	6 23	5 29	4 77	4.44	4 20	4.03	3 89	3 78	3 69	3 55	3 41	3 26	3 18	3 10	3 02	2 93	2 84	2.75
17	8.40	6.11	5.19	4 67	4 34	4.10	3 93	3 79	3 68	3 59	3 46	3 31	3 16	3 08	3 00	2 92	2 83	2 75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3 71	3.60	3 51	3 37	3 23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3 77	3 63	3 52	3 43	3 30	3 15	3.00	2.92	2.84	2 76	2 67	2 58	2.49
20	8.10	5 85	4.94	4.43	4 10	3 87	3 70	3 56	3 46	3 37	3.23	3 09	2 94	2.86	2.78	2.69	2 61	2.52	2.42
21	8.02	5 78	4 87	4.37	4.04	3 81	3.64	3 51	3 40	3 31	3 17	3 03	2 88	2 80	2.72	2 64	2 55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3 76	3 59	3.45	3 35	3 26	3.12	2 98	2 83	2 75	2 67	2 58	2.50	2 40	2.31
23	7.88	5.66	4.76	4 26	3.94	3 71	3 54	3 41	3 30	3.21	3.07	2 93	2 78	2.70	2.62	2.54	2 45	2 35	2.26
24	7 82	5.61	4.72	4.22	3.90	3.67	3 50	3 36	3.26	3.17	3.03	2 89	2 74	2.66	2.58	2.49	2.40	2 31	2.21
25	7.77	5.57	4.68	4.18	3.86	3 63	3 46	3.32	3.22	3.13	2.99	2.85	2 70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3 42	3 29	3 18	3 09	2 96	2 82	2 66	2.58	2.50	2 42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3 39	3 26	3 15	3 06	2.93	2.78	2.63	2 55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4 07	3.75	3.53	3.36	3.23	3 12	3 03	2.90	2 75	2.60	2 52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3 33	3.20	3.09	3.00	2.87	2 73	2 57	2 49	2.41	2.33	2.23	2.14	2.03
30	7.56	5 39	4.51	4 02	3.70	3 47	3 30	3 17	3 07	2.98	2 84	2 70	2 55	2.47	2.39	2.30	2 21	2.11	2.01
40	7.31	5.18	4 31	3 83	3.51	3.29	3 12	2 99	2 89	2 80	2 66	2.52	2 37	2.29	2 20	2 11	2 02	1.92	1.80
60	7.08	4.98	4 13	3 65	3.34	3.12	2 95	2.82	2 72	2 63	2.50	2 35	2 20	2 12	2.03	1 94	1 84	1 73	1.60
120	6.85	4.79	3.95	3.48	3.17	2 96	2 79	2 66	2.56	2 47	2.34	2.19	2.03	1 95	1 86	1.76	1 66	1 53	1.38
∞	6.63	4.61	3.78	3 32	3.02	2.80	2 64	2 51	2 41	2 32	2 18	2 04	1.88	1.79	1 70	1.59	1.47	1.32	1.00

Source: By permission of Prof. E. S. Pearson, from M. Merrington and C. M. Thompson, "Tables of Percentage Points of the Inverted Beta (F) Distribution," *Biometrika*, vol. 33, p. 73, 1943.

Table V. Poisson Distribution for Selected Values of m

$$P(y) = e^{-m}m^y/y!$$

y	$P(y)$ for specified values of m									
	$m = 0.1$	$m = 0.2$	$m = 0.3$	$m = 0.4$	$m = 0.5$	$m = 0.6$	$m = 0.7$	$m = 0.8$	$m = 0.9$	$m = 1.0$
0	.9048374	8187308	.7408182	6703200	.606531	.548812	.496585	.449329	406570	.367879
1	.0904837	.1637462	.2222455	2681280	.303265	.329287	347610	.359463	.365913	.367879
2	.0045242	.0163746	0333368	0536256	.075816	.098786	.121663	.143785	.164661	.183940
3	.0001508	.0010916	0033337	0071501	.012636	.019757	.028388	.038343	.049398	061313
4	.0000038	.0000546	.0002500	.0007150	.001580	.002964	004968	.007669	011115	.015328
5	.0000001	.0000022	0000150	.0000572	.000158	.000356	000696	.001227	.002001	.003066
6		.0000001	0000008	0000038	.000013	.000036	.000081	.000164	.000300	.000511
7				0000002	.000001	000003	.000008	.000019	.000039	.000073
8							.000001	.000002	.000004	.000009
9										.000001

y	$m = 2.0$	$m = 3.0$	$m = 4.0$	$m = 5.0$	$m = 6.0$	$m = 7.0$	$m = 8.0$	$m = 9.0$	$m = 10.0$
0	.135335	.049787	.018316	.006738	.002479	.000912	.000335	.000123	.000045
1	.270671	.149361	.073263	.033690	.014873	.006383	.002684	.001111	.000454
2	.270671	.224042	146525	.084224	044618	.022341	.010735	.004998	.002270
3	.180447	.224042	.195367	.140374	.089235	.052129	.028626	.014994	.007567
4	.090224	.168031	.195367	.175467	.133853	.091226	.057252	.033737	.018917
5	.036089	.100819	.156293	.175467	.160623	.127717	.091604	.060727	.037833
6	.012030	.050409	.104196	.146223	.160623	.149003	.122138	.091090	.063055
7	.003437	.021604	.059540	.104445	.137677	.149003	.139587	.117116	.090079
8	000859	008102	.029770	.065278	.103258	.130377	.139587	.131756	.112599
9	.000191	002701	.013231	.036266	.068838	.101405	.124077	.131756	.125110
10	.000038	.000810	.005292	.018133	.041303	.070983	.099262	.118580	.125110
11	.000007	.000221	.001925	.008242	.022529	.045171	.072190	.097020	.113736
12	.000001	.000055	.000642	.003434	.011264	.026350	.048127	.072765	.094780
13		.000013	.000197	.001321	.005199	.014188	.029616	.050376	.072908
14		.000003	.000056	.000472	.002228	.007094	.016924	.032384	.052077
15		.000001	.000015	.000157	.000891	.003311	.009026	.019431	.034718
16			.000004	.000049	.000334	.001448	.004513	.010930	.021699
17			.000001	.000014	.000118	.000596	002124	.005786	.012764
18				.000004	.000039	.000232	000944	.002893	.007091
19				.000001	.000012	.000085	000397	.001370	.003732
20					.000004	.000030	.000159	.000617	.001866
21					.000001	.000010	.000061	.000264	.000889
22						.000003	.000022	.000108	.000404
23						.000001	.000008	000042	.000176
24							.000003	.000016	.000073
25							.000001	.000006	.000029
26								.000002	.000011
27								.000001	.000004
28									.000001
29									.000001

SOURCE: Extracted by permission from E. C. Molina, *Poisson's Exponential Binomial Limit*, Copyright D. Van Nostrand Company, Inc., Princeton, N.J., 1949.

Table VI. Values of $z = \frac{1}{2} \ln \frac{1+r}{1-r}$

(For negative values of r put a minus sign in front of the tabled numbers.)

r	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0 0	00000	.01000	.02000	.03001	.04002	.05004	.06007	07012	08017	.09024
0 1	10034	.11045	.12058	13074	14093	15114	.16139	17167	.18198	.19234
0 2	.20273	.21317	22366	.23419	24477	25541	26611	27686	.28768	.29857
0.3	.30952	.32055	33165	.34283	35409	.36544	.37689	38842	.40006	.41180
0.4	.42365	.43561	.44769	45990	47223	.48470	.49731	.51007	.52298	.53606
0.5	.54931	.56273	57634	.59014	60415	.61838	.63283	64752	.66246	.67767
0.6	.69315	70892	.72500	74142	75817	77530	79281	.81074	.82911	.84795
0.7	.86730	.88718	.90764	92873	.95048	.97295	99621	1.02033	1.04537	1.07143
0.8	1.09861	1.12703	1.15682	1.18813	1.22117	1 25615	1 29334	1 33308	1.37577	1.42192
0.9	1.47222	1.52752	1.58902	1.65839	1 73805	1 83178	1 94591	2.09229	2.29756	2 64665

SOURCE: Extracted by permission from Wilfrid J. Dixon and Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 2d ed., McGraw-Hill Book Company, Inc., New York, 1957.

INDEX

INDEX

Ackoff, Russell L., 273
Additive time-series model, 172
Additivity, in randomized-blocks model, 154
  of sums of squares, 144, 163
Aggregative index, 195
Allocation, optimum, 240
  proportional, 240
Alpha (α), 31, 72, 256
Alternative hypothesis, 31
Analysis of variance, 136–169
  homogeneity assumption in, 149
  Latin squares, 166
  nested classification, 151
  randomized blocks, 151
  simple experiments, 144
  tests of hypotheses in, 144–160
  three-factor experiments, 166
  two-factor experiments, 155
Arithmetic computation, significance in, 37–38
Arithmetic mean, 40
Arnoff, E. Leonard, 273

Bar diagram, 52, 54
Bartlett's test for homogeneity, 151
Beta (β), 31, 72, 256
Binomial distribution, 21–23
  in quality control, 252
  reference to tables of, 101
  for single trial, 21
Binomial distribution, for sum, 22
Binomial test for frequencies, 101
Bivariate normal distribution, 128–129

Cause and effect, 10, 136–138
Central limit theorem, 64
Chi-square (χ^2) distribution, 80–84, 89
  empirical construction of, 88
  table of, 287
Chi-square test, for contingency tables, 109
  for frequencies, 104
  for variance, 81
Churchman, C. West, 273
Class interval, 53, 54
Cluster sampling, 227
Cochran, William G., 102
Coefficient, of correlation, 128–129, 216
    confidence limits on, 132
    multiple, 216
    simple, 128
    t test for significance of, 130
    tests of hypotheses for, 130–132, 217
  of regression (*see* Regression)
Column diagram, 53, 54
Combinations, 17–21
Conditional probability, 16
Confidence coefficient, 68
Confidence limits, 67–70
  on correlation coefficient, 132

Confidence limits, on difference between two means, 93–94
on estimate from line of regression, 127, 221
on mean, 69–70, 76–78
on proportions, 103
on regression coefficient, 126
as substitute for tests of hypothesis, 96–98
on variance, 83
Confounding, 138
Consumer's risk, 259
Consumer's standard, 257
Contingency tables, 108
tests of hypotheses for, 108–110
Continuity correction, 102
Continuous variable, 26, 52
Control charts, 246–254
for fraction defective, 252
for mean, 249
for median, 252
for range, 252
for variance, 249
Correlation, 128–134, 216–218
coefficient of (*see* Coefficient of correlation)
multiple, 216
partial, 223
simple, 115
Cross classification, 106–108
as substitute for regression, 114
Curvilinear regression, 134
Cycle, 171, 193

Decision rule, 8
examples, 31
relation to cost, 32
Deflation of time series, 188
Degrees of freedom, 48
Deviation (*see* Standard deviation)
Discrete variable, 26, 52
Dishonesty in surveys, 11
Dispersion, measure of, 45
Distribution-free tests, 109–111
Dodge, H. F., 261
Double sampling, 260
Dwyer, Paul S., 206

Econometric models, 224
Economic forecasting, 224
Errors, in decisions, 8
Types I and II, 31
experimental, 138
Expected mean squares, 163
Expected values, 40, 160, 271
Experimental data, 136
Experimental error, 138
Experimental material, 139

F distribution, 84–85, 90
empirical construction of, 88
table of, 288
F test, determination of sample size for, 86
of equality of two variances, 85
in multiple correlation, 217
in multiple regression, 214
Finite populations, 63, 228, 243
variance for, 146
Fisher, R. A., 110
Fisher's *z* transformation, 131, 217
and table of, 293
Forecasting, economic, 224
Frame in sampling, 227
Frequency distribution, 52

Gauss multipliers, 218
Gompertz curve, 184
Growth curves, 183

Histogram, 53, 54
Homogeneity assumption in analysis of variance, 149
Hypergeometric distribution, 24
Hypothesis, 29–31, 35, 70
(*See* also Tests of hypotheses)

Independence, 17
Index numbers, 194–196
Inspection in quality control, 245
Interaction, 155
Interval estimation, 68–69
  (*See also* Confidence limits)
Inventory problem, 276

Klein, Lawrence R., 224

Latin squares, 166
Least squares, 118, 120–122
Level of significance, 71
Linearity in regression models, 198
Logistic curve, 184
Lot-size problem, 268

Markoff theorem, 120–122
Matrix algebra, 200–203
Mean, arithmetic, 40
  confidence limits on, 69–70, 76–78
  control charts for, 249
  distribution of, 61
  model for, 40, 42
  properties of, 44, 45
  tests of hypotheses for, 73–80
  unbiasedness of, 161
  variance of, 162
Mean square, 144
  expected, 163
Measure of dispersion, 45
Median, 44
  control charts for, 252
Method of averages in fitting trends, 173
Military Standard 105A, 262
Misuses of statistics, 10–11
Models, additive time-series, 172
  econometric, 224
  general discussion, 266
  for mean, 40, 42
  multiple-regression, 198
  multiplicative time-series, 178
  nested classification, 150
Models, randomized-blocks, 151
  simple analysis of variance, 141
  simple regression, 116
  two-factor randomized-blocks design, 158
  two-factor randomized design, 156
Modified exponential trend, 184
Molina, E. C., 25
Monte Carlo method, 275–284
Morse, Philip M., 272
Moving average, 177
Multinomial distribution, 105*n*.
Multiple regression, 198–224
  coding of variables, 211
  computation from correlation coefficients, 222
  computing forms, 205, 206, 218
  F test in, 214
  meaning of regression coefficients, 212
  model, 198
  normal equations, 199
  prediction by, 209
  standardized regression coefficients, 213
  tests of hypothesis, 214
Multiple sampling, 261
Multiplicative time-series model, 178
Multistage sampling, 234
Mutually exclusive events, 16

Nested classification, 146
Noncentral *t*, 78–79
Nonlinear trend, 181
Nonparametric tests, 110–111
Normal distribution, 56–60
  as approximation to binomial, 101
  bivariate, 128–129
  table of areas, 285
  test, for frequencies, 102
    for means, 73
    for proportions, 103
  tests for correlation coefficients, 131–132
Normal equations, 118–199
Null hypothesis, 30

Observational data, 136
One-tailed test, 75
Operating-characteristic curve, 258
Optimum allocation, 240

Parameter, 3, 40
Partial correlation, 223
Permutations, 17–19
Point estimation, 67
Poisson distribution, 25, 254, 269, 292
Pooled estimate, of mean, 82
  of variance, 82
Population, definition of, 2
  finite, 46, 63, 228, 243
  standard deviation of, 46
Power of test, 32
Prior information, importance in decisions, 97
Probability, 13–28
  a priori, 13
  addition of, 15
  binomial, 21
  conditional, 16
  continuous, 26–68
  definition of, 13
  discrete, 26
  distributions, 20–28
  empirical, 13, 15
Producer's risk, 259
Producer's standard, 257
Proportional allocation, 240

Quality control, 244–263
  risks in, 255
Queuing problems, 264, 272

Rand Corporation, 6
Random numbers, 4–6
Random sampling, 226
Randomization, 139
Ratio estimation, 230
Regions of acceptance and rejection, 32, 73
Regression, curvilinear, 134
  multiple (*see* Multiple regression)
Regression, simple, 115–128
  coefficient of, 119
    confidence limits on, 126
    standard error of, 126, 221
    t test for significance of, 126, 220
    tests of hypotheses for, 125–127, 214–221
  cross classification as substitute, 114
  estimation of, 117–118
  model, 116
Reliability, 45
Replication, 139
Risks, in decision, 7
  in quality control, 255
Romig, H. G., 261

Sample, definition of, 2, 4
  systematic, 7
Sample size, determination of, for F test, 86
  for mean, variance known, 95
    variance unknown, 96
Sampling, 225–243
  double, 260
  frame in, 227
  multistage, 234–243
    equal probabilities without replacement, 235
    unequal probabilities with replacement, 237
  random, 226
    simple, 228
  ratio estimation in, 230–233
  sequential, 86, 261
  stratified, 236, 239
  three-stage, 242
Sampling tables, 261, 262
Scientific notation, 38
Seasonals, 171–193
  absolute units, 175
  method of ratios, to moving average, 192
  to trend, 187
Sequential sampling, 86, 261
Set, 13–15

Setup costs, 268
Siegel, Sidney, 110
Significance in arithmetic computation, 37–38
Significance tests (*see* Tests of hypothesis)
Spurious accuracy, 10
Square-root method, 206
Standard deviation, 45–57
  bias in, 48
  computation from grouped data, 55
  estimate of, 47
  of population, 46
Standard error, of difference between two means, 91
  of estimate, 126, 216
  of mean, 60–64
  of regression coefficient, 126, 221
  of z, 131
Statistic, definition of, 3
Statistical hypothesis, 29
Statistical proof, 30
Statistics, misuses of, 10–11
  as numerical facts, 2
Stein, Charles, 96
Stratified random sampling, 236, 239
Stratifying factors, 240
Suboptimization, 265
Summation notation, 41, 42
Systematic sample, 7

t, noncentral, 78–79
t distribution, 76, 89
  empirical construction of, 87
  table of, 286
t test, for difference between two means, 90–93
    dependent samples, 91–93
    independent samples, 90–91
  for mean, 76
  for significance, of correlation coefficients, 130
    of regression coefficient, 126, 220
Tchebycheff's inequality, 57
Tests of hypotheses, in analysis of variance, 144–160
Tests of hypotheses, confidence limits as substitute for, 96–98
  for contingency tables, 108–110
  for correlation coefficients, 130–132, 217
  for difference between two means, 90–93
    dependent samples, 91, 93
    independent samples, 90, 91
  for frequencies and proportions, 100–106
  for mean, 73–80
  nonparametric, 110–111
  for regression coefficients, 125–127, 214–221
  for two variances, 84–86
  for variance, 80–83
Three-stage sampling, 242
Trend, 171–187
  method, of averages, 173
    of least squares, 180
  nonlinear, 181
Tukey, J. W., 154
Two-tailed test, 74

Unbiasedness, of mean, 161
  of variance, 162
Universe, 2

Variable, 2
  continuous, 26, 52
  discrete, 26, 52
Variance, 47
  analysis of (*see* Analysis of variance)
  computing formula, 51
  control charts for, 249
  for finite population, 46
  grouped data, 55
  for infinite populations, 47
  of mean, 162
  tests of hypotheses for, 80–83
  unbiasedness of, 162

z transformation, 131, 217, 293