INTRODUCTION

TO THE

THEORY OF STATISTICS

BY

ALEXANDER McFARLANE MOOD

The RAND Corporation
Formerly Professor of Statistics, Iowa State College

FIRST EDITION

New York Toronto London

McGRAW-HILL BOOK COMPANY, INC.

1950

INTRODUCTION TO THE THEORY OF STATISTICS

THE MAPLE PRESS COMPANY, YORK, PA.

To

HARRIET

PREFACE

This book developed from a set of notes which I prepared in 1945. At that time there was no modern text available specifically designed for beginning students of mathematical statistics. Since then the situation has been relieved considerably, and had I known in advance what books were in the making it is likely that I should not have embarked on this volume. However, it seemed sufficiently different from other presentations to give prospective teachers and students a useful alternative choice.

The afore-mentioned notes were used as text material for three years at Iowa State College in a course offered to senior and first-year graduate students. The only prerequisite for the course was one year of calculus, and this requirement indicates the level of the book. (The calculus class at Iowa State met four hours per week and included good coverage of Taylor series, partial differentiation, and multiple integration.) No previous knowledge of statistics is assumed.

This is a statistics book, not a mathematics book, as any mathematician will readily see. Little mathematical rigor is to be found in the derivations simply because it would be boring and largely a waste of time at this level. Of course rigorous thinking is quite essential to good statistics, and I have been at some pains to make a show of rigor and to instill an appreciation for rigor by pointing out various pitfalls of loose arguments.

While this text is primarily concerned with the theory of statistics, full cognizance has been taken of those students who fear that a moment may be wasted in mathematical frivolity. All new subjects are supplied with a little scenery from practical affairs, and, more important, a serious effort has been made in the problems to illustrate the variety of ways in which the theory may be applied.

The problems are an essential part of the book. They range from simple numerical examples to theorems needed in subsequent chapters. They include important subjects which could easily take precedence over material in the text; the relegation of subjects to problems was based rather on the feasibility of such a procedure than on the priority of the subject. For example, the matter of correlation is dealt with almost entirely in the problems. It seemed to me inefficient to cover

multivariate situations twice in detail, i.e., with the regression model and with the correlation model. The emphasis in the text proper is on the more general regression model.

The author of a textbook is indebted to practically everyone who has touched the field, and I here bow to all statisticians. However, in giving credit to contributors one must draw the line somewhere, and I have simplified matters by drawing it very high; only the most eminent contributors are mentioned in the book.

My greatest personal debt is to S. S. Wilks, who kindled my interest in statistics and who was my mentor throughout my term of graduate study. Any merits which this book may have must be charged largely to his careful lectures and understanding direction of my studies.

My colleagues at Iowa State College have all contributed much to my understanding and general view of statistics. I am particularly aware of large debts to G. W. Brown, W. G. Cochran, and G. W. Snedecor. Among the many students who thoroughly revised the original notes by their excellent comments and suggestions I must mention H. D. Block, who gave the final manuscript a very careful and competent review. Margaret Kirwin and Ruth Burns accurately translated my scrawl into beautiful typescript. Bernice Brown and Miss Burns carefully proofread the entire set of galleys.

I am indebted to Catherine Thompson and Maxine Merrington, and to E. S. Pearson, editor of *Biometrika*, for permission to include Tables III and V, which are abridged versions of tables published in *Biometrika*. I am also indebted to Professors R. A. Fisher and Frank Yates, and to Messrs. Oliver and Boyd, Ltd., Edinburgh, for permission to reprint Table IV from their book "Statistical Tables for Use in Biological, Agricultural and Medical Research."

In the final chapter are some distribution-free tests which were developed jointly by G. W. Brown and myself at Iowa State College on a project sponsored by the Office of Naval Research. Professor Brown has very generously and graciously permitted me to include this material which should have first appeared in print under his name as well as mine. The tests referred to are presented in Sections 5, 6, 7, 8, and 9 of Chapter 16.

ALEXANDER MCFARLANE MOOD

SANTA MONICA, Calif.
January, 1950

CONTENTS

CHAPTER 10. SAMPLING DISTRIBUTIONS

CHAPTER 11. INTERVAL ESTIMATION

CHAPTER 12. TESTS OF HYPOTHESES

CHAPTER 13. REGRESSION AND LINEAR HYPOTHESES

CONTENTS

CHAPTER 1

INTRODUCTION

1.1. Statistics. In order to place this book in its proper perspective, it is necessary to consider first what statistics is. The lay conception of statistics ordinarily includes the collection of large masses of data and the presentation of such data in tables or charts; it may also include the calculation of totals, averages, percentages, and the like. In any case this conception is about thirty years out of date; these more or less routine operations are only an incidental part of statistics today.

We shall describe statistics as the technology of the scientific method. Statistics provides tools and techniques for research workers. These tools may be of quite general application and useful in any field of science—physical, biological, or social. On the other hand certain tools may be particularly designed for special fields of research.

We shall not embark on a discussion of the scientific method here, but we may recall its three main aspects: (1) the performance of experiments, (2) the drawing of objective conclusions from experiments, and (3) the construction of laws to simplify the description of the conclusions of large classes of experiments. Statistics is primarily concerned with the first two of these aspects; in fact, the field of statistics is commonly thought of as being divided into the two areas corresponding to these two aspects: (1) the design of experiments and investigations, (2) statistical inference. We shall continue our description of statistics by discussing these areas briefly in the following two sections.

1.2. The Design of Experiments and Investigations. An experiment is meant to study the effect of variation of certain factors or the relation between certain factors. Thus one may wish to study the relation between temperature and pressure in a fixed volume of a gas. Or one may wish to discover what if any effect on milk production results from altering the proportion of roughage in a cow's diet. Again one may wish to study the effect on the retail price of a certain commodity when a given public policy regarding the commodity is promulgated.

In the typical experiment the research worker is harassed by addi-

tional factors which influence the outcome of the experiment, factors which he would like to eliminate but cannot control completely. These extraneous factors are least important in the physical sciences, where the experimenter has good control over his experimental material. They are quite important in the biological sciences, where the geneticist must deal with animals each having its own peculiar genetic inheritance, the plant breeder must deal with whatever varieties happen to be available, do his experiments in whatever soil is at hand and in whatever weather conditions may occur. The extraneous factors become most troublesome in the social sciences, where the research worker frequently has no control at all over his experimental material. Studies in these sciences are often investigations rather than experiments.

Statistics is concerned with these extraneous factors—with designing the experiment so as to eliminate them if possible or to minimize their effects, with arranging the experiment in space or time so that the effects may be expected to cancel or partially cancel themselves, with designing the experiment so that the effects may be removed or partially removed in the analysis of the resulting data. The design may be nothing more than an obvious application of common sense. Thus suppose batches of the same material from several different sources are to be analyzed in order to determine whether they are sufficiently alike to be treated the same way in some manufacturing process. A number of specimens chosen at random from each batch are to be analyzed; two men are to do the individual analyses. It is plain that the specimens from each batch should be divided equally between the two analysts, else variations due to differences in the analysts' techniques will appear in the final results as differences between batches. Experimental designs range from such trivial devices as this to highly elaborate arrangements based on the mathematical theory of finite geometries.

In designing investigations, the problem is normally one of balancing extraneous factors by selecting representative samples. Thus suppose a political party, in order to judge how actively it should campaign in a given state, employs a public-opinion-polling agency to estimate the proportions of voters in the state who intend to vote for its candidate and the rival candidate. The polling agency will do this by interviewing a sample of voters in the state. It is clear that the factor in which the agency is interested (proportions of voters favoring the two candidates) will be widely influenced by a great many other factors in which it is not directly interested. For example, farmers as a group

and laborers as a group may feel quite differently about the candidates. The agency must control this factor by making the proportions of people in various occupational groups in the sample equal to those proportions in the state. It should make the proportions of people in various racial groups in the sample equal those for the state. The proportions of people at different economic levels should be the same. The proportions of people in different geographical areas should be the same. And so on. The sample should, in short, be as representative as possible of the population of the state. The statistician is concerned with ways of selecting such samples or, if this is impossible or impracticable, with ways of assessing the magnitudes of the effects of such extraneous factors and removing them in the final analysis of the results.

1.3. Statistical Inference. New knowledge in science is usually found by a logically hazardous process—the process of generalizing from particular results. The scientist, on perceiving a certain pattern in the results of one or more experiments, conjectures that the pattern may be characteristic of a large class of possible experiments. The conjecture or hypothesis would ordinarily be tested by performing other experiments; it might be further supported or it might be disproved. The latter outcome is by no means infrequent, for generalization or inductive thinking is well known to lead to uncertain conclusions.

The broad problem of statistical inference is to provide measures of the uncertainty of conclusions drawn from experimental data. This problem is attacked by means of the theory of probability, which forms the foundation of the theory of statistical inference. The tools of statistical inference enable the scientist to assess the reliability of his conclusions in terms of probability statements. To consider a simple example: Suppose a chemist has made three precise determinations of the atomic weight of chlorine, and suppose his results are 35.4563, 35.4578, 35.4575. He might conclude, for example, that the true atomic weight is between 35.456 and 35.458. It is the function of statistical inference to tell the chemist to what extent he may rely on this conclusion. The measure of reliability might be given by a statement of this form: "The odds are two to one that the conclusion is correct." If it is important that the chemist estimate the atomic weight within .002, he will likely be dissatisfied with such low odds and will make further determinations in order to decrease his chances of being wrong. He might, for example, feel that for his purposes he must be very confident of his conclusion and repeat his determinations

until there is only one chance in a hundred of his final conclusion being in error.

It is usually impossible to make an entirely valid generalization to arrive at a certain conclusion on the basis of experimental evidence. But it is possible to measure the uncertainty of such conclusions in probability terms and thus resolve to a considerable degree a very troublesome problem faced by every scientist.

The scope of statistical inference is as broad as experimentation itself. An experiment may be intended merely to evaluate a constant, as in the illustration just given, or it may be meant to evaluate parameters in a function, or perhaps to estimate a function itself, or a set of functions. An experiment may be designed to test a certain hypothesis suggested by a tentative theory—the hypothesis that two factors are unrelated, that a relation has a specified functional form. The experimenter may have to contend with relatively small effects from extraneous factors, as in the physical sciences, or with quite large ones, as in the social sciences. In any case the problem of statistical inference arises. If an experiment indicates that a certain hypothesis is false, the hypothesis may nevertheless remain tenable in the experimenter's mind if that conclusion is not supported by heavy odds. The certainty of a conclusion is often as important as the conclusion itself in the final evaluation of an experiment.

1.4. The Theory and Practice of Statistics. Another division of the field of statistics worth brief consideration is that between the theory and the methodology.

The theory of statistics is a branch of applied mathematics. It has its roots in an area of pure mathematics known as the theory of probability, and in fact the complete structure of statistical theory in a broad sense may be thought of as including the theory of probability. And it includes other things not part of the formal theory of probability—theoretical consequences of the principle of randomization, various principles of estimation, and principles of testing hypotheses. These principles may be regarded as axioms which augment the axioms of probability theory.

The statistician is, of course, engaged in producing tools for research workers. Faced with a particular experimental problem, he constructs a mathematical model to fit the experimental situation as best he can, analyzes the model by mathematical methods, and finally devises procedures for dealing with the problem. He is guided in this work by the principles of the theory of statistics.

The statistician is also engaged in developing and extending the

theory of statistics. There are many quite important problems of experimental design and statistical inference which remain untouched because the theory of statistics is not yet powerful enough to deal with them. The broad advance in the application of statistical methods during the past two decades was made possible by far-reaching developments in the theory which immediately preceded it.

It may be interesting to remark here on the origins of the theory of statistics. Certain areas of biological experimentation reached a point where what are now called statistical methods were imperative if further progress was to be made. The essentials of statistical theory were then evolved by the biologists themselves. This parallels the natural history of almost any branch of abstract knowledge, but it is nevertheless curious in the case of statistics. For the theory of statistics appears to be a very natural development of the theory of probability, which is several hundred years old; somehow it was almost completely overlooked by workers in that field. Incidentally the situation which created statistical theory still obtains; there are many areas of scientific experimentation ready and waiting for statistical methods which do not yet exist.

In contradistinction to the theory of statistics is the practice of statistics. There is a great body of tools and techniques for research workers which expands appreciably with the passing of each year. Until recent years the statistician was not much concerned with these tools, being content to pass them on to those who wished to use them. But as scientific research progresses experiments become more complex and the statistical tools become correspondingly complex and specialized. In some areas the time has come when it is impossible for the research worker to become familiar with all the tools that might be useful to him. Furthermore, as tools become more specialized, they become less flexible; to fit a particular experiment the tool often has to be modified, and this requires knowledge of statistical theory.

The use of statistical tools is not merely a matter of picking out the wrench that fits the bolt; it is more a matter of selecting the correct one of several wrenches which appear to fit the bolt about equally well but none of which fit it exactly. It is a long step from an algebraic formula to, for example, a nutrition experiment on hogs. There is nothing magic about the formula; it is merely a tool, and moreover a tool derived from some simple mathematical model which cannot possibly represent the actual situation with any great precision. In using the tool one must make a whole series of judgments relative to the nature and magnitude of the various errors engendered by the dis-

crepancies between the model and the actual experiment. These judgments cannot well be made by either the statistician or the experimenter, for they depend both on the nature of statistical theory and the nature of the experimental material.

To meet this development, the applied statistician has come on the scene. He is to be found in various industrial and academic research centers, and his function is, of course, to collaborate with the research workers in their experimentation and investigation. He must be completely familiar with both the theory and methodology of statistics even though his work is concerned not with the field of statistics at all but with the field of application. We merely wish to observe here that applied statistics has developed to the point where it may be regarded as a field of interest in itself.

1.5. The Scope of This Book. This book is concerned with the theory rather than the applications of statistics. In the course of the development many tools will be derived and discussed; a secondary purpose of the book is to make clear the conditions under which certain of the important statistical tools may be employed. But our primary purpose is the exposition of statistical theory.

The book is introductory in that no knowledge of statistics by the reader is presumed. And it is elementary in that no knowledge of mathematics beyond elementary calculus is presumed. This restriction of the mathematical level is necessarily costly. We shall have to omit entirely many interesting but more technical developments of the theory; the generality of theorems will be reduced; it will be necessary to make statements without proof from time to time; mathematical rigor will be sacrificed at many points; and cumbersome arguments will sometimes have to be used when very simple arguments at a higher mathematical level exist. All these sacrifices, however, will inhibit our presentation rather less than one might suppose. The essential aspects of the theory are entirely comprehensible without higher mathematics.

Since statistical theory is founded on probability theory, we shall begin the study with a consideration of probability concepts and the development of certain probability theorems which will be required. Next we shall consider mathematical models which have been found by experience to approximate many common experimental situations. It will then be possible to study mathematically the problems of statistical inference and of the design and analysis of experiments and investigations.

1.6. Reference System. The chapters are divided into numbered sections; the numbering begins anew in each chapter. In referring to a section contained in the same chapter as the reference, only the section number is given. In referring to a section in a different chapter, the chapter number is prefixed to the section number and separated from it by a period. Thus Sec. 5.3 refers to Sec. 3 of Chap. 5.

The equations are numbered anew in each section, and equation numbers are always enclosed in parentheses. Merely the equation number is given when referring to an equation in the same section as the reference; otherwise the section number is prefixed. Thus equation (4.6) refers to the sixth equation of the fourth section of the same chapter as the reference, and equation (9.1.12) refers to the twelfth equation of the first section of the ninth chapter.

CHAPTER 2

PROBABILITY AND COMBINATORIAL METHODS

2.1. Definition of Probability. Probability is a measure of the likelihood of occurrence of a chance event. A precise definition can be given in many ways, but for our immediate purposes, the following statement, known as the classical definition of probability, will suffice:

If an event can occur in N mutually exclusive and equally likely ways, and if n of these outcomes have an attribute A, then the probability of A is the fraction n/N.

We shall apply this definition to a few simple examples in order to illustrate its meaning.

If an ordinary die (one of a pair of dice) is tossed, there are six possible outcomes: any one of the six numbered faces may turn up. These six outcomes are mutually exclusive since two or more faces cannot turn up simultaneously. And, supposing the die to be *fair* or *true*, the six outcomes are equally likely; no one face is any more to be expected than another. Now suppose we want the probability that the result of a toss be an even number. Three of the six possible outcomes have that attribute. The probability that an even number will appear when a die is tossed is therefore $3/6$ or $1/2$. Similarly, the probability that a five will appear when a die is tossed is $1/6$. The probability that the result of a toss will be greater than two is $2/3$.

To consider another example, suppose a card is drawn at random from an ordinary deck of playing cards. The probability of drawing a spade is readily seen to be $13/52$ or $1/4$. The probability of drawing a number between five and ten inclusive is $24/52$ or $6/13$.

The application of the definition is straightforward enough in these simple cases, but it is not always so obvious. Careful attention must be paid to the qualifications "mutually exclusive" and "equally likely." Suppose one wished to compute the probability of getting two heads if a coin were tossed twice. He might reason that there were three possible outcomes for the two tosses: two heads, two tails, or one head and one tail. One of these outcomes has the desired attribute; therefore the probability is $1/3$. This reasoning is faulty because the three given outcomes are not equally likely. The third

outcome can occur in two ways since the head may appear on the first toss and the tail on the second, or the head may appear on the second toss and the tail on the first. Thus there are four equally likely outcomes: HH, HT, TH, TT. The first of these has the desired attribute while the others do not. The correct probability is therefore ¼. The result would be the same if two coins were tossed simultaneously.

Again suppose one wished to compute the probability that a card drawn from an ordinary deck will be an ace or a spade. In enumerating the favorable outcomes he might count 4 aces and 13 spades, and reason that there are 17 possible outcomes with the desired attribute. This is clearly incorrect because the events are not mutually exclusive. The occurrence of an ace does not preclude the occurrence of a spade.

We note that a probability is always a number between zero and one. The ratio n/N must be a proper fraction since the total number of possible outcomes cannot be smaller than the number of outcomes with a specified attribute. If an event is certain to happen, its probability is one; while if it is certain not to happen, its probability is zero. Thus, the probability of obtaining an eight in tossing a die is zero. The probability that the outcome of tossing a die will be less than ten is one.

The probabilities determined by the classical definition are called a priori probabilities. When one states that the probability of obtaining a head in tossing a coin is one-half, he has arrived at this result purely by deductive reasoning. The result does not require that any coin be tossed, or even be at hand. We say that if the coin is true, the probability of a head is one-half, but this is little more than saying the same thing in two different ways. Nothing is said about how one can determine whether or not a particular coin is true.

The fact that we shall deal with ideal objects in developing the theory of probability will not trouble us, because that is a common requirement of mathematical systems. Geometry, for example, deals with conceptual perfect circles, lines with zero width, and so forth, but it is a useful branch of knowledge which can be applied to diverse practical problems.

There are some rather troublesome defects in the classical, or a priori, approach. It is obvious, for example, that the definition of probability must be modified somehow when the total number of possible outcomes is infinite. One might seek, for example, the probability that a positive integer drawn at random be even. The intuitive answer to this question is ½. If one were pressed to justify this result on the basis of the definition, he might reason as follows: Suppose we limit our-

selves to the first 20 integers; 10 of these are even so that the ratio of favorable events to the total number is $^{10}/_{20}$ or $\frac{1}{2}$. Again, if the first 200 integers are considered, 100 of these are even, and the ratio is also $\frac{1}{2}$. In general, the first $2N$ integers contain N even integers; if we form the ratio $N/2N$ and let N become infinite so as to encompass the whole set of positive integers, the ratio remains $\frac{1}{2}$.

The above argument is plausible and the answer is plausible, but it is no simple matter to make the argument stand up. It depends, for example, on the natural ordering of the positive integers, and a different ordering could produce a different result. Thus, one could just as well order the integers in this way: 1, 3, 2; 5, 7, 4; 9, 11, 6; . . . , taking the first pair of odd integers, then the first even integer; the second pair of odd integers, then the second even integer; and so forth. With this ordering, one could argue that the probability of drawing an even integer is $\frac{1}{3}$. The integers can also be ordered so that the ratio n/N will oscillate back and forth and never approach any definite value as N increases.

There is another difficulty with the classical approach to the theory of probability which is deeper even than that arising in the case of an infinite number of outcomes. Suppose we have a coin known to be biased in favor of heads (it is loaded so that a head is more likely to appear than a tail). The two possible outcomes of tossing the coin are not equally likely. What is the probability of a head? The classical definition leaves us completely helpless here.

In a situation like the above we shall simply assume that there does exist some definite though unknown number which gives the desired probability. And we shall assume that the number obeys the same laws as the probabilities arising from the classical definition.

We have pointed out these difficulties merely to indicate the limitations of our approach. A complete discussion of these points belongs properly in a textbook on the theory of probability. There are other methods of defining probabilities which are logically more satisfactory than the one we have chosen, but ours has the advantage of simplicity. And as yet there is no general agreement among writers on the theory of probability as to what is the most satisfactory set of axioms for the theory.

2.2. Permutations and Combinations. The evaluation of a priori probabilities requires the enumeration of all possible outcomes of a given chance event. This sort of enumeration can often be facilitated by certain combinatorial formulas which will be developed now. They are based on the following two basic principles:

(a) *If an event A can occur in a total of m ways and if a different event B can occur in n ways, then the event A or B can occur in $m + n$ ways provided A and B cannot occur simultaneously.*

(b) *If an event A can occur in a total of m ways and if a different event B can occur in n ways, then the event A and B can occur in mn ways.*

These two ideas may be illustrated by letting A correspond to the drawing of a spade from a deck of cards and B correspond to the drawing of a heart. Each of these events can be done in 13 ways. The number of ways in which a heart or a spade can be drawn is obviously $13 + 13 = 26$. To illustrate the second principle, suppose two cards are drawn from the deck in such a way that one is a spade and the other is a heart. There are $13 \times 13 = 169$ ways of doing this, since with the ace of spades we may put any one of the 13 hearts, or with the king of spades we may put any one of the 13 hearts, and so on for all 13 of the spades.

The two principles may clearly be generalized to take account of more than two events. Thus, if three mutually exclusive events A B, and C can occur in m, n, and p ways, respectively, then the event A or B or C can occur in $m + n + p$ ways, and the event A and B and C can occur in mnp ways.

We shall now use the second of these principles to enumerate the number of arrangements of a set of objects. Let us consider the number of arrangements of the letters a, b, c. We can pick any one of the three to place in the first position; either of the remaining two may be put in the second position, and the third position must be filled by the unused letter. The filling of the first position is an event which can occur in three ways; the filling of the second position is an event which can occur in two ways, and the third event can occur in one way. The three events can occur together in $3 \times 2 \times 1 = 6$ ways. The six arrangements, or *permutations*, as they are called, are

$$abc,\ acb,\ bac,\ bca,\ cab,\ cba$$

In this simple example the elaborate method of counting was hardly worth while because it is easy enough to write down all the six permutations. But if we had asked for the number of permutations of six letters, we should have had

$$6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

permutations to write down.

It is obvious now that in general the number of permutations of n

different objects is

$$n(n-1)(n-2)(n-3)\cdots(2)(1)$$

The row of dots indicates omission of intermediate factors. This product of an integer by all the positive integers smaller than it, is usually denoted more briefly by $n!$ (read n factorial). Thus $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$, etc. Since

$$(n-1)! = \frac{n!}{n}$$

it is common to define $0!$ as one, so that the relation will be consistent when $n = 1$.

Let us now enumerate the number of permutations that may be made from n objects if only r of the objects are used in any given permutation. Reasoning as before, the first position may be filled in n ways, the second position may be filled in $n - 1$ ways, and so forth. When we come to the rth position, we will have used $r - 1$ of the objects so that $n - (r - 1)$ will remain from which we can choose. The number of permutations of n objects taken r at a time is therefore $n(n-1)(n-2)\cdots(n-r+1)$. The symbol $P_{n,r}$ is used to denote this number.

$$P_{n,r} = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} \qquad (1)$$

Thus the number of permutations of the four letters a, b, c, d taken two at a time is $P_{4,2} = 4 \times 3 = 12$. On putting $r = n$ in equation (1), we get the result stated earlier: that the number of permutations of n objects taken n at a time is $n!$.

With the aid of equation (1) we can now solve the following problem: In how many different ways can r objects be selected from n objects? $P_{n,r}$ counts all the possible selections as well as all the arrangements of each selection or *combination*. Two combinations are different if they are not made up of the same set of objects. Thus abc and abd are different three-letter combinations, while abc and bac are different permutations of the same combination. Let the symbol $\binom{n}{r}$ denote the number of different combinations. Then it is clear that $P_{n,r}$ equals $\binom{n}{r}$ times $r!$, since each combination of r objects has $r!$ arrangements. Therefore

$$\binom{n}{r} = \frac{P_{n,r}}{r!} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} \qquad (2)$$

Another common symbol for this number is $C_{n,r}$, but we shall not use it in this text. The number of combinations of five objects taken three at a time is

$$\binom{5}{3} = \frac{5 \times 4 \times 3}{3!} = \frac{60}{6} = 10$$

The number $\binom{n}{r}$ may be given a different interpretation. It is the number of ways in which n objects may be divided into two groups, one group containing r objects, and the other group containing the other $n - r$ objects. Now suppose we wish to divide n objects into three groups containing n_1, n_2, n_3 objects, respectively, with

$$n_1 + n_2 + n_3 = n$$

We shall first divide them into two groups containing n_1 and $n_2 + n_3$ objects. This may be done in $\binom{n}{n_1}$ ways. Then we may divide the second group into two groups containing n_2 and n_3 objects. This may be done in $\binom{n_2 + n_3}{n_2}$ ways. Using the second principle of enumeration, the total number of ways of doing the two divisions together is

$$\binom{n}{n_1}\binom{n_2 + n_3}{n_2} = \frac{n!}{n_1!(n_2 + n_3)!}\,\frac{(n_2 + n_3)!}{n_2!n_3!} = \frac{n!}{n_1!n_2!n_3!}$$

This type of argument may be carried further to find the number of ways of dividing n objects into k groups containing $n_1, n_2, \cdots, n_k$ objects with $n_1 + n_2 + \cdots + n_k = n$. This number is readily found to be

$$\frac{n!}{n_1!n_2! \cdots n_k!} \tag{3}$$

Thus the number of ways of dividing four objects into three groups containing 1, 1, and 2 objects is

$$\frac{4!}{1!1!2!} = 12$$

The expression (3) also has a second interpretation. It is the number of different permutations of n objects when n_1 of the objects are alike and of one kind, n_2 are alike and of a second kind, and so forth. Referring to the numerical example above, there are 12 permutations of the letters a, b, c, c. In order to see that expression (3) gives the correct number, consider n different objects (for example, the letters

$a, b, c, \cdots, p$) arranged in a definite order. And consider a division of this set of objects into k groups, the first group containing n_1 objects, the second n_2, and so forth. Now in the original arrangement of objects, replace all the objects selected for the first group by ones, all those selected for the second group by twos, and so forth. The result will be a permutation of n_1 ones, n_2 twos, $\cdots$, n_k k's. A little reflection will convince one that every division of the letters into the k groups corresponds to a different permutation of the integers, and that this is the total set of permutations, because if there were another, there would be another division of the letters into k groups.

We have derived three formulas in this section, not only because they are useful but because their derivation serves to illustrate the application of the two principles of enumeration given at the beginning of the section. It is the methods that are important. The formulas will aid in solving many problems, but they are useless in many others, and one must then fall back on the elementary principles.

Illustrative example: If two cards are drawn from an ordinary deck, what is the probability that one will be a spade and the other a heart?

Since nothing is said about the order in which the spade and the heart should occur, this is a problem in combinations. To compute the probability, we must find the total number of possible outcomes of two-card draws, and then find the number of these that have the specified attribute. The total number of two-card combinations that can be made up from 52 cards is $\binom{52}{2} = 1326$. And we have seen before that there are $13 \times 13 = 169$ different combinations with the required attribute. The probability is therefore $^{169}/_{1326} = {}^{13}/_{102}$.

This problem could also be solved by regarding the different two-card permutations as the set of possible outcomes. The denominator of the ratio would then be $P_{52,2} = 2652$. To get the numerator, we consider that each of the 169 two-card combinations has two permutations and get $2 \times 169 = 338$ as the number of permutations with the required attribute. Or we may start at the beginning as follows: The number of permutations in which the spade occurs first and the heart second is $13 \times 13 = 169$ by principle (b). And the number with the heart first and the spade second is the same. Either of these sets of permutations satisfies the specification. By principle (a) the required number is $169 + 169 = 338$. Again we find the probability is $^{13}/_{102}$.

Illustrative example: What is the probability that of four cards drawn from an ordinary deck, at least three will be spades?

Here again we are interested in combinations. The total number

of possible four-card combinations is $\binom{52}{4} = 270{,}725$. To get the numerator: the specification, at least three spades, means either three or four. The number of four-card hands containing exactly three spades is $\binom{13}{3}\,39 = 11{,}154$; the first factor is the number of three-card combinations of three spades, and the second is the number of ways a card may be selected from the other three suits; the product is taken in accordance with principle (*b*). The number of hands with all cards spades is $\binom{13}{4} = 715$. By principle (*a*), the number of hands with the required attribute is $11{,}154 + 715 = 11{,}869$. The required probability is 11,869/270,725.

One might attempt to find the numerator by the following method: The number of three-card combinations of spades is $\binom{13}{3} = 286$. The fourth card may be either a spade or not a spade, and after three spades have been selected, the fourth card may be selected from the whole set of 49 remaining cards. Thus the required number of hands is $49 \times 286 = 14{,}014$. This argument is faulty because the hands with four spades have been counted more than once. A specific three-card combination of spades is AKQ, and when the jack of spades is drawn from the remaining 49 cards, we have the combination AKQJ. But we also count this combination when the AQJ is considered and the king is drawn from the remaining 49 cards. It is now clear that the hands with four spades have been counted four times in the above figure. We can obtain the correct result by subtracting from it three times the number of hands with four spades. The result is

$$14{,}014 - 3\binom{13}{4} = 11{,}869$$

as before.

Illustrative example: Seven balls are tossed into four numbered boxes so that each ball falls in a box and is equally likely to fall in any of the boxes. What is the probability that the first box will contain two balls?

Since the first ball may fall in any one of four ways, the second may fall in any one of four ways, and so forth, the total number of possible outcomes is, by principle (*b*), 4^7. To enumerate the number of outcomes with the desired attribute, let us first divide the seven balls into two groups, one containing two and the other five balls. This

may be done in $\binom{7}{2}$ ways. Now the group of two will be put into the first box and the other five distributed among the other three boxes. This may be done, by the same reasoning as above, in 3^5 ways. The number of favorable outcomes is therefore $\binom{7}{2}3^5$, and the desired probability is

$$\frac{\binom{7}{2}3^5}{4^7} = \frac{7 \times 3^6}{4^7} \cong .3115$$

(The symbol $\cong$ is used to denote approximate equality.)

2.3. Stirling's Formula. In finding numerical values of probabilities, one is often confronted with the evaluation of long factorial expressions which are troublesome to compute by direct multiplication. If an adding machine is available, and there are not a great number of factors in the expression, it is often convenient to use logarithms. However, when the factors become numerous, this method also becomes tedious, and much labor may be saved by using Stirling's formula, which gives an approximate value of $n!$. It is

$$n! \cong \sqrt{2\pi}\, e^{-n} n^{n+\frac{1}{2}} \tag{1}$$

where e is the Napierian base, $2.71828 \cdots$. A much more accurate approximation may be obtained by replacing the factor e^{-n} by $e^{-[n-(1/12n)]}$, but this refinement is rarely used. To indicate the accuracy of the formula, we may compute $10!$, which is actually 3,628,800. Formula (1) using five-place logarithms gives

$$10! \cong 3{,}599{,}000$$

The more refined formula gives:

$$10! \cong 3{,}629{,}000$$

The error in (1) for $n = 10$ is a little less than 1 per cent, and the percentage error decreases as n increases.

2.4. Sum and Product Notations. A sum of terms such as $n_3 + n_4 + n_5 + n_6 + n_7$ is often designated by the symbol $\sum_{i=3}^{7} n_i$. The Σ is the capital Greek letter sigma, and in this connection it is often called the *summation sign*. The letter i is called the *summation index*.

The term following the Σ is called the *summand*. The $i = 3$ below Σ indicates that the first term of the sum is obtained by putting $i = 3$ in the summand. The 7 above the Σ indicates that the final term of the sum is obtained by putting $i = 7$ in the summand. The other terms of the sum are obtained by giving i the integral values between the limits 3 and 7. Thus

$$\sum_{j=2}^{5} (-1)^{j-2} j x^{2j} = 2x^4 - 3x^6 + 4x^8 - 5x^{10}$$

An analogous notation is obtained by substituting the capital Greek letter Π for Σ. In this case the terms resulting from substituting the integers for the index are multiplied instead of added. Thus

$$\prod_{a=1}^{5} \left[c + (-1)^a \frac{a}{b}\right] = \left(c - \frac{1}{b}\right)\left(c + \frac{2}{b}\right)\left(c - \frac{3}{b}\right)\left(c + \frac{4}{b}\right)\left(c - \frac{5}{b}\right)$$

Using this notation, expression (2.3) derived previously may be written $n! \Big/ \prod_{i=1}^{k} n_i!$.

2.5. The Binomial and Multinomial Theorems. The expansion of the binomial expression $(x + y)^n$ is given in elementary algebra courses, and a proof of the correctness of the expansion is ordinarily by induction. We shall here expand the binomial by a simple combinatorial method which readily generalizes to the multinomial case. If we write the binomial in the form $(x + y)(x + y)(x + y) \cdots (x + y)$, which has n factors, the problem of finding the coefficient of one of the terms, say $x^{n-a}y^a$, reduces to the problem of finding the number of ways of dividing the n factors into two groups. The first term of the expansion is x^n, which is obtained by selecting the x from each of the factors. The next term is some coefficient times $x^{n-1}y$. This term arises by selecting the x from $n - 1$ of the factors and the y from the remaining one. The one from which y is taken may be chosen in any of n ways; hence the coefficient of $x^{n-1}y$ is n. In general, to get the coefficient of $x^{n-a}y^a$, we must count the number of ways of dividing the n factors into two groups so that one group contains a factors and the other $n - a$ factors; y is selected from each factor of the first group and x from each factor of the second group. The number of ways of dividing the n factors into two such groups is $\binom{n}{a}$, which is the desired coeffi-

cient. The binomial expansion is therefore

$$(x + y)^n = x^n + nx^{n-1}y + \binom{n}{2} x^{n-1}y^2 + \cdots + y^n$$

$$= \sum_{i=0}^{n} \binom{n}{i} x^{n-i}y^i \qquad ($$

The multinomial theorem follows directly. If the expression

$$(x_1 + x_2 + \cdots + x_k)^n$$

is multiplied out, one will obtain terms of the form

$$Cx_1^{n_1}x_2^{n_2} \cdots x_k^{n_k}$$

where C is some coefficient and the exponents satisfy the relation

$$\sum_{i=1}^{k} n_i = n$$

We wish to determine C. Terms of the given form arise when x_1 i selected from n_1 of the n factors, x_2 is selected from n_2 of the remainin factors, and so forth. The number of ways of getting such a term i equal to the number of ways of dividing the n factors into k group containing $n_1, n_2, \cdots, n_k$ factors. This is expression (3) of Sec. 2 Thus the general term of the multinomial expansion is

$$\frac{n!}{n_1!n_2! \cdots n_k!} x_1^{n_1}x_2^{n_2} \cdots x_k^{n_k} \qquad \text{or} \qquad n! \prod_{i=1}^{k} \frac{x_i^{n_i}}{n_i!}$$

and we may write

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{n_1,n_2,\cdots,n_k} n! \prod_{i=1}^{k} \frac{x_i^{n_i}}{n_i!} \qquad (2)$$

We have indicated only that the summation is over the indices $n_1, n_2,$ $\cdots, n_k$. The range of each index is zero to n, but they cannot all be summed independently over that range because we must have $\sum_{i=1}^{k} n_i = n$. The summation is over all sets of values of $n_1, n_2, \cdots,$ n_k such that their sum is n and such that each n_i is an integer in the range zero to n inclusive. The sum is very troublesome to write down when n is large. We shall illustrate it for a simple case.

$$(x_1 + x_2 + x_3)^4 = \sum_{n_1,n_2,n_3} \frac{4!}{n_1!n_2!n_3!} x_1^{n_1}x_2^{n_2}x_3^{n_3}$$

The sets of values of (n_1, n_2, n_3) which satisfy $n_1 + n_2 + n_3 = 4$ are (4, 0, 0), (3, 1, 0), (3, 0, 1), (2, 2, 0), (2, 1, 1), (2, 0, 2), (1, 3, 0), (1, 2, 1), (1, 1, 2), (1, 0, 3), (0, 4, 0), (0, 3, 1), (0, 2, 2), (0, 1, 3), (0, 0, 4). The sum therefore has 15 terms, the first few of which are

$$(x_1 + x_2 + x_3)^4 = \frac{4!}{4!}x_1^4 + \frac{4!}{3!}x_1^3x_1 + \frac{4!}{3!}x_1^3x_3 + \frac{4!}{2!2!}x_1^2x_2^2 + \cdots + \frac{4!}{4!}x_3^4$$
$$= x_1^4 + 4x_1^3x_2 + 4x_1^3x_3 + 6x_1^2x_2^2 + \cdots + x_3^4$$

A set of numbers such as (3, 1, 0) is called a three-part *partition* of four. (2, 6) is a two-part partition of eight. The 15 triplets of numbers listed above form the complete set of *ordered three-part partitions* of four. The partitions are called *ordered* because the same combination of three parts in a different order is counted as a different partition. If it is not specified that the partitions be ordered, the unordered ones are assumed; thus, the three-part partitions of four are simply (4, 0, 0), (3, 1, 0), (2, 2, 0), (2, 1, 1). In terms of the idea of partitions, the multinomial sum (2) may be described briefly as follows: the sum is taken over all ordered k-part partitions of n, the parts being $(n_1, n_2, \cdots, n_k)$.

2.6. Combinatorial Generating Functions. The enumeration of possible outcomes and of outcomes with a certain attribute can become quite a complex problem. In fact, it is easy to state problems in which the enumeration is practically impossible. One of the most powerful devices for solving enumeration problems involves the use of what are called *generating functions*. The subject of combinatorial generating functions is a field of mathematics in itself, and we shall consider only a few simple cases here. We wish merely to indicate the nature of this method of analysis.

Let us consider the last illustration given in Sec. 2 where seven balls were tossed into four boxes, and consider the function

$$(x_1 + x_2 + x_3 + x_4)^7$$

The coefficient of a term such as $x_1^2x_2^4x_3$ in the expansion of this multinomial is given by formula (2.3) as $7!/2!4!1!0!$, which is just the number of ways of dividing seven objects into four groups so that the first contains two objects, the second four, and so forth. So any term in the multinomial expansion gives a description of a possible outcome; a factor such as x_i^5 means five balls have fallen in the ith box, and the numerical coefficient of the term gives the number of ways in which

that outcome can occur. If the x's are now all replaced by ones, the terms become simply $7!\Big/\prod_{i=1}^{4} n_i!$, and to get the whole set of possible outcomes, we need to sum this expression over all sets of the n_i whose sum is seven. This sum by the multinomial theorem is just

$$(1 + 1 + 1 + 1)^7 = 4^7$$

If we want the probability that the first box contains two balls, we shall sum $7!/\Pi n_i!$ over all sets of n_i which have $n_1 = 2$. Let us rewrite the term as

$$\frac{7!}{2!5!}\frac{5!}{n_2!n_3!n_4!}$$

and now we wish to sum this over all sets such that $n_2 + n_3 + n_4 = 5$. If we multiply $5!/n_2!n_3!n_4!$ by $1^{n_2}1^{n_3}1^{n_4}$, we have the general term of $(1 + 1 + 1)^5$; hence the desired sum is $7!/2!5!$ times 3^5.

The function $(x_1 + x_2 + x_3 + x_4)^7$ is a simple type of generating function; it is an algebraic expression which is given an interpretation in terms of the physical problem at hand. It may be used to answer any of the questions that may be asked about the physical problem to which it is related. Thus, if the number of ways in which the first two boxes can each contain at least two balls is required, we would add the coefficients of all terms in the generating function which have x_1 and x_2 with powers greater than or equal to two.

Now let us consider another problem. An urn contains five black and four white balls. The balls are all drawn one by one from the urn, and the first three drawn are placed in a black box while the last six are placed in a white box. What is the probability that the number of black balls in the black box plus the number of white balls in the white box is equal to five?

We may solve this problem by considering the balls of each color to be numbered. The total number of ways of dividing the nine objects into two groups, the first containing three and the second six, is $\binom{9}{3}$. To get five balls to match the color of the box containing them, we must clearly have two black balls in the black box and three white ones in the white box. The black box may be filled $\binom{5}{2}\binom{4}{1}$ ways since there are $\binom{5}{2}$ ways of picking two black ones from the five black

ones to be among the first three drawn, and $\binom{4}{1}$ ways of choosing one white ball to be among the first three drawn. The probability is $\binom{5}{2}\binom{4}{1}\Big/\binom{9}{3}$.

The following generating function may be related to this problem:

$$(x_1t + x_2)^5(x_1 + x_2t)^4$$

Here x_1 corresponds to the black box and x_2 to the white one. The first factor corresponds to the five black balls, and the second to the four white balls. We shall consider the coefficient of the term involving $x_1^3x_2^6$. It will be a polynomial in t, and if t were put equal to one, the polynomial would have the value $\binom{9}{3}$, since then we should have the coefficient of $x_1^3x_2^6$ in $(x_1 + x_2)^9$. The coefficient of t^r in the polynomial is the number of ways in which r balls can fall in boxes of the same color as the balls. In forming a term in $x_1^3x_2^6$, we may choose certain of the x_1's from the factor $(x_1t + x_2)^5$ and the remainder from the other factor. Those chosen from the first factor represent black balls, and those chosen from the second represent white balls. Thus, when a black ball is associated with the black box, we get a factor t, and when a white ball is associated with the white box, we also get a factor t. The power of t then gives the total number of times a ball is associated with a box of its color. On expanding the generating function, one would find the coefficient of $x_1^3x_2^6t^5$ to be $\binom{5}{2}\binom{4}{1}$ as before.

The generating function is of no value for this simple problem, but it becomes useful if more than two colors are considered. Thus suppose an urn contained n_1 balls of a given color, n_2 of a second color, and n_3 of a third color; and suppose m_1 are drawn and placed in a box of the first color, m_2 are then drawn and placed in a box of the second color, and the remaining balls, say m_3 of them, are placed in a box of the third color. Let n be the total number of balls; then

$$n = n_1 + n_2 + n_3 = m_1 + m_2 + m_3$$

The coefficient of $x_1^{m_1}x_2^{m_2}x_3^{m_3}t^r$ in the function

$$(x_1t + x_2 + x_3)^{n_1}(x_1 + x_2t + x_3)^{n_2}(x_1 + x_2 + x_3t)^{n_3}$$

gives the number of ways in which r balls match color of the box containing them. The coefficient is difficult to calculate in this case, but

to find it is a straightforward procedure, while to find it without the generating function is considerably more troublesome.

We shall consider one other kind of generating function. If five dice are tossed, what is the probability that the sum of the spots will be 15?

Since the first die may fall in six ways, the second may fall in six ways, and so forth, the total number of possible outcomes is 6^5. Now we need the number of these outcomes that have a sum equal to 15. In the case of two dice, it is easy to write down all possible combinations which give a specified sum. Thus to obtain a sum of five, the two dice may fall (1, 4), (2, 3), (3, 2), (4, 1). These are the ordered two-part partitions of five when zero is excluded as a part. In our problem we must enumerate all the ordered five-part partitions of 15 which have all parts between one and six inclusive.

In problems involving partitions of numbers, there is a generating function which will usually materially simplify the enumeration. For the particular problem of counting the ways of getting 15 with five dice, let us consider this function:

$$(x + x^2 + x^3 + x^4 + x^5 + x^6)^5 \tag{1}$$

It is a polynomial in x in which the term of lowest degree is x^5 and the term of highest degree is x^{30}. Let us suppose that the function is written as the product of five factors instead of as a fifth power. The first factor will be associated with the first die, the second factor with second die, and so on. In the expansion of the function there will be a number of terms x^{15}; one, for example, will arise when x is selected from each of the first three factors and x^6 is selected from the remaining two factors. This situation corresponds to the appearance of a one on the first three dice, and a six on the other two. It is readily seen that there is a one-to-one correspondence between the ways x^{15} can arise in the expansion and the ways the five dice can total 15. Hence our required number is the coefficient of x^{15} in the expansion of the function. This coefficient may be found most easily by use of the following identity:

$$\frac{1 - x^n}{1 - x} = 1 + x + x^2 + \cdots + x^{n-1} \tag{2}$$

which may be verified by multiplying both sides by $1 - x$. Using this identity, the generating function may be put in the form:

$$\frac{x^5(1 - x^6)^5}{(1 - x)^5}$$

We may omit the factor x^5 and find the coefficient of x^{10} in what remains. Now we need another identity:

$$\frac{1}{(1-x)^n} = 1 + \binom{n}{1}x + \binom{n+1}{2}x^2 + \binom{n+2}{3}x^3 + \cdots \qquad |x| < 1$$

$$= \sum_{i=0}^{\infty} \binom{n+i-1}{i} x^i \tag{3}$$

which reduces our problem to that of finding the coefficient of x^{10} in

$$(1 - x^6)^5 \sum_{i=0}^{\infty} \binom{4+i}{i} x^i$$

If the first factor is expanded, all but the first two terms have x to a higher power than 10 and may be neglected. And now the problem becomes that of finding the coefficient of x^{10} in

$$(1 - 5x^6) \sum_{i=0}^{\infty} \binom{4+i}{i} x^i$$

which has two terms in x^{10}: one when the 1 is multiplied by the term given by $i = 10$ in the sum, and the other when the $-5x^6$ is multiplied by the term given by $i = 4$ in the sum. The coefficient is therefore $\binom{14}{10} - 5\binom{8}{4}$, and the probability we set out to find is

$$\frac{\binom{14}{10} - 5\binom{8}{4}}{6^5} = \frac{651}{7776}$$

$$\cong .0837$$

These examples will serve to indicate the kind of attack that may be made on enumeration problems by means of generating functions. The method is powerful, but we cannot develop it here. We merely wish to point out the existence of the method.

2.7. Marginal and Conditional Probability. Suppose that there are n equally likely possible outcomes of a chance event, and that they may be classified according to two criteria. Thus the event may be the selection of a ball from an urn in which all the balls are colored and all are numbered; the possible outcomes may be classified according to color, or according to number. In general, suppose there is an A classification with r classes which we denote by $A_1, A_2, \cdots, A_r$, and

a B classification with s classes denoted by $B_1, B_2, \cdots, B_s$. The n outcomes may then be classified in a two-way table as follows:

	B_1	B_2	$\cdots$	B_s
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_r	n_{r1}	n_{r2}		n_{rs}

Here we have indicated that n_{11} of the n outcomes have both the attribute A_1 and the attribute B_1; n_{12} have both the attribute A_1 and the attribute B_2; and in general n_{ij} of the outcomes have the attributes A_i and B_j. The sum of all n_{ij} is n. As an example we may consider the drawing of a card from an ordinary deck of playing cards. The 52 outcomes may be classified according to suit (say A_1, A_2, A_3, A_4), according to denomination (say $B_1, B_2, \cdots, B_{13}$). In this example every n_{ij} is one.

The probability that the event will have a given specification, A_1 and B_3, for example, will be denoted by $P(A_1, B_3)$, and the value of this probability is obviously n_{13}/n. In general,

$$P(A_i, B_j) = \frac{n_{ij}}{n}$$

We may be interested in only one of the criteria of classification, say A, and indifferent to the B classification. In this case B is omitted from the symbol, and the probability of A_2, say, is written $P(A_2)$, and

$$P(A_2) = \frac{n_{21} + n_{22} + n_{23} + \cdots + n_{2s}}{n}$$

$$= \sum_{j=1}^{s} \frac{n_{2j}}{n}$$

This is called a *marginal probability*, and the term marginal is used whenever one or more criteria of classification are ignored. It is clear that

$$P(A_i) = \sum_{j=1}^{s} \frac{n_{ij}}{n}$$

or

$$P(A_i) = \sum_{j=1}^{s} P(A_i, B_j) \tag{1}$$

since $n_{ij}/n = P(A_i, B_j)$. Also the marginal probability of B_j is

$$P(B_j) = \sum_{i=1}^{r} P(A_i, B_j) \tag{2}$$

Thus the probability that a chance event has a specified attribute is the sum of all the probabilities of events that have that attribute. The probability that a card be an ace is the sum of the probabilities that it be the ace of spades, the ace of hearts, the ace of diamonds, and the ace of clubs.

In a more general situation, suppose there are three criteria of classification, A, B, and C. Let n_{ijk} of the n possible outcomes have the specification A_i, B_j, C_k; and let the C classification be $C_1, C_2, \cdots, C_t$, with the A and B classes the same as before. The complete classification would be a three-way table consisting of t layers of two-way tables, each layer corresponding to a C_k. The marginal probability of, say, A_i and C_k is

$$P(A_i, C_k) = \sum_{j=1}^{s} P(A_i, B_j, C_k) \tag{3}$$

and the marginal probability of C_k is

$$P(C_k) = \sum_{i=1}^{r} \sum_{j=1}^{s} P(A_i, B_j, C_k) \tag{4}$$

$$= \sum_{i=1}^{r} P(A_i, C_k) \tag{5}$$

$$= \sum_{j=1}^{s} P(B_j, C_k) \tag{6}$$

The extension of these ideas to more than three criteria of classification is apparent.

Returning to the original two-way classification, suppose the outcome of a chance event is examined for one attribute but not for the other. We wish to find the probability that the other attribute has a specified value. The event, for example, may be observed to have the attribute B_3. What is the probability that it also has the attribute A_2? The total number of outcomes for A given that B_3 has occurred, is

$\sum_{i=1} n_{i3}$, and the number of favorable outcomes for A_2 are n_{23}. Thu

the probability of A_2, given that B_3 has occurred, is $n_{23} \Big/ \sum_{i=1} n_{i3}$

This is called a *conditional probability* and is denoted by the symbo $P(A_2|B_3)$. In general

$$P(A_i|B_j) = \frac{n_{ij}}{\sum_{i=1}^{r} n_{ij}}$$

$$P(B_j|A_i) = \frac{n_{ij}}{\sum_{j=1}^{s} n_{ij}}$$

On dividing both the numerator and denominator of the fraction on the right by n, we have

$$P(A_i|B_j) = \frac{P(A_i, B_j)}{P(B_j)} \tag{7}$$

$$P(B_j|A_i) = \frac{P(A_i, B_j)}{P(A_i)} \tag{8}$$

or in another form

$$P(A_i, B_j) = P(A_i|B_j)P(B_j) \tag{9}$$

$$= P(B_j|A_i)P(A_i) \tag{10}$$

The last equation may be stated: the probability that an outcome will have the attribute A_i and B_j is equal to the marginal probability of A_i multiplied by the conditional probability of B_j given that A_i has occurred.

The idea of conditional probability has a straightforward extension to situations involving more than two criteria of classification. In the case of three criteria, for example, it may be shown directly that

$$P(A_i, B_j|C_k) = \frac{P(A_i, B_j, C_k)}{P(C_k)} \tag{11}$$

$$P(A_i|B_j, C_k) = \frac{P(A_i, B_j, C_k)}{P(B_j, C_k)} \tag{12}$$

also that

$$P(A_i, B_j, C_k) = P(A_i, B_j|C_k)P(C_k) \tag{13}$$

$$= P(A_i|B_j, C_k)P(B_j, C_k) \tag{14}$$

$$= P(A_i|B_j, C_k)P(B_j|C_k)P(C_k) \tag{15}$$

and other similar relations could be obtained by permuting the letters A, B, C. Thus

$$P(B_j|A_i, C_k) = \frac{P(A_i, B_j, C_k)}{P(A_i, C_k)} \tag{16}$$

and

$$P(A_i, B_j, C_k) = P(B_j|A_i, C_k)P(A_i|C_k)P(C_k) \tag{17}$$

or

$$P(A_i, B_j, C_k) = P(B_j|A_i, C_k)P(C_k|A_i)P(A_i) \tag{18}$$

We shall not take the space to write out all such possible relations, but the student would do well to do so. These relations are fundamental in the theory of statistics and must be well understood.

In defining conditional probability we have used a rather specialized model. But it is apparent that the idea is quite general. Let X be any subset of the whole set of possible outcomes, and let Y be any subset of X; then

$$P(Y|X) = \frac{P(Y)}{P(X)}$$

for if N is the total number of outcomes, n is the number in X, and m is the number in Y, then $P(Y|X) = m/n$, $P(Y) = m/N$, and

$$P(X) = \frac{n}{N}$$

2.8. Two Basic Laws of Probability. The two laws correspond to the two principles of enumeration discussed in Sec. 2. The additive law of probability states that

If A and B are mutually exclusive subsets of the whole set of possible outcomes of a chance event, then the probability that the event occurs in A or B is equal to the probability that it occurs in A plus the probability that it occurs in B.

Symbolically, we may write this as

$$P(A \text{ or } B) = P(A) + P(B) \tag{1}$$

This law follows directly from principle (a) of Sec. 2. In general, if $A_1, A_2, \cdots, A_h$ are mutually exclusive subsets of the whole set of outcomes, then

$$P(A_1 \text{ or } A_2 \text{ or } A_3 \cdots \text{ or } A_h) = \sum_{i=1}^{h} P(A_i) \tag{2}$$

The marginal probability defined by (7.1) is a special case of this relation. The specification A_i is fulfilled by the subsets A_i, B_1; A_i, B_2;

$\cdots$; A_i, B_s; hence

$$P(A_i) = P(A_i, B_1 \text{ or } A_i, B_2 \cdots \text{ or } A_i, B_s)$$
$$= \sum_{j=1}^{s} P(A_i, B_j)$$

If the two subsets A and B of (1) are not mutually exclusive, then (1) is no longer true. In this case, certain outcomes have both the attribute A and the attribute B. We may interpret this in terms of the two-way classification given at the beginning of Sec. 7. Suppose we want the probability that the outcome is in A_1 or B_2. A_1 consists of the first row of the table and B_2 consists of the second column. The outcomes in A_1B_2 satisfy both specifications, and thus the two sets A_1 and B_2 are not mutually exclusive. The probability that the outcome falls in A_1 or B_2 is easily calculated by adding all n_{ij} in the first row and second column and dividing by n.

$$P(A_1 \text{ or } B_2) = \frac{\sum_1^s n_{1j} + \sum_2^r n_{i2}}{n}$$
$$= \frac{\sum_1^s n_{1j} + \sum_1^r n_{i2} - n_{12}}{n}$$
$$= P(A_1) + P(B_2) - P(A_1, B_2) \tag{3}$$

This gives us a more general law of addition of probabilities.

If A and B are subsets of the set of outcomes of a chance event, the probability that the event occurs in A or B is equal to the probability that it occurs in A plus the probability it occurs in B minus the probability that it occurs in both A and B.

The situation is illustrated in Fig. 1, where the outcomes of a chance event are represented by points in a plane and two subsets are enclosed by two circles A and B. Certain outcomes fall in the lenticular region common to both circles, and in adding the outcomes in both circles, these points are counted twice and must therefore be subtracted once. Symbolically, the additive law is

$$P(A \text{ or } B) = P(A) + P(B) - P(A, B) \tag{4}$$

We may generalize this law to account for more than two subsets; thus

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

as is easily verified by drawing a figure similar to Fig. 1 in which three circles intersect so as to have a region common to all three. The general law for h subsets, which may be proved by induction, is

$$P(A_1 \text{ or } A_2 \cdots \text{ or } A_h) = \sum_{i=1}^{h} P(A_i) - \sum_{i,j} P(A_i, A_j) + \sum_{i,j,k} P(A_i, A_j, A_k) - \cdots \pm P(A_1, A_2, \cdots, A_h) \quad (6)$$

where the second sum is over all combinations of the numbers 1, 2, $\cdots$, h taken two at a time, the third is over all combinations of the

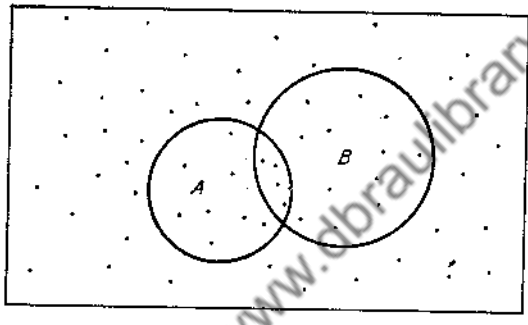


FIG. 1.

numbers taken three at a time, and so forth. If all the subsets are mutually exclusive, then all the probabilities in the sums beyond the first sum are zero, and (6) reduces to (2).

We have essentially derived the multiplicative law of probabilities in defining conditional probability in the preceding section.

If some of the outcomes of a chance event can have both the attributes A and B, the probability of such an occurrence is equal to the probability of A multiplied by the conditional probability of B given that A has occurred, or it is equal to the probability of B multiplied by the conditional probability of A given that B has occurred.

In symbols,

$$P(A, B) = P(A)P(B|A) \quad (7)$$

$$= P(B)P(A|B) \quad (8)$$

We may refer to the model given in preceding section; or we may use the model of Fig. 1. Let n be the number of points in Fig. 1; let m_1 be the number of points in A (including those common to B), m_2 be the number in B, and m_3 be the number common to A and B. Then

$$P(A, B) = \frac{m_3}{n}$$

$$P(A) = \frac{m_1}{n}$$

$$P(B) = \frac{m_2}{n}$$

$$P(A|B) = \frac{m_3}{m_2}$$

$$P(B|A) = \frac{m_3}{m_1}$$

whence (7) and (8) follow directly.

In general we may show by induction that

$$P(A_1, A_2, \cdots, A_h) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)P(A_4|A_1, A_2, A_3) \cdots P(A_h|A_1, A_2, \cdots, A_{h-1}) \quad (9)$$

and there are $h!$ such relations which may be obtained by permuting the letters in the right-hand side of (9). The two relations for $h = 2$ are given by (7) and (8).

2.9. Compound Events. The multiplicative law of probabilities is particularly useful in simplifying the computation of probabilities for compound events. A compound event is one that consists of two or more single events as when a die is tossed twice, or three cards are drawn one at a time from a deck. The following simple example will illustrate the method:

Two balls are drawn, one at a time, from an urn containing two black, three white, and four red balls. What is the probability that the first is red and the second is white? (The first is not replaced before the second is drawn.)

The outcomes of this compound event may be classified according to two criteria: the color of the first ball, and the color of the second ball. We may therefore construct a table like that at the beginning of Sec. 7. The A classification corresponds to the color of the first ball, and we shall let A_1, A_2, A_3 correspond to the colors black, white, and red, respectively. Similarly the classes B_1, B_2, B_3 will correspond to the same colors for the second ball. The total number of outcomes is $n = 9 \times 8 = 72$. It is not $\binom{9}{2} = 36$, because we are considering permutations, not arrangements; i.e., we are not asking that one ball be red and one white; we require that the colors appear in a specific

order. The complete table of outcomes is

	B_1	B_2	B_3
A_1	2	6	8
A_2	6	6	12
A_3	8	12	12

and the probability asked for in the problem is

$$P(A_3, B_2) = 12/72 = 1/6$$

By using the multiplicative law of probabilities, we need only consider the two separate events one at a time. Here we must use the law in the form

$$P(A_3, B_2) = P(A_3)P(B_2|A_3)$$

Now $P(A_3)$ is simply the probability of drawing a red ball in a single draw, which is $4/9$, and $P(B_2|A_3)$ is the probability of drawing a white one, given that a red one has already been drawn, which is $3/8$. The product of these two numbers gives the required probability

$$P(A_3, B_2) = 4/9 \times 3/8 = 1/6$$

The validity of the above technique is not obvious. It is not immediately evident that the marginal probability $P(A_3)$ can be computed by completely disregarding the second event, nor that the conditional probability corresponds to the simple physical event described above.

For a compound event consisting of two single events we need only consider a 2×2 table. Let A_1 correspond to a success on the first event, and A_2 correspond to a failure, and let m_1 be the number of ways the first event can succeed, and m_2 be the number of ways it can fail. Let B_1 and B_2 be similarly defined for the second event. Let m_{11} and m_{12} be the numbers of ways the second event can succeed and fail if the first succeeds, and let m_{21} and m_{22} be the number of ways the second can succeed or fail if the first event fails. The 2×2 table is

	B_1	B_2
A_1	m_1m_{11}	m_1m_{12}
A_2	m_2m_{21}	m_2m_{22}

The total number of possible outcomes is

$$n = m_1m_{11} + m_1m_{12} + m_2m_{21} + m_2m_{22}$$

The required probability is

$$P(A_1, B_1) = \frac{m_1m_{11}}{n} \tag{1}$$

The marginal probability $P(A_1)$ is

$$\frac{m_1m_{11}}{n} + \frac{m_1m_{12}}{n} = \frac{m_1(m_{11} + m_{12})}{m_1(m_{11} + m_{12}) + m_2(m_{21} + m_{22})} \tag{2}$$

Now the probability of a success on the first event without regard to the second is simply $m_1/(m_1 + m_2)$, which is not equal to the above expression unless

$$m_{11} + m_{12} = m_{21} + m_{22}$$

i.e., unless the total number of outcomes for the second event is the same regardless of whether or not the first event is a success. The conditional probability is $m_{11}/(m_{11} + m_{12})$ and gives the probability of a success for the second event under the assumption that the first was a success.

We might be inclined to conclude that the conditional-probability approach is correct only if the number of outcomes for the second event is independent of the outcome of the first event. Precisely the opposite is true. The correct probability is

$$P(A_1, B_1) = \frac{m_1}{m_1 + m_2} \frac{m_{11}}{m_{11} + m_{12}} \tag{3}$$

and not the value m_1m_{11}/n given in equation (1).

The value computed by the conditional approach is always correct, while that computed by enumeration of outcomes is correct only if the number of outcomes for the second event is independent of the outcome of the first event.

A simple example will clarify the situation. Suppose a coin is tossed, and if a head appears, a black ball is placed in an urn, while if a tail appears, a black ball and a white ball are placed in the urn. Then a ball is drawn from the urn. If a head is tossed, the ball will necessarily be black. Using H, T, B, W to represent heads, tails, black, and white, the three possible outcomes of this sequence are HB, TB, TW. These three outcomes are clearly not equally likely. If the experiment

were repeated a number of times, we should expect the outcome HB to occur twice as often as either of the other two. $P(\text{HB}) = \frac{1}{2}$, not $\frac{1}{3}$.

In general, the possible outcomes of a compound event are not equally likely if the number of outcomes of the second event depends on the outcome of the first; hence the definition of probability is not applicable. However, if the definition can be applied to the constituent events separately, then it is possible to compute the probability of the compound event by using the method of conditional probabilities. Unfortunately, it is not possible to give a formal proof of these statements. We must simply rely on our intuition, or rather on the import of whatever experimental evidence we may possess. Such evidence may be obtained, for example, by performing the above-described experiment a number of times.

Illustrative example: To illustrate further the method of conditional probabilities, let us compute the probability that of five cards drawn from an ordinary deck, exactly two will be aces.

We shall suppose the deck consists of four A's, representing aces, and 48 N's, representing not aces. To use conditional probabilities, we must assume the five cards are drawn one at a time, and we must assume a particular order such as A, A, N, N, N. We shall use equation (8.9) with $h = 5$.

$$P(A, A, N, N, N) = P(A)P(A|A)P(N|A, A)P(N|A, A, N)P(N|A, A, N, N)$$

Now $P(A) = \frac{4}{52}$; with one ace removed from the deck, $P(A|A) = \frac{3}{51}$; with two aces removed from the deck, $P(N|A, A) = \frac{48}{50}$. Proceeding thus,

$$P(A, A, N, N, N) = \tfrac{4}{52} \times \tfrac{3}{51} \times \tfrac{48}{50} \times \tfrac{47}{49} \times \tfrac{46}{48}$$

This is the probability for the given order, but the problem did not specify any order, so we must consider all possible orders. There are $5!/(2!3!) = 10$ permutations of two A's and three N's, so we have 10 probabilities to evaluate, and the required probability, by the additive law, is the sum of these 10 probabilities. It is soon apparent, however, that all the probabilities are equal. Thus, for example,

$$P(N, A, N, N, A) = \tfrac{48}{52} \times \tfrac{4}{51} \times \tfrac{47}{50} \times \tfrac{46}{49} \times \tfrac{3}{48}$$

which is the same as the above number except that the numerators are permuted. Clearly this will be the case for all permutations. Hence

the required probability is

$$10P(A, A, N, N, N) = \frac{10 \times 4 \times 3 \times 47 \times 46}{52 \times 51 \times 50 \times 49}$$
$$\cong .0399$$

Independent Events. If the conditional probability $P(B|A)$ is equal to the marginal probability $P(B)$, the events A and B are said to be *independent.* The outcome of B is not influenced in any way by A. Thus a die may be tossed twice, and we may seek the probability that the results will be two and three in that order

$$P(2, 3) = P(2)P(3|2) = P(2)P(3) = \tfrac{1}{6} \times \tfrac{1}{6}$$

In the illustrative example involving two aces in five cards, the five constituent events of the compound event will be independent if we require that each card drawn be replaced in the deck and the deck shuffled before the next card is drawn. The probability that the second card will be an ace is then $\frac{4}{52}$ instead of $\frac{3}{51}$. The probability that two aces will appear when five cards are drawn with replacement is

$$10(\tfrac{4}{52})^2(\tfrac{48}{52})^3 \cong .0465$$

In general,

If the constituent events of a compound event are mutually independent, the probability of the compound event is equal to the product of the probabilities of the constituent events.

We may write this in the form

$$P(A_1, A_2, \cdots, A_k) = \prod_{i=1}^{k} P(A_i) \tag{4}$$

provided that

$$P(A_i) = P(A_i|A_j \cdots A_p) \qquad \text{for all } i, j, \cdots, p$$

It is important to remember that this probability is the probability of occurrence of the separate events in a specific order.

The additive law of probability given by equation (8.6) can also be used to simplify materially certain problems in compound events. A striking example is provided in the following:

Illustrative example: Six cards are drawn with replacement from an ordinary deck. What is the probability that each of the four suits will be represented at least once among the six cards?

We shall solve the problem by finding first the probability that all the suits do not appear. Let A symbolize the appearance of all the suits, and B symbolize the nonappearance of at least one of the suits.

Since either A or B is certain to happen,

$$P(A \text{ or } B) = 1$$

and since A and B are mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) = 1$$

and

$$P(A) = 1 - P(B)$$

Thus, if we can find $P(B)$, $P(A)$ can be determined at once.

To get $P(B)$, let us classify the possible outcomes favorable to B into four sets: B_1 is the set of all outcomes in which spades are absent; B_2 is set for which hearts are absent; B_3, diamonds absent; B_4, clubs absent. These sets are overlapping; an outcome which consists of only spades and hearts falls in B_3 and in B_4. Clearly

$$P(B) = P(B_1 \text{ or } B_2 \text{ or } B_3 \text{ or } B_4)$$

and employing equation (8.6)

$$P(B) = \Sigma P(B_i) - \Sigma P(B_i, B_j) + \Sigma P(B_i, B_j, B_k) - P(B_1, B_2, B_3, B_4)$$

in which the sums are taken over all combinations of the subscripts. The probability $P(B_1)$ that a spade will not appear in the six draws is $(\frac{3}{4})^6$, and the value is the same for all B_i; hence

$$\Sigma P(B_i) = 4(\tfrac{3}{4})^6$$

The probability $P(B_1, B_2)$ that neither spades nor hearts will appear in the six draws is $(\frac{1}{2})^6$ and is the same for all six pairs of the four suits taken two at a time; hence

$$\Sigma P(B_i, B_j) = 6(\tfrac{1}{2})^6$$

Similarly

$$\Sigma P(B_i, B_j, B_k) = 4(\tfrac{1}{4})^6$$

and

$$P(B_1, B_2, B_3, B_4) = 0$$

since the simultaneous nonappearance of every suit is impossible. The required probability is, therefore,

$$P(A) = 1 - 4(\tfrac{3}{4})^6 + 6(\tfrac{1}{2})^6 - 4(\tfrac{1}{4})^6$$
$$\cong .381$$

A slight alteration of this example will illustrate another useful technique.

Illustrative example: Cards are drawn one at a time with replacement from an ordinary deck until all suits have appeared at least once. What is the probability that six draws will be required?

Referring to the preceding example, let P_n denote the probability that all suits will be represented at least once if n cards are drawn. Clearly

$$P_n = 1 - 4(\tfrac{3}{4})^n + 6(\tfrac{1}{2})^n - 4(\tfrac{1}{4})^n$$

Now suppose we knew the answer to the present problem for a general value of n. Let p_n denote this probability (that exactly n draws will be required to produce all the suits).

If n cards are drawn, the first appearance of each suit at least once may occur on the fourth draw, or the fifth, or the sixth, and so forth. Since these outcomes are mutually exclusive, we have

$$P_n = p_4 + p_5 + p_6 + \cdots + p_n$$

From this relation we conclude that

$$p_n = P_n - P_{n-1}$$

and in particular that

$$\begin{aligned} p_6 &= 1 - 4(\tfrac{3}{4})^6 + 6(\tfrac{1}{2})^6 - 4(\tfrac{1}{4})^6 - [1 - 4(\tfrac{3}{4})^5 + 6(\tfrac{1}{2})^5 - 4(\tfrac{1}{4})^5] \\ &= (\tfrac{3}{4})^5 - 3(\tfrac{1}{2})^5 + 3(\tfrac{1}{4})^5 \\ &\cong .147 \end{aligned}$$

2.10. A Priori and Empirical Probabilities. In introducing the theory of probability we have relied heavily on the combinatorial definition given in the first section of the chapter. However, we have seen that this approach has severe limitations, and the question arises as to how useful such a theory may be.

A theory of statistics based on a priori probability would indeed have very limited usefulness. While there are a few practical situations in which such a theory could be used (the field of genetics provides one important area), the great majority of fields of application occur where a priori probabilities do not exist. Our theory must be generalized, and we shall do it quite arbitrarily. We shall simply assume the existence of certain probabilities, and we shall assume that they obey the same laws as do combinatorial probabilities. We may consider the coin, mentioned earlier, which is known to be loaded in favor of heads. We shall assume that there is a number which gives the correct probability of a head, though one cannot say what the number is. We can, however, estimate the number. We may toss the coin a large number of times and divide the number of heads by the total

number of tosses. This if 62 heads appear in 100 tosses, we would estimate the probability to be .62. This estimate is called an *empirical probability*. We shall not make the error of stating that the correct probability of a head is .62, because we know that if the coin were tossed 100 times again, the number of heads might well differ from 62. The empirical probability is merely an estimate of what we think of as the true probability. We shall see later that the estimate can be made more and more accurate by increasing the number of trials in the experiment.

We may observe that we do not need to postulate the existence of a probability for every imaginable situation. We may as well limit ourselves to operationally meaningful situations. That is, we shall not assume the existence of a probability unless it is possible to set up an experiment by means of which the assumed probability can be estimated. Referring to the question, mentioned in the first section, of drawing an even number from the whole set of positive integers, we do not need to assume that such a probability exists. For there is no way to estimate it; we cannot build an urn large enough to hold balls numbered 1, 2, 3, $\cdots$ ad infinitum or even procure the balls. Clearly this kind of limitation in the theory will not limit its practical application.

Our position then is this: We develop the theory by thinking about ideal coins, ideal dice, ideal random drawings from an urn, and so forth. And we admit the existence of probabilities which have no a priori basis, provided they can be estimated. We speak of the probability of a head being one-half when a coin is tossed. But faced with an actual coin, we refuse to say what the probability of a head is. If the coin appears homogeneous and fairly symmetrical, we may guess that the probability is somewhere near one-half, but we shall not be surprised if a long series of trials indicates that the probability is somewhere between .57 and .58, for example. We shall not hesitate to make statements of the following kind: whatever the probability p may be, the probability of a tail is $1 - p$, the probability of two heads when the coin is tossed twice is p^2, the probability of a head and a tail in either order when the coin is tossed twice is $2p(1 - p)$, and so forth. Thus, we shall use our laws of probability on p.

The justification for these assumptions (that noncombinatorial probabilities exist, and that they obey the same laws as combinatorial probabilities) is simply that they work. A great mass of experimental evidence supports the assumptions, while no evidence has ever been brought forward which seriously controverts them.

2.11. Notes and References. The development of the theory of probability began in the seventeenth century and has continued steadily to the present day. It is therefore an old and now fairly extensive branch of applied mathematics. The subject had its origin in games of chance, but it brought forth such a variety of interesting problems that many eminent mathematicians were attracted to it. Today there is likely more work being done in this field than ever before, and this is due in large part to the rapid developments in statistics.

An excellent modern textbook on probability theory is J. V. Uspensky, "Introduction to Mathematical Probability," McGraw-Hill Book Company, Inc., New York, 1937.

2.12. Problems

1. An urn contains three white balls and seven black ones. What is the probability that one drawn at random will be white?

2. If two coins are tossed, what is the probability that a head and a tail will appear?

3. If a three-volume set of books is placed on a shelf in random order, what is the probability that they will be in the correct order?

4. What is the probability of obtaining three heads if three coins are tossed? What is the probability that at least two heads will appear?

5. An urn contains three white balls and two black ones. What is the probability that two balls drawn from the urn will both be white?

6. How many three-digit numbers can be formed with the integers 1, 2, 3, 4, 5, if duplication of the integers is not allowed? If duplication is allowed?

7. How many three-digit numbers can be formed from 0, 1, 2, 3, 4 if duplication is not allowed? How many of these are even?

8. In how many ways can a committee of three be chosen from nine men?

9. There are five roads from A to B and six roads from B to C. In how many ways can one go from A to C via B?

10. How many different sums of money can be formed with one each of the six kinds of coins minted by the United States Treasury?

11. In how many ways can six girls and four boys be divided into two groups of two boys and three girls?

12. In a baseball league of eight teams, how many games will be necessary if each team is to play every other team twice at home?

13. How many football teams can be formed with 12 men who can play any line position and 8 men who can play any back position?

14. How many signals can a ship show with five different flags if there are five significant positions on the flagpole?

15. How many license plates can be made if they are to contain five symbols, the first two being letters and the last three integers?

16. How many diagonals are there in a twelve-sided polygon?

17. How many dominoes are there in a set from double 0 to double 12?

18. What is the probability of getting a seven with a pair of dice?

19. What is the probability that two cards drawn from an ordinary deck will be spades?

20. What is the probability that a five-card hand will contain exactly two aces? At least two aces?

21. What is the probability that a bridge hand will be a complete suit?

22. An urn contains four white, five red, and six black balls. Another contains five white, six red, and seven black balls. One ball is selected from each urn. What is the probability they will be of the same color?

23. Show that $\binom{n}{r} = \binom{n}{n-r}$.

24. In how many ways can n different objects be divided into k groups containing $n_1, n_2, \cdots, n_k$ objects, if

$$n_1 + n_2 + \cdots + n_k = n - m?$$

25. An urn contains m white and n black balls. k balls are drawn and laid aside, their color unnoticed. Then another ball is drawn. What is the probability that it is white?

26. Six dice are tossed. What is the probability that every possible number will appear?

27. Seven dice are tossed. What is the probability that every number appears?

28. What is the probability of getting a total of five points with three dice?

29. An urn contains ten balls numbered from one to ten. Four balls are drawn, and suppose x is the second smallest of the four numbers drawn. What is the probability that $x = 3$?

30. If n balls are tossed into k boxes so that each ball is equally likely to fall in any box, what is the probability that a specified box will contain m balls?

31. Show that $\sum_{i=1}^{n} CX_i = C \sum_{i=1}^{n} X_i$.

32. Show that $\prod_{i=1} CX_i^a = C^n \left(\prod_{i=1} X_i\right)^a$.

33. Show that $\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1} \sum_{j=1} X_i X_j$.

34. Show that $\prod_{i=1}^{2n+1} (X + n + 1 - i) = \prod_{i=1}^{n} (X^2 - i^2)$.

35. Find the coefficient of x^6y^2 in the expansion of the binomial $(x^2 - ay)^5$.

36. Find the coefficient of $x^2y^2z^3$ in the expansion of the trinomial $(2x - y - z)^7$.

37. If six balls are tossed into three boxes so that each is equally likely to fall in any box, what is the probability that all boxes will be occupied?

38. The corners of a regular tetrahedron are numbered one, two, three, four. Five tetrahedra are tossed. What is the probability that the sum of the upturned corners will be 12?

39. The spades and hearts are removed from a deck of cards and placed face up in a row. The remaining cards are shuffled and dealt face up in a row beneath the row of spades and hearts. What is the probability that all the clubs will be beneath spades? What is the probability that among the 26 pairs of cards, 16 pairs will consist of cards of the same color?

40. Six cards are drawn from an ordinary deck. What is the probability that there will be one pair (two aces, or two fives, for example) and four scattered cards? That there will be two pairs and two scattered cards?

41. The face cards are removed from an ordinary deck and the remainder divided into the four suits. A card is drawn at random from each suit. What is the probability that the total of the four numbers drawn is 20?

42. An urn contains three black balls, three white ones, and two red ones. Three balls are drawn and placed in a black box, then three more are drawn and placed in a white box, and the remaining two are put in a red box. What is the probability that all but two of the balls will fall in boxes corresponding to their colors?

43. An urn contains four white and five black balls; a second urn contains five white and four black ones. One ball is transferred from the first to the second urn; then a ball is drawn from the second urn. What is the probability it is white?

44. In the above problem suppose two balls, instead of one, are transferred from the first to the second urn. Find the probability that a ball then drawn from the second urn will be white.

45. If it is known that at least two heads appeared when five coins were tossed, what is the probability that the exact number of heads was three?

46. If a bridge player has seven spades, what is the probability that his partner has at least one spade? At least two spades?

47. If a bridge player and his partner have eight spades between them, what is the probability that the other five spades are split three and two in the opposing hands?

48. A bridge player and his partner hold all spades except K, 3, 2. What is the probability that they are split K and 3, 2 in the opposing hands? What is the probability that K or K, 2 or K, 3 or K, 3, 2, appears in a specified one of the two opposing hands?

49. A person repeatedly casts a pair of dice. He wins if he casts an eight before he casts a seven. What is his probability of winning? NOTE: $1 + x + x^2 + x^3 + \cdots = 1/(1 - x)$, if $|x| < 1$.

50. In a dice game a player casts a pair of dice twice. He wins if the two numbers thrown do not differ by more than two with the following exceptions: if he gets a 3 on the first throw, he must produce a 4 on the second throw; if he gets an 11 on the first throw, he must produce a 10 on the second throw. What is his probability of winning?

51. The game of craps is played with two dice as follows: In a particular game one person throws the dice. He wins on the first throw if he gets 7 or 11 points; he loses on the first throw if he gets 2, 3, or 12 points. If he gets 4, 5, 6, 8, 9, or 10 points on the first throw, he continues to throw the dice repeatedly until he produces either a 7 or the number first thrown; in the latter case he wins, in the former he loses. What is his probability of winning?

52. In simple Mendelian inheritance, a physical characteristic of a plant or animal is determined by a single pair of genes. The color of peas is an example. Letting y and g represent yellow and green, peas will be green if the plant has the color-gene pair (g, g); they will be yellow if the color-gene pair is (y, y) or (y, g). In view of this last combination, yellow is said to be dominant to green. Progeny get one gene from each parent and are equally likely to get either gene from each parent's pair. If (y, y) peas are crossed with (g, g) peas, all the resulting peas will be (y, g) and yellow because of dominance. If (y, g) peas are crossed with (g, g) peas, the probability is .5 that the resulting peas will be yellow and is .5 that they will be green. In a large number of such crosses one would expect about half the resulting peas to be yellow, the remainder to be green. In crosses between (y, g) and (y, g)

peas, what proportion would be expected to be yellow? What proportion of the yellow peas would be expected to be (y, y)?

53. Peas may be smooth or wrinkled, and this is a simple Mendelian character. Smooth is dominant to wrinkled so that (s, s) and (s, w) peas are smooth while (w, w) peas are wrinkled. If (y, g) (s, w) peas are crossed with (g, g) (w, w) peas, what are the possible outcomes and what are their associated probabilities? For the (y, g) (s, w) by (g, g) (s, w) cross? For the (y, g) (s, w) by (y, g) (s, w) cross?

54. Albinism in human beings is a simple Mendelian character. Let a and n represent albino and nonalbino; the latter is dominant, so that normal parents cannot have an albino child unless both are (a, a). Suppose that in a large population the proportion of n genes is p and the proportion of a genes is $q = 1 - p$, so that q^2 of the individuals are albinos. Assuming that albinism is not a factor in the selection of marriage partners or in the number of children of a particular marriage, what proportion of individuals of the next generation would be expected to be albinos? If albinos married only albinos and had as many children on the average as nonalbinos, what proportion of individuals in the next generation would be expected to be albinos? What would happen eventually to the population if albinos continued generation after generation to mate only with albinos (assume number of individuals in each generation is the same)?

55. It is known that an urn was filled by casting a die and putting white balls in the urn equal in number to that obtained on the throw of the die. Then black balls were added in a number determined by a second throw of the die. It is also known that the total number of balls in the urn is eight. What is the probability that the urn contains exactly five white balls?

56. Urn A contains two white and two black balls; urn B contains three white and two black balls. One ball is transferred from A to B; one ball is then drawn from B and turns out to be white. What is the probability that the transferred ball was white?

57. Each of six urns contains 12 black and white balls; one has 8 white balls, two have 6 white balls, and three have 4 white balls. An urn is drawn at random, and three balls are drawn without replacement from that urn. Two of the three are white; the other is black. What is the probability that the urn drawn contained 6 white and 6 black balls?

58. Three newspapers, A, B, C, are published in a certain city. It is estimated from a survey that of the adult population:

20% read A
16% read B
14% read C
8% read both A and B
5% read both A and C
4% read both B and C
2% read all three

What percentage reads at least one of the papers? Of those that read at least one, what percentage reads both A and B?

59. Twelve dice are cast. What is the probability that each of the six faces will appear at least once?

60. A die is cast repeatedly until each of the six faces appears at least once. What is the probability that it must be cast ten times?

CHAPTER 3

DISCRETE DISTRIBUTIONS

3.1. Introduction. In Chap. 2 we were concerned with finding the probability of a specific outcome for a certain chance event. In this chapter we shall be concerned with a complete set of probabilities. A simple example will introduce the idea. What is the probability that x heads will appear if four coins are tossed? Denoting the probability by $f(x)$ (this is the functional notation):

$$f(x) = \frac{\binom{4}{x}}{2^4} \qquad 0 \leq x \leq 4 \tag{1}$$

We have a function which tells us directly what the probability is for any value of x in its possible range, which is zero to four inclusive. The function gives the complete set of probabilities for the given character (number of heads). We may calculate the function by giving x each of its possible values, and we may then plot the function, as in Fig. 2, using vertical lines of length equal to $f(x)$ on some scale. Since one of the values of x is certain to occur, the sum of the set of probabilities must be one, because the probability of zero or one, or two, or three, or four heads, is equal to the sum of the separate probabilities.

$$\sum_{x=0}^{4} f(x) = 1 \tag{2}$$

The function of $f(x)$ is called a *discrete probability density function*, or *distribution function*. We shall usually refer to it more briefly as simply a *density* or a *distribution*. It is useful to think of $f(x)$ as giving the relative frequency of occurrence of the separate values of x. Thus, suppose the four coins were tossed a very large number of times. We should expect no heads to appear ($x = 0$) in about one-sixteenth of the tosses; we should expect one head to appear ($x = 1$) in about one-fourth of the tosses, and so forth. The graph of the density makes a number of things immediately evident: that the most likely number of heads is two, that one head can be expected to occur about four times as often as no heads, that three heads can be expected to occur about

as often as one head, and so forth. The word "about" is used because we are familiar with the fluctuations that accompany chance events. Thus, if a single coin is to be tossed ten times, we expect five heads and five tails on the average, but actually some other division of heads and tails is quite likely to occur in a given trial.

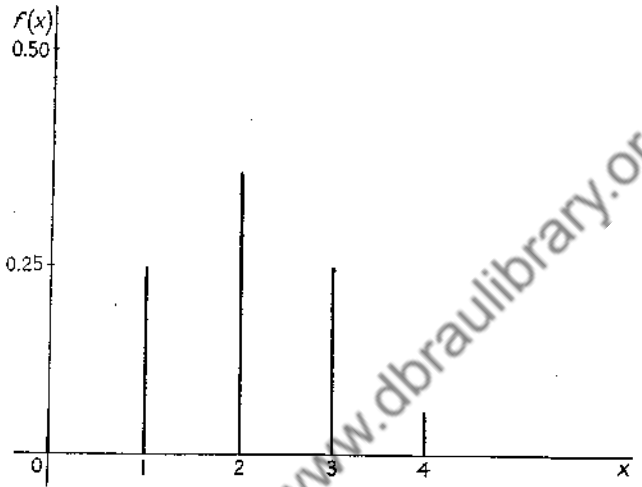


FIG. 2.

The results of an actual experiment in tossing four coins are given in the following table. Four coins were tossed 160 times and the number of heads counted on each toss.

RESULTS OF TOSSING FOUR COINS 160 TIMES

Number of heads	Actual occurrences	Expected occurrences
0	6	10
1	41	40
2	56	60
3	45	40
4	12	10
	160	160

The agreement between actual and expected occurrences is none too good (it is to be remembered that the probability of a head may not have been exactly one-half for each of the four coins actually used), but still the general character of the distribution of actual outcomes was fairly well indicated by the distribution function $f(x)$.

Knowing the density function of some attribute x, we can supply the answer to any probability question pertaining to x. Thus, referring again to our particular example, the probability of two heads is

$$P(x = 2) = f(2) = \frac{\binom{4}{2}}{2^4} = \frac{3}{8}$$

The probability that the number of heads will be less than three is

$$P(x < 3) = \sum_{x=0}^{2} f(x) = 11/16$$

The probability that the number of heads will be between one and three inclusive is

$$P(1 \leq x \leq 3) = \sum_{x=1}^{3} f(x) = 7/8$$

Given that the number of heads on a specific outcome is less than four, the conditional probability that the number is not more than two is

$$P(x \leq 2 | x < 4) = \frac{\sum_{x=0}^{2} f(x)}{\sum_{x=0}^{3} f(x)} = \frac{11}{15}$$

The symbol $P(\cdot \cdot \cdot)$ will always be used as it has been used here and may be read "the probability that . . . ." Thus in the last equation, the symbol represents this phrase: the probability that x is less than or equal to two given that x is less than four. A vertical bar used in the symbol will always mean "given that" or "when it is known that" and will precede the specified condition of a conditional probability.

3.2. Discrete Density Functions. The essential properties of discrete density functions have already been suggested in the preceding section, and we need only to describe them in somewhat more general language.

The set of possible outcomes of a chance event are classified into a number, say k, of mutually exclusive classes according to some attribute. Associated with each class is a value of a *random variable*, or *variate*, x. The density function is a function of x which gives the probability that any specified value of x will occur.

The variate x may naturally describe the attribute, as was the case in the coin-tossing illustration, or it may simply be a code. Thus in

drawing balls from an urn, the classification may be according to color. We could define a random variable x by arbitrarily setting a correspondence between values of x and colors: $x = 1$ corresponds to black; $x = 2$ corresponds to red; and so forth. When a red ball is drawn, the variate has the value two.

The density function may be a mathematical expression involving x, as was the case in the preceding section, or it may be only a table of values. Thus if an urn contains three black, two red, and five white balls, we may code the colors 1, 2, 3, respectively, and find the probabilities .3, .2, and .5. We do not bother to construct a mathematical expression which will take on these values when x is put equal to 1, 2, and 3, but merely tabulate the function:

x:	1	2	3
$f(x)$:	.3	.2	.5

The word *discrete* is used to distinguish the variate from *continuous* variates, which will be discussed in the next chapter. A variate x is discrete if it can take on only isolated values, i.e., if successive possible values of x are separated on the x axis. The distinction will be brought out in more detail in the next chapter.

The set of probabilities represented by a density function will always have a sum equal to one because we shall speak of a density only when (1) all the possible outcomes are included among the separate classes of outcomes, (2) the classes are mutually exclusive.

3.3. Multivariate Distribution. When the outcome of a chance event can be characterized in more than one way, the probability density function is a function of more than one variable. Thus when a card is drawn from an ordinary deck, it may be characterized according to its suit and to its denomination. Let $x = 1, 2, 3, 4$ correspond to the suits in some order (say, spades, hearts, diamonds, clubs), and let $y = 1, 2, 3, \cdots, 13$ correspond to the denominations, A, 2, $\cdots$, 10, J, Q, K. The probability of drawing a particular card will be denoted by $f(x, y)$ and clearly

$$f(x, y) = 1/52 \qquad 1 \leq x \leq 4,\ 1 \leq y \leq 13 \tag{1}$$

This function may be plotted over a plane as in Fig. 3; the probabilities are represented by vertical lines at the points (x, y) in the horizontal plane where the probabilities are defined. In this case, since the function is a constant, the lines are of equal height.

To consider another example: Let four balls be drawn from an urn containing five black, six white, and seven red balls. Let x be the

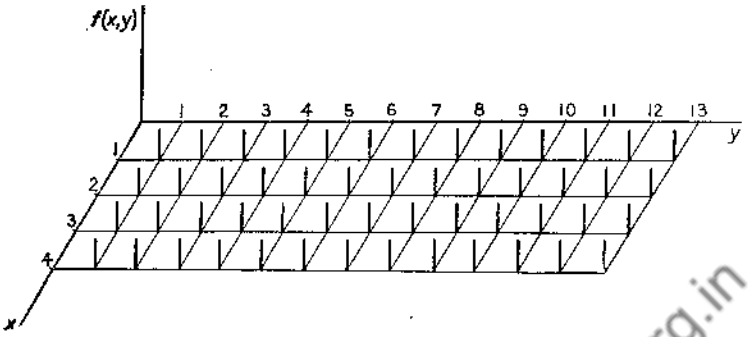


FIG. 3.

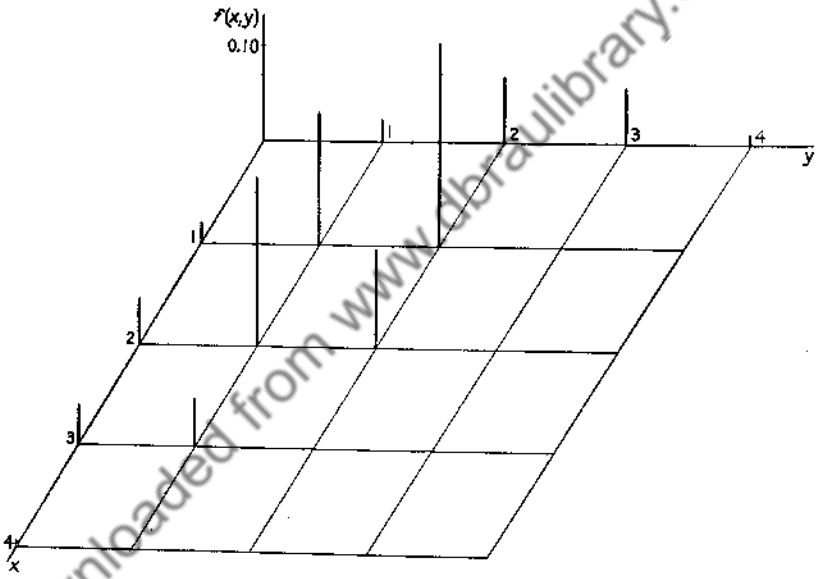


FIG. 4.

number of white balls drawn and y be the number of red balls drawn. The density is

$$f(x, y) = \frac{\binom{6}{x}\binom{7}{y}\binom{5}{4-x-y}}{\binom{18}{4}} \qquad 0 \le x + y \le 4 \tag{2}$$

and its graph is shown in Fig. 4. In this example, we might consider defining a third random variable, z, to be the number of black balls

drawn, and obtain a trivariate distribution. But z is exactly determined by x and y since $z = 4 - x - y$. No new information can be obtained by adding z to the set of random variables characterizing the outcomes, and, in fact, if z were included in the distribution function, the set of probabilities represented by that function, $f(x, y, z)$, would be exactly the same set that we have already obtained using x and y.

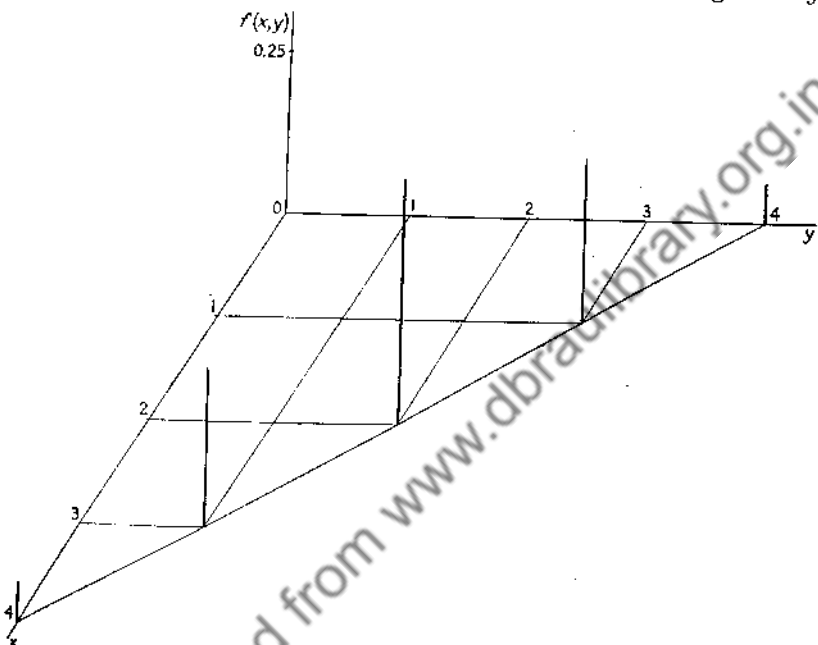


FIG. 5.

A simpler example of functional dependence is that of tossing a coin, say four times. Let x be the number of heads and y be the number of tails. Since $x + y$ must be equal to four, the variables are functionally dependent; knowing one, the other is exactly determined. The density is

$$f(x, y) = \binom{4}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^y \qquad x + y = 4$$

and its graph is given in Fig. 5. It gives us no more information than the function used as an example in Sec. 1; the set of probabilities is exactly the same as before.

We have used the terms dependent and independent in two entirely different connections. In Chap. 2 we defined two events to be inde-

pendent if the conditional probability of one, given the other, was equal to the marginal probability of the first. We shall in the future refer to this kind of independence as *independence in the probability sense.* Returning to the urn example: x and y are *functionally independent* (since y is not uniquely determined when x is known), but they are *dependent in the probability sense* (as we shall see).

In the urn example, the marginal density of x is found by applying the definition in Sec. 2.7 (i.e., Sec. 7 of Chap. 2), and is

$$f(x) = \sum_{y=0} f(x, y) = \frac{\binom{6}{x}\binom{12}{4-x}}{\binom{18}{4}} \qquad 0 \le x \le 4 \tag{3}$$

The sum may be performed by means of an algebraic identity, but here it is simpler to consider the problem anew as one involving 6 white balls and 12 that are not white. Similarly the marginal density of y is

$$f(y) = \sum_{x=0}^{x} f(x, y) = \frac{\binom{7}{y}\binom{11}{4-y}}{\binom{18}{4}} \qquad 0 \le y \le 4 \tag{4}$$

This function is plotted in Fig. 6. The height of the line at $y = 0$, which represents $f(0)$, is equal to the sum of the lengths of the vertical lines along the x axis in Fig. 4; $f(1)$ is the sum of the lengths of the vertical lines along the line $y = 1$ in Fig. 4, and so forth.

The conditional density of x, given y, is defined exactly as in Sec. 2.7 and is denoted by

$$\begin{aligned} f(x|y) &= \frac{f(x, y)}{f(y)} \\ &= \frac{\binom{6}{x}\binom{5}{4-x-y}}{\binom{11}{4-y}} \qquad 0 \le x \le 4 - y \end{aligned}$$

Similarly

$$f(y|x) = \frac{\binom{7}{y}\binom{5}{4-x-y}}{\binom{12}{4-x}} \qquad 0 \le y \le 4 - x$$

If x were given some specific value, say $x = 1$, we could plot the density

$f(y|1)$ by giving y its successive values: 0, 1, 2, 3. The vertical lines would have the same relative heights as those along the line $x = 1$ in Fig. 4; their lengths would be increased by the factor $1/f(x)$ evaluated for $x = 1$ so that the sum of their lengths would be one. We observe that $f(y|x)$ is not equal to the marginal distribution of y, so that y and x are not independent in the probability sense. Of course, the fact that $f(y|x)$ involves x is sufficient evidence that the two variates are dependent in the probability sense. If, however, we had an example in which

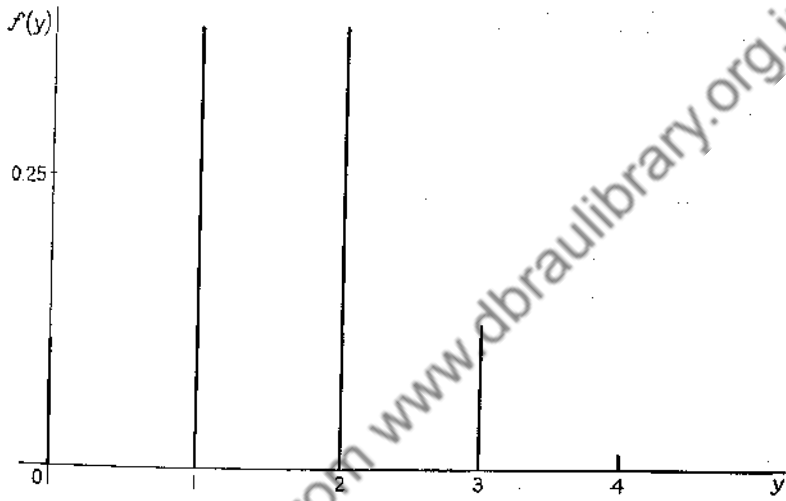


FIG. 6.

$f(y|x)$ did not involve x, it would still be possible for the two variates to be dependent because the range of y might depend on x. If both $f(y|x)$ and the range of y do not involve x, then the two variates will obviously be independent in the probability sense.

As an example of a distribution involving several variates, suppose 12 cards are drawn without replacement from an ordinary deck, and let x_1 be the number of aces, x_2 be the number of deuces, x_3 be the number of treys, and x_4 be the number of fours. The distribution of these variates is given by a function of four variates and is, in fact,

$$f(x_1, x_2, x_3, x_4) = \frac{\binom{4}{x_1}\binom{4}{x_2}\binom{4}{x_3}\binom{4}{x_4}\binom{36}{12 - x_1 - x_2 - x_3 - x_4}}{\binom{52}{12}}$$

where the range of each variate is $0 \leq x_i \leq 4$ subject to the restriction

that $\Sigma x_i \leq 12$. There are a large number of marginal and conditional distributions associated with this distribution; a few examples are

$$f(x_2, x_3) = \frac{\binom{4}{x_2}\binom{4}{x_3}\binom{44}{12 - x_2 - x_3}}{\binom{52}{12}} \qquad \begin{array}{l} 0 \leq x_i \leq 4 \\ x_2 + x_3 \leq 8 \end{array}$$

$$f(x_4) = \frac{\binom{4}{x_4}\binom{48}{12 - x_4}}{\binom{52}{12}} \qquad 0 \leq x_4 \leq 4$$

$$f(x_2, x_4 | x_1, x_3) = \frac{\binom{4}{x_2}\binom{4}{x_4}\binom{36}{12 - x_1 - x_2 - x_3 - x_4}}{\binom{44}{12 - x_1 - x_3}} \qquad \begin{array}{l} 0 \leq x_i \leq 4 \\ x_2 + x_4 \leq 12 - x_1 - x_3 \end{array}$$

the first two being marginal distributions and the third a conditional distribution. The distribution $f(x_1, x_2, x_3, x_4)$ itself may in this case be regarded as a marginal distribution of some more detailed distribution, for example, the six-variate distribution of $x_1, x_2, x_3, x_4, x_5, x_6$, where x_5 and x_6 are the numbers of fives and sixes that appear among the 12 cards drawn.

We cannot plot the four-variate distribution; in fact, we have used all three dimensions of conceptual space in plotting bivariate distributions. This could have been avoided by using a different device; we might have used dots of different sizes rather than vertical lines and thus pictured the bivariate distributions in two dimensions. This method would not have given as clear a representation of the relative magnitude of the probabilities. Using the dots, we could get a pictorial representation of a trivariate distribution, but for more than three variates no simple graphical representation is possible.

The probability that random variables will fall in any region of their space is obtained by summing the density function over all points in the region. Suppose a bivariate density $f(x, y)$ is defined for $x = 0, 1, 2, \cdots, r$ and $y = 0, 1, 2, \cdots, s$. The probability that $x < 5$ and $y \leq 3$ is obtained by summing $f(x, y)$ over the region defined by the inequalities (the rectangle in Fig. 7).

$$P(x < 5, y \leq 3) = \sum_{x=0}^{4} \sum_{y=0}^{3} f(x, y)$$

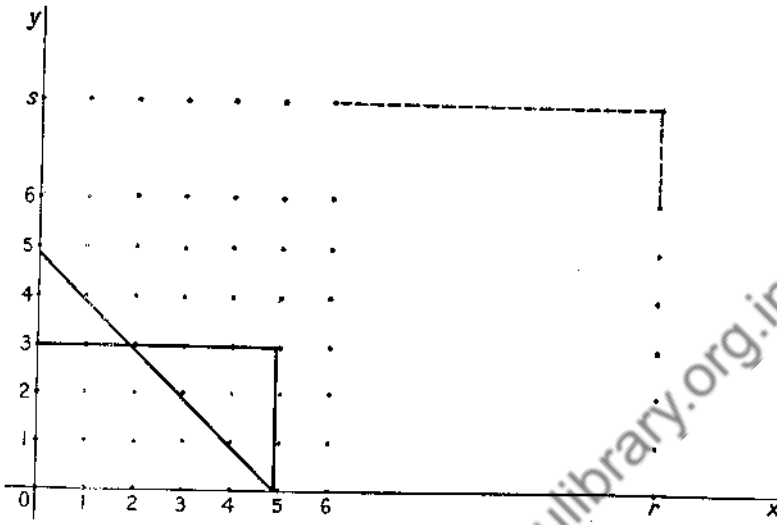


FIG. 7.

The probability that the sum of x and y is less than 5 is equal to the sum of $f(x, y)$ over all points within the triangle bounded by the line $x + y = 5$.

$$\begin{aligned} P(x + y < 5) &= f(0, 0) + f(1, 0) + f(2, 0) + f(3, 0) + f(4, 0) \\ &\quad + f(0, 1) + f(1, 1) + f(2, 1) + f(3, 1) \\ &\quad + f(0, 2) + f(1, 2) + f(2, 2) \\ &\quad + f(0, 3) + f(1, 3) \\ &\quad + f(0, 4) \\ &= \sum_{x=0}^{4} \sum_{y=0}^{4-x} f(x, y) = \sum_{y=0}^{4} \sum_{x=0}^{4-y} f(x, y) \end{aligned}$$

Some other examples are

$$P(x + y = 5) = \sum_{x=0}^{5} f(x, 5 - x)$$

$$P(x \leq 2 | y = 3) = \sum_{x=0}^{2} f(x|3)$$

$$= \frac{\sum_{x=0}^{2} f(x, 3)}{\sum_{x=0}^{r} f(x, 3)}$$

$$P(x \leq 2|y > 3) = \frac{\sum_{x=0}^{2} \sum_{y=4}^{8} f(x, y)}{\sum_{x=0}^{7} \sum_{y=4}^{8} f(x, y)}$$

$$P(x + y = 2|x^2 + y^2 \leq 5)$$
$$= \frac{f(0, 2) + f(1, 1) + f(2, 0)}{f(0,0) + f(0,1) + f(0,2) + f(1,0) + f(1,1) + f(1,2) + f(2,0) + f(2,1)}$$

For three variables, the regions may be troublesome to visualize, and for more than three variables, we must rely on the analytical description of the region to determine the required sums. Some relatively easy examples are

$$P(x \leq 3, y \leq 4, 2 \leq z \leq 6) = \sum_{x=0}^{3} \sum_{y=0}^{4} \sum_{z=2}^{6} f(x, y, z)$$

$$P(x + y = 4|z = 2) = \sum_{x=0}^{4} f(x, 4 - x|2)$$

$$P(x + y + z \leq 6) = \sum_{x=0}^{6} \sum_{y=0}^{6-x} \sum_{z=0}^{6-x-y} f(x, y, z)$$

$$P(x + y + z = 6) = \sum_{x=0}^{6} \sum_{y=0}^{6-x} f(x, y, 6 - x - y)$$

3.4. The Binomial Distribution. The binomial distribution is probably the most frequently used discrete distribution in applications of the theory of statistics. It is the distribution associated with repeated trials of the same event. Suppose we denote by p the probability of success of some event. The event may be the occurrence of a head when a coin is tossed, in which case $p = \frac{1}{2}$; it may be the occurrence of a seven when two dice are cast, in which case $p = \frac{1}{6}$; it may be the occurrence of at least two aces when five cards are drawn from an ordinary deck, in which case

$$p = \frac{\left[\binom{4}{2}\binom{48}{3} + \binom{4}{3}\binom{48}{2} + \binom{4}{4}\binom{48}{1}\right]}{\binom{52}{5}}$$

Or more generally, p may represent the probability of occurrence of some actual event to which no numerical a priori probability can be assigned.

Whatever the event, if the probability of its occurrence is p, the probability of its nonoccurrence is $1 - p$, since we cannot suppose that the event can both occur and not occur in a given trial. It will be convenient to denote $1 - p$ by q, and in speaking of a given trial we shall say the probability of a success is p and the probability of a failure is q.

$$p + q = 1$$

Now suppose that n trials are made. We shall be concerned with the number of successes, x, that occur among the n trials. The variate x has the density

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad 0 \leq x \leq n \tag{1}$$

since there are $\binom{n}{x}$ orders in which x success and $n - x$ failures can occur, while the probability for any particular order is $p^x q^{n-x}$. This distribution is the binomial distribution. It is a discrete distribution of one random variable, x.

The function contains two other variables p and n (q is not counted because it is determined by p) of a different character. Their variation is between different binomial distributions; for a specific binomial distribution, p and n must be given numerical values. Variables of this kind are called *parameters*. The function actually represents a *two-parameter family* of distributions, and a specific member of the family is given when p and n are given specific values. The parameter n is called a *discrete parameter*, since it can have only the isolated values 1, 2, 3, $\cdots$; it would be meaningless to speak of, say, 2.53 trials. But p is a *continuous parameter*, since it can conceivably have any value in the range zero to one. Thus it is possible for p to be .5, say, in the case of a true coin, or possibly .5000037 in the case of a slightly biased coin. Any arbitrarily chosen number between zero and one is an allowable value of p.

Two particular binomial distributions are plotted in Fig. 8. In (*a*), $p = .4$ and $n = 4$; in (*b*), $p = .8$ and $n = 3$. In general, the binomial density will have a maximum value determined as follows: Let m be the integral part of the number $(n + 1)p$ and let e be the fractional part. Thus if $n = 7$ and $p = .3$, we have $m = 2$ and $e = .4$. The largest value of $f(x)$ occurs when x is put equal to m; m is called the *modal value* or simply the *mode* of x. To prove that this value of x does maximize $f(x)$, let us assume for the moment that e is not zero, and let us form the ratio $f(x + 1)/f(x)$. We wish to show that this ratio is less than one when x is greater than or equal to m, and greater

than one when x is less than m. We are thinking of a situation like that illustrated in Fig. 9. Now

$$\frac{f(x+1)}{f(x)} = \frac{p}{q}\frac{n-x}{x+1}$$

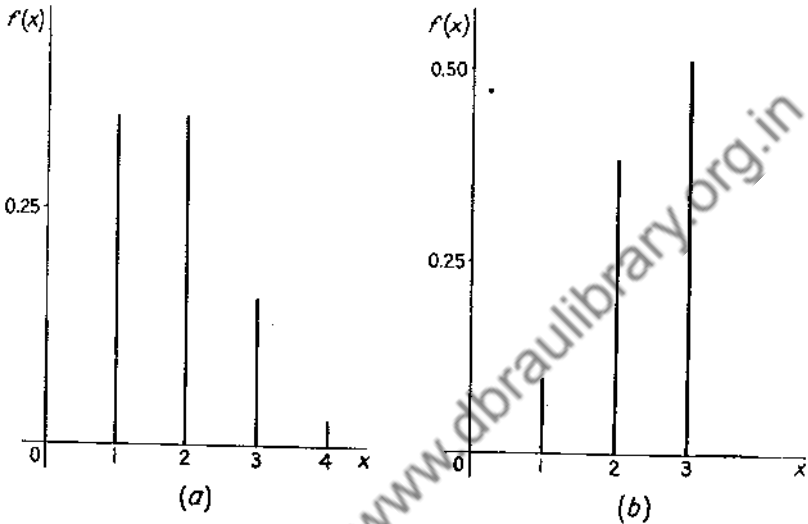


FIG. 8.

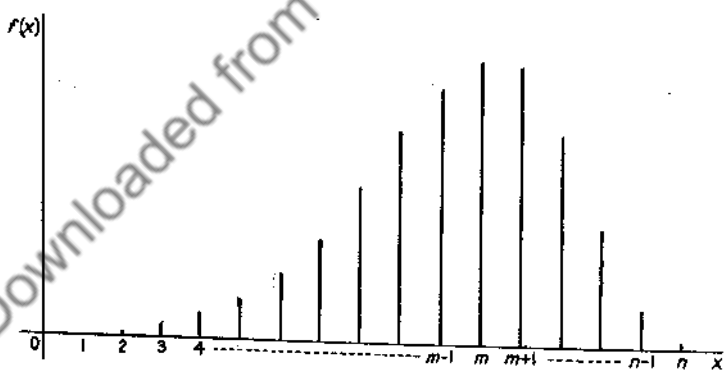


FIG. 9.

and if x is greater than or equal to m, then

$$\frac{p}{q}\frac{n-x}{x+1} \leq \frac{p}{q}\frac{n-m}{m+1}$$

On substituting $(n+1)p - e$ for m, the right-hand expression may be written

$$\frac{p}{q}\frac{n-m}{m+1} = \frac{(n+1)-[(1-e)/q]}{(n+1)+[(1-e)/p]}$$

which is certainly less than one. If x is less than $m-1$,

$$\begin{aligned}\frac{p}{q}\frac{n-x}{x+1} &> \frac{p}{q}\frac{n-(m-1)}{m}\\ &> \frac{p}{q}\frac{(n+1)q+e}{(n+1)p-e}\\ &> \frac{n+1+e/q}{n+1-e/p}\end{aligned}$$

and is therefore greater than one. We have omitted the case

$$x = m-1$$

here

$$\begin{aligned}\frac{f(x+1)}{f(x)} &= \frac{p}{q}\frac{n-m+1}{m}\\ &= \frac{(n+1)+e/q}{(n+1)-e/p}\end{aligned}$$

which is again greater than one if e is not zero. If $e = 0$, the ratio is equal to one, and $f(m) = f(m-1)$; there are two largest values of $f(x)$ which are equal and which occur at $x = m$ and at $x = m-1$. This situation is illustrated in Fig. 8(a) where $(n+1)p = 2$ is an exact integer, so that $f(1)$ and $f(2)$ are two equal maximum values of $f(x)$.

For large values of n the appearance of the binomial distribution is generally like that of Fig. 9. In Fig. 8(b) the mode is at $x = n$ when $p = .8$ and $n = 3$, but as n increases, the mode moves away from the extreme right end of the range; thus, if $n = 100$, we have

$$101 \times .8 = 80.8$$

so that the mode is 80 and is well away from the extreme value of $x = 100$.

The computation of binomial probabilities becomes troublesome when n is large. Approximate methods can be developed for computing $\binom{n}{x} p^x q^{n-x}$, but we shall omit these because the computation of single terms is rarely required. In most applications, partial sums are needed. Thus we may require the probability that x be greater than

an integer a,

$$P(x > a) = \sum_{x=a+1}^{n} f(x)$$

Methods of computing such sums will be given in Chaps. 7 and 11.

3.5. The Multinomial Distribution. The multinomial distribution is associated with repeated trials of an event which can have more than two outcomes. Thus the outcome of tossing a die may be any one of the six numbers 1, 2, $\cdots$, 6. If the event refers to the appearance of aces when, say, seven cards are drawn, there are five possible outcomes: 0, 1, 2, 3, or 4 aces.

In general, suppose there are k possible outcomes of a chance event, and let the probabilities of these outcomes be denoted by $p_1, p_2, \cdots, p_k$. Obviously we must have

$$\sum_{i=1}^{k} p_i = 1 \tag{1}$$

just as $p + q = 1$ in the binomial case. Suppose the event is repeated n times, and let x_1 be the number of times the outcome associated with p_1 occurs, let x_2 be the number of times the outcome associated with p_2 occurs, and so forth. The density for the random variables $x_1, x_2, \cdots, x_k$ is

$$f(x_1, x_2, \cdots, x_{k-1}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \tag{2}$$

where the range of each x_i is zero to n inclusive, subject to the restriction that $\sum_{i=1}^{k} x_i = n$. We have written the function as one involving only $k - 1$ of the x_i's since only $k - 1$ of them are functionally independent; x_k is exactly determined by the relation $\sum_{1}^{k} x_i = n$ when the $x_1, \cdots, x_{k-1}$ are specified. Thus this is a multivariate distribution involving $k - 1$ variates. The x_k on the right-hand side of (2) is to be interpreted as merely a symbol for the expression

$$n - x_1 - x_2 - \cdots - x_{k-1}$$

The expression (2) is a k-parameter family of distributions, the parameters being $n, p_1, p_2, \cdots, p_{k-1}$. The other variable p_k is, like

q in the binomial distribution, exactly determined by

$$p_k = 1 - p_1 - p_2 - \cdots - p_{k-1}$$

A particular case of a multinomial distribution is obtained by putting, e.g., $n = 3$, $k = 3$, $p_1 = .2$, $p_2 = .3$ to get

$$f(x_1, x_2) = \frac{3!}{x_1!x_2!(3 - x_1 - x_2)!}(.2)^{x_1}(.3)^{x_2}(.5)^{3-x_1-x_2}$$

This function is plotted in Fig. 10.

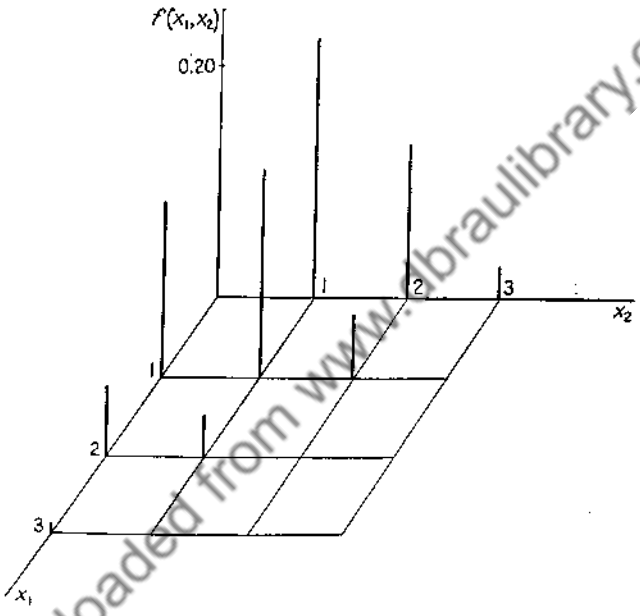


FIG. 10.

It may be shown by a direct generalization of the argument used in the preceding section that the maximum value of $f(x_1, x_2, \cdots, x_{k-1})$ occurs when the x_i are put equal to m_i, the integral parts of $(n + 1)p_i$.

3.6. The Poisson Distribution. The Poisson density is represented by the function

$$f(x) = \frac{e^{-m}m^x}{x!} \qquad x = 0, 1, 2, 3, \cdots \tag{1}$$

which has an infinite range. Since the exponential e^m has the series expansion

$$e^m = 1 + m + \frac{m^2}{2!} + \cdots + \frac{m^x}{x!} + \cdots$$

it follows that

$$\sum_{x=0}^{\infty} f(x) = 1$$

The distribution has useful application in situations where a large number of objects are distributed over a large area. To consider a concrete example, suppose a volume V of fluid contains a large number N of small organisms. It is assumed that the organisms have no social instincts, and that they are as likely to appear in any part of the fluid as in any other part with the same volume. Now suppose a drop of volume D is to be examined under a microscope, what is the probability that x organisms will be found in the drop? We assume that V is very much larger than D. Since the organisms are assumed to be distributed throughout the fluid with uniform probability, it follows that the probability that any given one of them may be found in D is D/V. And since they are assumed to have no social instincts, the occurrence of one in D has no effect on whether or not another occurs in D. The probability that x of them occur in D is therefore

$$\binom{n}{x}\left(\frac{D}{V}\right)^x\left(\frac{V-D}{V}\right)^{n-x} \tag{2}$$

We are also assuming here that the organisms are so small that the question of crowding may be neglected; all N of them would occupy no appreciable part of the volume D. The Poisson density is an approximation to the above expression, which is simply a binomial density in which $p = D/V$ is very small.

The Poisson distribution is obtained by letting V and N become infinite in such a way that the density of organisms $N/V = d$ remains constant. Rewriting (2) in the form

$$\frac{N(N-1)(N-2)\cdots(N-x+1)}{x!N^x}\left(\frac{ND}{V}\right)^x\left(1-\frac{ND}{NV}\right)^{N-x}$$
$$= \frac{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{x-1}{N}\right)(Dd)^x\left(1-\frac{Dd}{N}\right)^{N-x}}{x!}$$

the limit as N becomes infinite is readily seen to be

$$\frac{e^{-Dd}(Dd)^x}{x!}$$

which is the same form as (1) if we put $Dd = m$. This derivation shows that m is the average value of x, since D, the volume of the

portion examined, multiplied by the over-all density d gives the average number expected in the volume D.

We have gone into some detail in discussing this distribution because it is often erroneously applied to data which do not fulfill the assumptions required by the distribution. Thus it cannot be used, for example, in studying the distribution of insect larvae over some large crop area, because insects lay their eggs in clusters so that if one is found in a given small area, others are likely to be found there also.

The Poisson density function is perhaps best thought of as an approximation to the binomial density, $\binom{N}{x} p^x q^{N-x}$, when Np is large relative to p and N is large relative to Np. It is particularly useful when N is unknown.

3.7. Other Discrete Distributions. The *hypergeometric distribution* is

$$f(x) = \frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}} \tag{1}$$

Equation (3.3) gives a special example. Equation (3.2) is an example of a bivariate hypergeometrical distribution.

The *uniform distribution* is

$$f(x) = \frac{1}{n} \qquad x = 1, 2, \cdots, n \tag{2}$$

The casting of a die provides an example.

The *negative binomial distribution* is

$$f(x) = p^r \binom{x+r-1}{r-1} q^x \qquad x = 0, 1, 2, \cdots \tag{3}$$

and $\Sigma f(x) = 1$ since

$$\sum_{x=0}^{\infty} \binom{x+r-1}{r-1} q^x = \frac{1}{(1-q)^r} = \frac{1}{p^r}$$

An example is provided by letting p be the probability of success and q be the probability of failure of a given event. Let $f(x)$ be the probability that exactly $x + r$ trials will be required to produce r successes. The last trial must be a success, and its probability is p. Among the other $x + r - 1$ trials there must be $r - 1$ successes, and the prob-

ability of this is

$$\binom{x+r-1}{r-1} p^{r-1} q^x$$

The product of these two probabilities gives the desired probability, $f(x)$, and is the same as (3).

3.8. Problems. Specify range of variates for every distribution. Do not obtain numerical answers which require lengthy computations.

1. Five cards are dealt from an ordinary deck. What is the density function for the number of spades?

2. Ten balls are tossed into four boxes so that each ball is equally likely to fall in any box. What is the density for the number of balls in the first box?

3. A coin is tossed until a head appears. What is the density for the number of tosses?

4. What is the density for the number that appears when a die is cast?

5. Two dice are cast. What is the density of the sum of the two numbers which appear?

6. Cards are drawn from an ordinary deck without replacement until a spade appears. What is the density for the number of draws?

7. Ten dice are cast. What is the density of the number of ones and twos?

8. An urn contains m black and n white balls. k balls are drawn without replacement. What is the density of the number of white balls? Specify the range for the various relative sizes of m, n, and k.

9. Three coins are tossed n times. Find the joint density of x, the number of times no heads appear; y, the number of times one head appears; and z, the number of times two heads appear.

10. A machine makes nails with an average of 1 per cent defective. What is the density of the number of defectives in a sample of 50 nails?

11. An urn contains 10 white and 20 black balls. Balls are drawn one by one, without replacement, until 5 white ones have appeared. Find the density of the total number drawn.

12. Seven cards are drawn without replacement from an ordinary deck. Find the joint density of the number of aces and the number of kings.

13. Show that

$$\sum_{i=0}^{c} \binom{a}{i}\binom{b}{c-i} = \binom{a+b}{c}$$

by equating coefficients of x^c in

$$(1 + x)^a(x + 1)^b = (1 + x)^{a+b}$$

Hence verify algebraically that the sum of the hypergeometric density is one.

14. Use the result of Prob. 13 to find the marginal density of the number of aces from the result of Prob. 12.

✓**15.** In a town with 5000 adults, a sample of 100 are asked their opinion of a proposed municipal project; 60 are found to favor it and 40 to oppose it. If in fact the adults of the town were equally divided on the proposal, what would be the probability of obtaining a majority of 60 or more favoring it in a sample of 100?

✓**16.** A distributor of bean seeds determines from extensive tests that 5 per cent of a large batch of seeds will not germinate. He sells the seeds in packages of 200 and guarantees 90 per cent germination. What is the probability that a given package will violate the guarantee?

✓**17.** A manufacturing process is intended to produce electrical fuses with no more than 1 per cent defective. It is checked every hour by trying 10 fuses selected at random from the hour's production. If one or more of the 10 fails, the process is halted and carefully examined. If in fact its probability of producing a defective fuse is .01, what is the probability that the process will needlessly be examined in a given instance?

18. Referring to the above problem, how many fuses (instead of 10) should be tested if the manufacturer desires that the probability be about 0.95 that the process will be examined when it is producing 10 per cent defectives?

19. A has two pennies; B has one. They match pennies until one of them has all three. What is the density of the number of trials required to end the game?

20. Referring to the above problem, what is the density of the number of trials given that A wins?

21. A die is cast ten times. What is the probability that the number of ones and twos will not differ by more than two from its modal value?

✓**22.** A Poisson distribution has a double mode at $x = 1$ and $x = 2$; what is the probability that x will have one or the other of these two values?

✓**23.** Red-blood-cell deficiency may be determined by examining a specimen of the blood under a microscope. Suppose a certain small fixed volume contains on the average 20 red cells for normal persons.

What is the probability that a specimen from a normal person will contain less than 15 red cells?

✓**24.** An insurance company finds that 0.005 per cent of the population dies from a certain kind of accident each year. What is the probability that the company must pay off on more than 3 of 10,000 insured risks against such accidents in a given year?

✓**25.** A telephone switchboard handles 600 calls on the average during a rush hour. The board can make a maximum of 20 connections per minute. Use the Poisson distribution to estimate the probability that the board will be overtaxed during any given minute.

26. A die is cast until a six appears. What is the probability that it must be cast more than ten times?

27. Two dice are cast ten times. Let x be the number of times no ones appear, and let y be the number of times two ones appear. What is the probability that x and y will each be less than 3?

28. In Prob. 27 what is the probability that $x + y$ will be 4? What is the probability that $x + y$ will be between 2 and 4 inclusive?

29. A die is cast twenty times. What is the probability that there will be at least twice as many ones and twos as there are threes?

✓**30.** Ten cards are drawn without replacement from an ordinary deck. What is the probability that the number of spades will exceed the number of clubs?

✓**31.** Suppose a neutron passing through plutonium is equally likely to release 1, 2, or 3 other neutrons, and suppose these second-generation neutrons are in turn each equally likely to release 1, 2, or 3 third-generation neutrons. What is the density of the number of third-generation neutrons?

32. Using the density of Prob. 12, find the conditional density of the number x of aces, given the number y of kings.

33. Using the density of Prob. 9, find the conditional density of x and z, given y.

Determine the sums required to compute the following probabilities using density functions with as many variates as needed. Assume all variates take the values: $0, 1, 2, \cdots, m$.

34. $P(2x + y \leq 3)$

35. $P(x^2 + y^2 = 25)$

36. $P(x^2 < 5 | 1 \leq y \leq 6)$

37. $P(x > 2y - a), 0 < a < m$

38. $P(x > y > z)$

39. $P(x + y = 5 | y = 3)$

40. $P(x + y = 5 | z = 3)$

41. $P(x \leq 3, y \leq 4, z \geq 5, w \geq 6)$

42. $P(a \leq x \leq b | y = z), 0 < a < b < m$

43. $P(x > 2y | x > z)$

CHAPTER 4

DISTRIBUTIONS FOR CONTINUOUS VARIATES

4.1. Continuous Variates. A continuous variate is one that is not restricted to have only isolated values; it may have any value in a certain interval or collection of intervals.

To consider an example, suppose a rifle is perfectly aimed at the center of a square target and fired several times after being clamped in that position. The bullets will not all strike the center, because minor variations in the weight of the bullets, shape of the bullets, in the effect of humidity and temperature on the powder, and other factors, will cause variations in the trajectories of the bullets. After a few shots the appearance of the target might be represented by Fig. 11. Let a random variable x be defined as the horizontal deviation of the center of a hit from a vertical line through the center of the target. Clearly x may have any value in its possible range of variation.

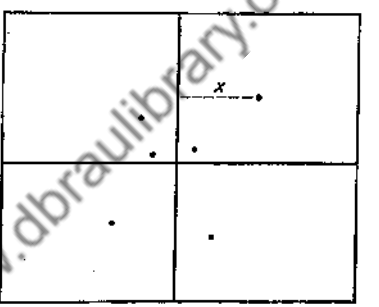


FIG. 11.

The number of possible values of x is infinite. In fact, any finite interval, however small, contains an infinite number of points. The interval .001 to .002, for example, contains among others the points .0011, .00111, .001111, .0011111, and so on. This fact raises some difficulties about defining the probability of x. In order to understand the problem, we must digress briefly to consider the number of points in an interval.

The number of positive integers is infinite; it is called a *denumerable infinity*. The symbol A_0 will be used to denote a denumerable infinity. Any set of objects which can be put into one-to-one correspondence with the positive integers will be said to contain A_0 objects. Thus the set of even integers contains A_0 elements, for we can set up the correspondence

$$2, 4, 6, 8, 10, \cdots, 2n, \cdots$$
$$1, 2, 3, 4, 5, \cdots, n, \cdots$$

The set of numbers .5, 1, 1.5, 2, 2.5, . . . also has A_0 elements, since we can set up the correspondence

$$\begin{array}{l} \frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \frac{4}{2}, \frac{5}{2}, \cdots, \frac{n}{2}, \cdots \\ 1, 2, 3, 4, 5, \cdots, n, \cdots \end{array}$$

The set of unreduced proper fractions is also denumerable, since we may set up the correspondence

$$\begin{array}{l} \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \cdots, \frac{j}{r+1}, \cdots \\ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \cdots, n, \cdots \end{array}$$

where r is the largest integer for which $r(r-1)/2 < n$ and

$$j = n - \frac{r(r-1)}{2}$$

Thus for $n = 9$, we have $r = 4$, $j = 3$.

This last example shows that the number of rational numbers (fractions) on the interval zero to one is at most a denumerable set. Actually, in our sequence, every reduced fraction is counted A_0 times. Thus $\frac{2}{3}$, for example, appears as

$$\frac{2}{3}, \frac{4}{6}, \frac{6}{9}, \frac{8}{12}, \cdots, \frac{2n}{3n}, \cdots$$

which is obviously a denumerable set. In the theory of sets, it is shown that every infinite subset of a denumerable set is also denumerable. This theorem together with our last example shows that the number of rational points on the interval zero to one is a denumerable set. It can also be shown that the number of rational points on the whole x axis is denumerable.

The total number of points on a finite interval, say the interval from zero to one on the x axis, is called a *continuous infinity*. This infinity is very much larger than a denumerable infinity and will be denoted by A_1. We shall not prove that A_1 is larger than A_0, but it becomes reasonable when we attempt to count the points on the unit interval. Every point on the unit interval may be represented by an infinite decimal. Thus the point $\frac{1}{3}$ may be represented by

$$.33333 \cdots$$

and $\frac{1}{4}$ may be represented by

$$.2500000 \cdots \quad \text{or by} \quad .2499999 \cdots$$

Conversely every infinite decimal corresponds to a distinct point on the unit interval. We can count the number of possible decimal expansions as follows: The first place can be filled in 10 ways, the second in 10 ways, the third in 10 ways, and so forth. The first n places can therefore be filled in 10^n different ways. The number of infinite decimal sequences is therefore 10^{A_0}, since there are A_0 places in the sequence. When we compare 10^5 with 5, 10^{20} with 20, 10^{1000} with 1000, it becomes reasonable to suppose that 10^{A_0} is of an entirely different order from A_0. This number, 10^{A_0}, is A_1. Actually there are more decimal expansions than points, because of certain duplications, as illustrated above for the point at $\frac{1}{4}$, but these duplications are denumerable and may be neglected relative to A_1. Any finite number n raised to the power A_0 can be shown to be equal to any other raised

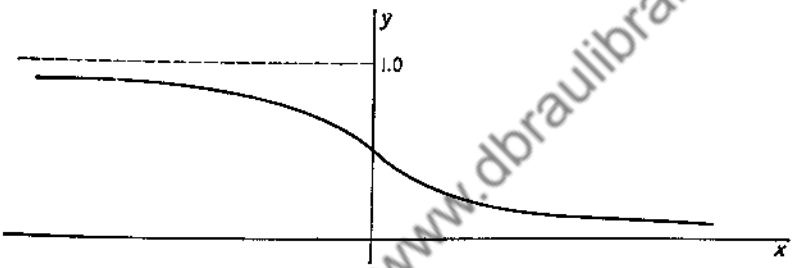


FIG. 12.

to that power. Since the number of points on the unit interval, A_1, satisfies the relation

$$9^{A_0} \leq A_1 \leq 10^{A_0}$$

it follows that $A_1 = 10^{A_0}$ since $9^{A_0} = 10^{A_0}$. The equality sign here is used to mean one-to-one correspondence.

We can easily show that the number of points on the whole x axis is A_1. We may set up a correspondence by means of the function

$$y = \frac{1}{x+2} \qquad \text{if } x \geq 0$$

$$= 1 - \frac{1}{2-x} \qquad \text{if } x \leq 0$$

which is plotted in Fig. 12. Corresponding to every value of x there is a unique value of y between zero and one, and conversely there is a unique value of x for every value of y between zero and one. Thus we have a one-to-one correspondence between the points on the infinite x axis and the unit interval on the y axis. The number of points on

the x axis is therefore A_1. It can also be shown that the number of points in any finite interval, however large or small, is A_1. The correspondence is set up as in Fig. 13. Let I and J be any two intervals of different lengths, and let P be the point of intersection of two lines joining their end points as illustrated. Any point x of I is made to correspond to the point y of J which lies on the line joining x and P. Thus any interval can be related to the unit interval.

Even more bizarre results than these could be obtained by pursuing the theory of sets further. Thus, for example, the number of points in a finite or infinite plane is also A_1. But we have enough results for our immediate purposes. The important idea is the distinction between the two infinities—denumerable and continuous. There are a denumerable infinity of rational points in any interval, but the total

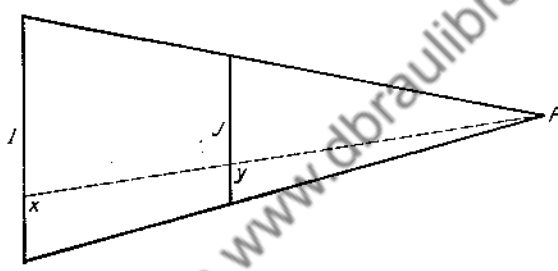


FIG. 13.

number of points is A_1, and the number of rational points is entirely negligible relative to the total number. We could remove all the rational points and essentially the whole interval would still remain.

We can now distinguish precisely between discrete and continuous variates. A discrete variate is one which can take on a finite number of values or a denumerable infinity of values. A continuous variate is one which can take on a continuous infinity of values.

4.2. Probability Functions for Continuous Variates. In the case of discrete variates it is possible to have a finite probability associated with each admissible point, even when the number of points is infinite, and yet have the sum of the probabilities equal to one. Thus if x is the number of tosses required to obtain a head with a coin, we have seen that the density of x is

$$f(x) = (\tfrac{1}{2})^x \qquad x = 1, 2, 3, 4, \cdots$$

and

$$\sum_{x=1}^{\infty} f(x) = 1$$

In the case of a continuous variate this is not possible. No matter how rapidly we try to make the probabilities converge to zero, their sum will nevertheless be infinite unless practically all the points (all but a denumerable set) are given probability zero. Referring back to the horizontal deviations of rifle shots on a target, it is clear that all values of x within a small interval will be about equally likely, and it cannot reasonably be assumed that most of these points have probability zero while some few others have finite probabilities.

We have encountered a difficulty which, it is to be pointed out, is purely logical. From a practical point of view the difficulty is obscured by the fact that we could not actually distinguish between a deviation of .5 inch and one of .500003 inch. We are limited by the accuracy of whatever measuring device we use, and a deviation can be identified only within a certain interval. Thus if we can measure only to within a hundredth of an inch, we might measure a deviation to be 4.26 inches. This would be interpreted to mean that the deviation lies somewhere in the interval 4.25 to 4.27 inches and might better be written $4.26 \pm .01$ to indicate this fact.

The logical problem is met by dealing with intervals rather than individual points. Let us first examine some empirical probabilities for intervals. Suppose the rifle is fired 100 times at the target of Fig. 11, and suppose the target area is divided into strips by drawing vertical lines on it 1 inch apart. Letting the deviations x be negative to the left of the central line, suppose the vertical lines are drawn at $x = \pm 1, \pm 2, \pm 3$, and so on. Now for a given strip, say the one with $0 < x < 1$, the number of shots in that strip divided by 100 will be the empirical probability that a deviation will be between zero and one. We may tabulate a hypothetical distribution of shots and compute the empirical probabilities as in the accompanying table. The

Strip	Number of shots	Empirical probability
$-5 < x < -4$	1	.01
$-4 < x < -3$	1	.01
$-3 < x < -2$	6	.06
$-2 < x < -1$	13	.13
$-1 < x < 0$	24	.24
$0 < x < 1$	27	.27
$1 < x < 2$	16	.16
$2 < x < 3$	7	.07
$3 < x < 4$	3	.03
$4 < x < 5$	2	.02

empirical distribution represented by this table could be plotted by using vertical lines as was done with discrete distributions. However, we shall not plot a line at say the mid-point of each interval but shall prefer to use a rectangle with height equal to the probability divided by the width of the interval, and with a width equal to the width of the interval. This is done to indicate that the probability refers to the whole interval rather than to any single point in the interval. The result is shown in Fig. 14.

Referring to Fig. 14, we note that the area of one of the rectangles is equal to the empirical probability for the interval corresponding to it, since the height of the rectangle is equal to the probability and the base is one. We shall focus attention on the areas rather than the heights. The sum of the areas of all the rectangles is one. For

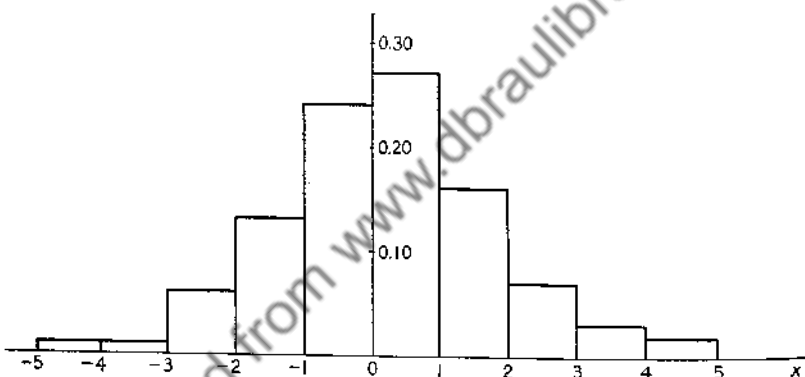


FIG. 14.

intervals other than those chosen originally, we may also estimate probabilities. Thus we would estimate the probability that $0 < x < 2$ by adding the areas of the two rectangles over that interval to get .43. To estimate the probability that, say, $-.25 < x < 1.5$, we would compute the area over that interval to get

$$.06 + .27 + .08 = .41$$

If a second 100 shots were fired at the target, we could obtain another empirical distribution, which would in all likelihood be different from the first though its general appearance might be similar. In constructing a theory of probability, we like to think of these empirical probabilities as being estimates of some "true" probability. To this end we assume the existence of a curve $f(x)$ such as that plotted in Fig. 15. We may not be able to specify the function, but we assume

that there is some function which will give the correct probability for any interval. The probabilities are given by areas under the curve, not by values of the function. Thus

$$P(0 < x < 1) = \int_0^1 f(x)dx$$

and this is the probability that is estimated by the area of the rectangle over the interval $0 < x < 1$ in Fig. 14.

The function $f(x)$ is thought of as a smooth curve rather than a step function for the following reasons: In the first place it is recognized that the choice of intervals in any actual experiment is purely arbitrary. In the rifle experiment we could just as well have used intervals ½ inch

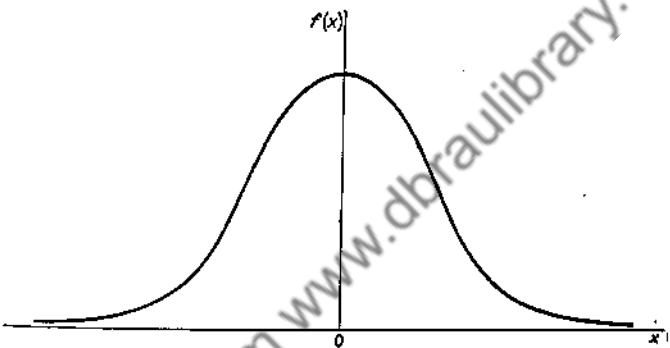


FIG. 15.

long, or intervals with end points at 1.2, 2.2, 3.2, for example, or we could have used intervals of different lengths—0 to .5, .5 to 1.5, 1.5 to 3, for example. So the steps of the empirical distribution have no particular significance. In the second place, suppose we consider two small intervals at a division point, say $1.9 < x < 2$ and $2 < x < 2.1$. Since the second interval is farther removed from center than the first, we should expect its probability to be somewhat smaller, but it is not reasonable to suppose a deviation is more than twice as likely to appear in the first interval, as is indicated in Fig. 14. The smooth curve gives a more reasonable relation between the two probabilities. In the third place, experiments with a large number of trials usually indicate that there are no abrupt changes in the distribution curve. Thus if the rifle were fired, say, 1000 times, and if intervals $\frac{1}{10}$ inch wide were used, the steps would likely be much smaller than those of Fig. 14 and approximate a smooth curve.

In general, a probability density function for a continuous variate will be a function $f(x)$ defined over the range of the variate, and the

range may be finite or infinite. It is often convenient to think of the variate as always having an infinite range; when the range is actually finite, $f(x)$ may be defined to be zero outside the range. The function must be positive or zero, and the area under the curve must be one. Symbolically, the requirements for a density function are

(a) $$f(x) \geq 0$$

(b) $$\int_{-\infty}^{+\infty} f(x)dx = 1$$

The probability that the variate x falls in any interval $a < x < b$ is given by the integral

$$P(a < x < b) = \int_a^b f(x)dx$$

Since the area over a point is zero (a geometric line has no area), it is customary to define the probability that x has any particular value to be zero. We may, in fact, argue that the probability is zero as follows: To compute the probability that x will be some number a, let us find the probability for a small interval of width $2c$ about a:

$$P(a - c < x < a + c) = \int_{a-c}^{a+c} f(x)dx$$

The integral is equal to $2cf(a')$ where a' is a properly chosen point in the interval $a - c$ to $a + c$. (A point a' is determined by constructing a rectangle of area $\int_{a-c}^{a+c} f(x)dx$ over the interval. The top side of the rectangle will intersect the curve $f(x)$ at one or more points if the curve is continuous, as we suppose it is. Any one of these points may be chosen as a'. a' is obviously dependent on c and will approach a as c approaches zero.) Now we shall let c approach zero and define

$$\begin{aligned} P(x = a) &= \lim_{c \to 0} P(a - c < x < a + c) \\ &= \lim_{c \to 0} 2cf(a') = 0 \end{aligned}$$

We have defined an interval by the expression $a < x < b$, but we could equally well have used $a \leq x < b$ or $a < x \leq b$ or $a \leq x \leq b$ without changing the probability associated with the interval. A matter of one or two points does not change the probability for a continuous variate because the probability associated with a single point is zero. In fact, a denumerable set of points could be omitted from the interval without affecting the probability associated with it.

In specific ideal situations, we may be able to say what the exact function $f(x)$ is, just as we did in dealing with a priori probabilities. But in practical situations, $f(x)$ will ordinarily be unknown.

Any positive function over any arbitrarily chosen range may be regarded as a density function for some hypothetical variate over that range, provided the function is multiplied by a constant which will make the integral of the function over the range equal to one. Thus, $3 + 2x$, for example, may be made a density function over the range $2 < x < 4$. Since

$$\int_2^4 (3 + 2x)dx = 18$$

the following function is a density function:

$$\begin{aligned} f(x) &= 0 && x < 2 \\ &= \tfrac{1}{18}(3 + 2x) && 2 < x < 4 \\ &= 0 && x > 4 \end{aligned}$$

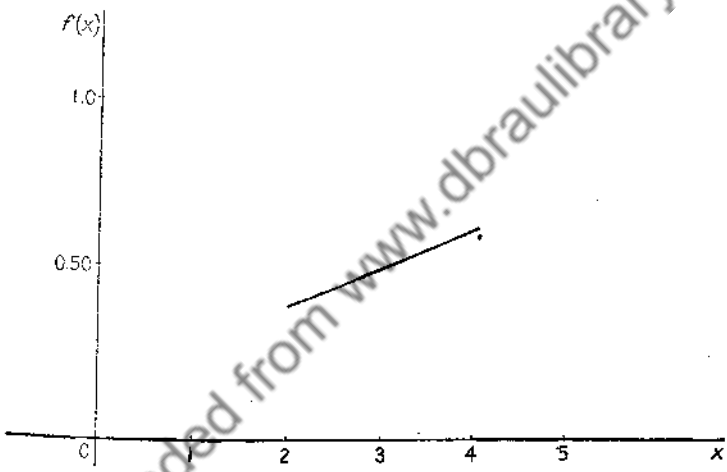


FIG. 16.

The function is obviously positive or zero, and

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{2} 0\,dx + \int_2^4 \tfrac{1}{18}(3 + 2x)dx + \int_4^{\infty} 0\,dx \\ &= 0 + 1 + 0 \\ &= 1 \end{aligned}$$

The probability that a variate having this density will fall in the interval $2 < x < 3$, for example, is

$$\begin{aligned} P(2 < x < 3) &= \int_2^3 \tfrac{1}{18}(3 + 2x)dx \\ &= \tfrac{4}{9} \end{aligned}$$

The function is plotted in Fig. 16.

4.3. Multivariate Distributions. Going back to the rifle experiment, we may characterize each shot not only by its horizontal deviation x but by its vertical deviation y measured perpendicularly from a horizontal line through the center of the target. Suppose a large number of shots are fired, and suppose the target is divided into 1-inch squares by means of horizontal and vertical lines 1 inch apart. We could count the number of hits in each square and compute an empirical probability for each square. By plotting columns with heights equal to the empirical probabilities over each square, we might get a

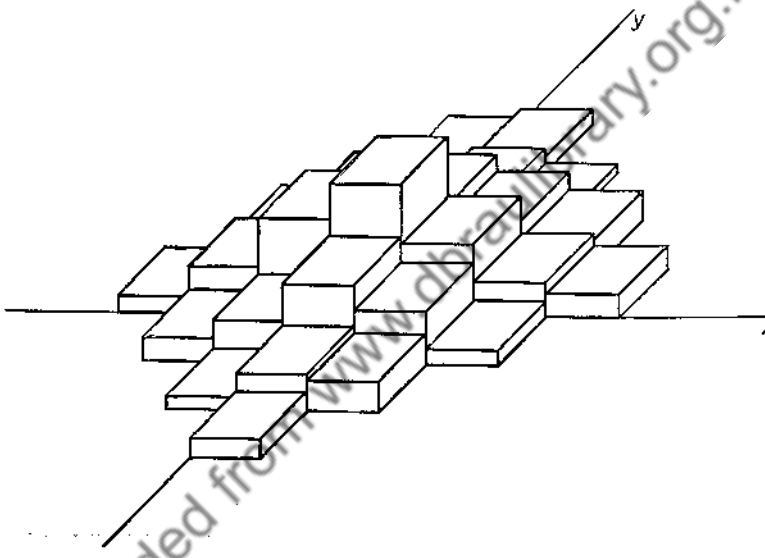


FIG. 17.

result like that illustrated in Fig. 17. The volume of a column estimates the probability that a shot will fall in the square over which the column is constructed.

We shall naturally idealize this situation by postulating the existence of a function $f(x, y)$ which would plot as a smooth surface over the x, y plane. The probability that a shot falls in a given region is represented by the volume under the surface over that region. One quarter of such a surface is illustrated in Fig. 18. The probability that x and y fall in the rectangular region $0 < x < a$, $0 < y < b$ illustrated in the figure is

$$P(0 < x < a, 0 < y < b) = \int_0^a \int_0^b f(x, y)dy\, dx \tag{1}$$

As in the case of one variable, we require

$$f(x, y) \geq 0 \tag{2}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dy\, dx = 1 \tag{3}$$

The function $f(x, y)$ is called the *joint density function* for x and y.

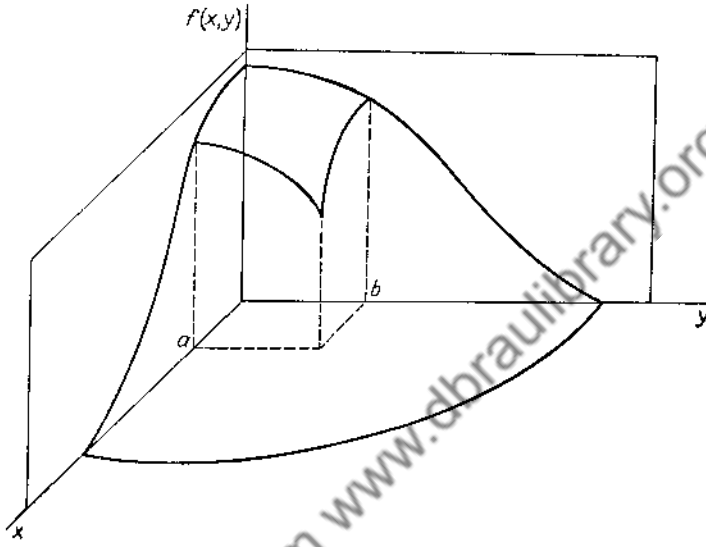


FIG. 18.

As an illustration, the function $6 - x - y$ is positive over the rectangle $0 < x < 2$, $2 < y < 4$, for example; hence it may be used to define a joint density function over that region. Since

$$\int_0^2 \int_2^4 (6 - x - y)dy\, dx = 8$$

The following is a density function:

$$\begin{aligned} f(x, y) &= \tfrac{1}{8}(6 - x - y) \qquad 0 < x < 2,\ 2 < y < 4 \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{4}$$

If x and y are random variables having this density, the probability that they will fall in the region $x < 1$, $y < 3$, for example, is

$$\begin{aligned} P(x < 1, y < 3) &= \int_{-\infty}^{1} \int_{-\infty}^{3} f(x, y)dy\, dx \\ &= \int_0^1 \int_2^3 \tfrac{1}{8}(6 - x - y)dy\, dx \\ &= \tfrac{3}{8} \end{aligned}$$

The probability that $x + y$ will be less than three is

$$P(x + y < 3) = \int_0^1 \int_2^{3-x} \tfrac{1}{8}(6 - x - y)dy\, dx$$
$$= \tfrac{5}{24}$$

The probability that $x < 1$ when it is known that $y < 3$ is

$$P(x < 1|y < 3) = \frac{P(x < 1,\ y < 3)}{P(y < 3)}$$

We have already computed the numerator of this expression, and the denominator is

$$P(y < 3) = \int_0^2 \int_2^3 \tfrac{1}{8}(6 - x - y)dy\, dx$$
$$= \tfrac{5}{8}$$

hence

$$P(x < 1|y < 3) = \frac{\frac{3}{8}}{\frac{5}{8}} = \frac{3}{5}$$

The extension of these ideas to the case of more than two variates is apparent. In general, any function $f(x_1, x_2, \cdots, x_k)$ may be regarded as a density function of k random variables, provided that

$$f(x_1, x_2, \cdots, x_k) \geq 0 \tag{5}$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_k)dx_1\, dx_2 \cdots dx_k = 1$$

The probability that a point $(x_1, x_2, \cdots, x_k)$ falls in any given region of the k-dimensional space is obtained by integrating the density function over that region.

The function

$$f(x_1, x_2, x_3, x_4) = 16x_1x_2x_3x_4 \qquad 0 < x_i < 1 \tag{6}$$
$$= 0 \qquad \text{otherwise}$$

is a density function since it satisfies the two requirements. The probability that a point falls in the region $x_1 < \frac{1}{2}$, $x_4 > \frac{1}{3}$ is

$$P(x_1 < \tfrac{1}{2}, x_4 > \tfrac{1}{3}) = \int_{\frac{1}{3}}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{1}{2}} f(x_1, x_2, x_3, x_4)dx_1\, dx_2\, dx_3\, dx_4$$
$$= \int_{\frac{1}{3}}^{1} \int_0^1 \int_0^1 \int_0^{\frac{1}{2}} 16x_1x_2x_3x_4\, dx_1\, dx_2\, dx_3\, dx_4$$
$$= \tfrac{2}{9}$$

4.4. Cumulative Distributions. Since in the case of continuous variates the probabilities are given by integrals, it is often convenient

to deal with the integrals of the densities rather than the densities themselves. Let $f(x)$ be a density function for one variate (such as is plotted in Fig. 15, for example) and let

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{1}$$

This function $F(x)$ is the probability that the value of an observation will be less than x. Thus

$$F(a) = P(x < a) \tag{2}$$

$F(x)$ is called the *cumulative distribution function* of x, or simply the *cumulative distribution*. The graph of a cumulative distribution

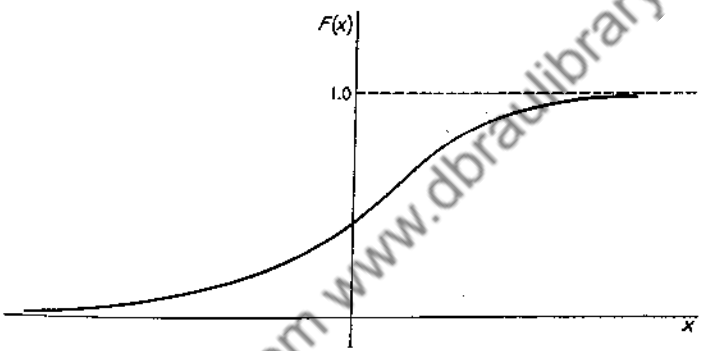


FIG. 19.

function is illustrated in Fig. 19. Any function $F(x)$ may be regarded as the cumulative distribution of a random variable, provided that

$$F(x) \text{ is a nondecreasing function} \tag{3}$$

$$F(-\infty) = 0 \tag{4}$$

$$F(\infty) = 1 \tag{5}$$

and given the cumulative distribution, one can find the density by differentiating it:

$$f(x) = \frac{dF(x)}{dx} \tag{6}$$

The probability that x falls in an interval $a < x < b$ is, in terms of the cumulative distribution,

$$\begin{aligned} P(a < x < b) &= P(x < b) - P(x < a) \\ &= F(b) - F(a) \end{aligned} \tag{7}$$

Referring to the example at the end of Sec. 2, where

$$f(x) = \tfrac{1}{18}(3 + 2x) \qquad 2 < x < 4$$
$$= 0 \qquad \text{otherwise}$$

we find

$$F(x) = 0 \qquad x < 2$$
$$= \int_2^x \tfrac{1}{18}(3 + 2x)dx = \tfrac{1}{18}(x^2 + 3x - 10) \qquad 2 < x < 4$$
$$= 1 \qquad x > 4$$

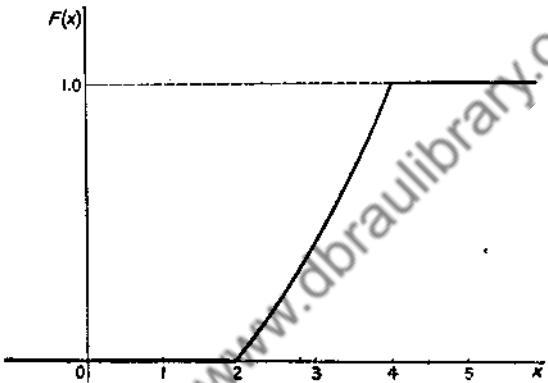


FIG. 20.

and the probability is

$$P(2 < x < 3) = F(3) - F(2)$$
$$= \tfrac{1}{18}(9 + 9 - 10) - 0$$
$$= \tfrac{4}{9}$$

The function is plotted in Fig. 20.

For several variates the cumulative distribution is defined similarly:

$$F(x_1, x_2, \cdots, x_k) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_k} f(t_1, t_2, \cdots, t_k)dt_k\, dt_{k-1} \cdots dt_1 \qquad (8)$$

where $f(x_1, x_2, \cdots, x_k)$ is the density. The value of the cumulative distribution at the point $(a_1, a_2, \cdots, a_k)$, for example, is the probability

$$P(x_1 < a_1, x_2 < a_2, \cdots, x_k < a_k) = F(a_1, a_2, \cdots, a_k) \qquad (9)$$

Any function $F(x_1, x_2, \cdots, x_k)$ may be regarded as a cumulative distribution of k variates, provided that

$$F(x_1, x_2, \cdots, x_k) \text{ is nondecreasing in every variate} \tag{10}$$

$$F(\infty, \infty, \cdots, \infty) = 1 \tag{11}$$

$$F(x_1, \cdots, -\infty, \cdots, x_k) = 0 \tag{12}$$

and this last condition is intended to indicate that F vanishes if any one of the variates approaches minus infinity. Given the cumulative distribution F, the density may be found by differentiating F with respect to each of its variates:

$$f(x_1, x_2, \cdots, x_k) = \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}\cdots\frac{\partial}{\partial x_k}F(x_1, x_2, \cdots, x_k) \tag{13}$$

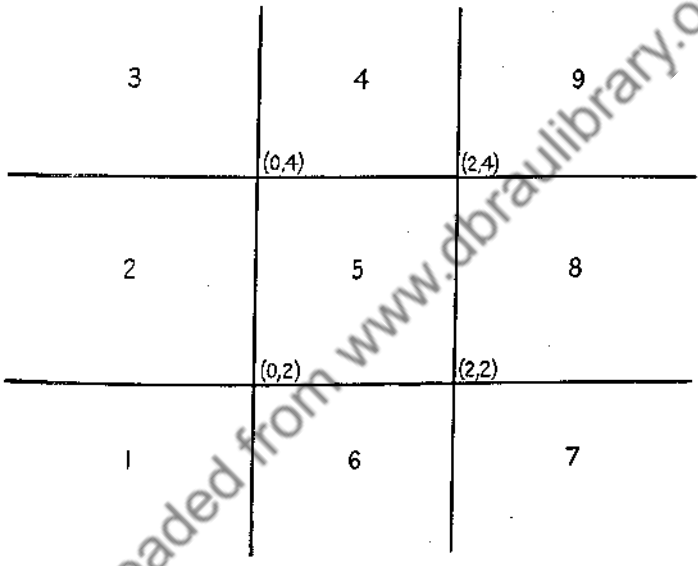


FIG. 21.

To illustrate a cumulative distribution for two variates, we may use the density given in equation (3.4):

$$\begin{aligned} f(x, y) &= \tfrac{1}{8}(6 - x - y) \qquad 0 < x < 2,\ 2 < y < 4 \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{14}$$

There are nine regions in the x, y plane to be taken account of in defining $F(x, y)$; the nine regions are indicated in Fig. 21, in which the coordinates of the points of intersection of the lines are given. (The left vertical line coincides with the y axis.) This complication arises because of the piecewise definition of $f(x, y)$. We could simply state

that

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)dt\, ds \tag{15}$$

but a more detailed characterization of the function will be required if it is to be useful. In region 1 of Fig. 21 $f(x, y)$ is zero; hence

$$F(x, y) = 0 \qquad x < 0,\ y < 2$$

In region 2, although y is greater than two, we have $x < 0$, so that (15) is still zero since $f(s, t)$ never becomes positive over the range of integration. The same is true in regions 3, 6, 7. For x, y in region 5, the integrand is not zero when $0 < s < x$, $2 < t < y$, and we have

$$\begin{aligned} F(x, y) &= \int_0^x \int_2^y \tfrac{1}{8}(6 - s - t)dt\, ds \\ &= \int_0^x \frac{1}{8}\left[(6 - s)(y - 2) - \frac{y^2}{2} + 2\right] ds \\ &= \tfrac{1}{16}x(y - 2)(10 - y - x) \qquad 0 < x < 2,\ 2 < y < 4 \end{aligned} \tag{16}$$

For any point in region 4, the integrand in (15) is positive when $0 < s < x$, $2 < t < 4$; hence

$$F(x, y) = \int_0^x \int_2^4 f(s, t)dt\, ds$$

and this integral may be computed by putting $y = 4$ in (16) to get

$$F(x, y) = \tfrac{1}{8}x(6 - x) \qquad 0 < x < 2,\ y > 4$$

Similarly, in region 8, $F(x, y) = F(2, y)$ when $x > 2$, so that

$$F(x, y) = \tfrac{1}{8}(y - 2)(8 - y) \qquad x > 2,\ 2 < y < 4$$

and in region 9, $F(x, y) = 1$. Combining these results,

$$\begin{aligned} F(x, y) &= 0 && x < 0 \text{ or } y < 2 \\ &= \tfrac{1}{16}x(y - 2)(10 - y - x) && 0 < x < 2,\ 2 < y < 4 \\ &= \tfrac{1}{8}x(6 - x) && 0 < x < 2,\ y > 4 \\ &= \tfrac{1}{8}(y - 2)(8 - y) && x > 2,\ 2 < y < 4 \\ &= 1 && x > 2,\ y > 4 \end{aligned} \tag{17}$$

The function is plotted in Fig. 22.

The probability that a point (x, y) will fall in any rectangle, say $a_1 < x < b_1$, $a_2 < y < b_2$, may be written in terms of the cumulative distribution as follows:

$$\begin{aligned} P(a_1 < x < b_1,\ a_2 < y < b_2) &= P(x < b_1,\ y < b_2) - P(x < a_1,\ y < b_2) \\ &\quad - P(x < b_1,\ y < a_2) \\ &\quad + P(x < a_1,\ y < a_2) \\ &= F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) \\ &\quad + F(a_1, a_2) \qquad (18) \end{aligned}$$

Thus, in the above example,

$$\begin{aligned} P(0 < x < 1,\ 3 < y < 4) &= F(1, 4) - F(0, 4) - F(1, 3) + F(0, 3) \\ &= \tfrac{5}{8} - 0 - \tfrac{3}{8} + 0 \\ &= \tfrac{1}{4} \end{aligned}$$

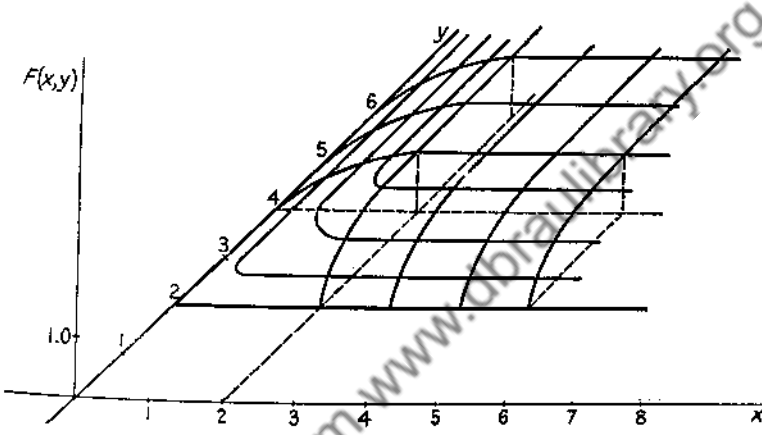


FIG. 22.

These distributions can become quite complex for several variables, and in fact many important problems in applied statistics remain unsolved merely because the integrations required for their solution are too complex to perform. Modern developments in high-speed computing machines promise to remedy this situation within the next few years.

In this book we shall ordinarily use small letters to denote probability density functions and the corresponding capital letters to represent their cumulative forms. Thus,

$$G(x) = \int_{-\infty}^{x} g(t)dt$$

or if the variate is discrete,

$$G(x) = \sum^{x} g(t)$$

The word *density* will refer specifically to $g(x)$, while the phrase *cumulative distribution* will refer specifically to $G(x)$. The word *distribution*

will be used as a more general term and may refer to either the density or its cumulative form.

4.5. Marginal Distributions. Associated with any distribution of more than one variable are several marginal distributions. Let $f(x, y)$ be a density for two continuous variates. We may be interested in only one of the variates, say x. We therefore seek a function of x which when integrated over an interval, say $a < x < b$, will give the probability that x will lie in that interval. In the x, y plane such an

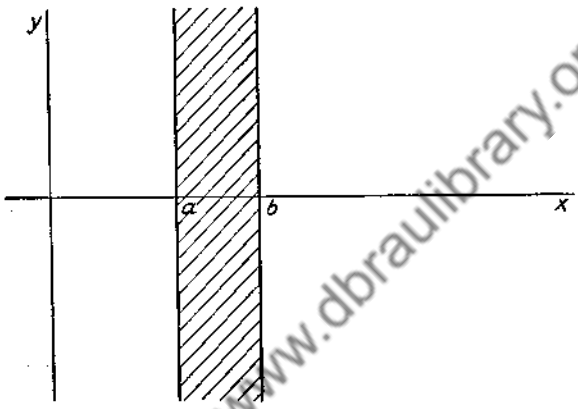


FIG. 23.

interval corresponds to a strip as illustrated in Fig. 23. The specification $a < x < b$ is satisfied by any point in the strip; hence

$$P(a < x < b) = \int_a^b \int_{-\infty}^{\infty} f(x, y)dy\, dx \tag{1}$$

Whatever the specification on x, the limits of integration for y are $-\infty$ to $+\infty$, so we may define a function, say

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)dy \cdot \tag{2}$$

and this function is the required marginal density, since

$$P(a < x < b) = \int_a^b f_1(x)dx \tag{3}$$

for any pair of values a and b. Similarly the marginal density of y is

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)dx \tag{4}$$

In general, given any density $f(x_1, x_2, \cdots, x_k)$, one may find the marginal density of any subset of the variates by integrating the func-

tion with respect to all the other variates between the limits $-\infty$ and $+\infty$. Thus the marginal density of x_1, x_2, and x_4, for example, is

$$f_{124}(x_1, x_2, x_4) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_k)dx_3\, dx_5\, dx_6 \cdots dx_k \quad (5)$$

Referring to the distribution defined in equation (3.4), the marginal density of x is

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x, y)dy & -\infty < x < \infty \\ &= \int_2^4 \tfrac{1}{8}(6 - x - y)dy & 0 < x < 2 \\ &= \tfrac{1}{4}(3 - x) & 0 < x < 2 \\ &= 0 & x < 0 \text{ or } x > 2 \end{aligned} \quad (6)$$

The cumulative marginal distribution is easily found if the cumulative distribution is given. For two variables, the cumulative marginal distribution of x is

$$\begin{aligned} F_1(x) &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x, y)dy\, dx = \int_{-\infty}^{x} f_1(x)dx \\ &= F(x, \infty) \end{aligned} \quad (7)$$

Thus we need only let the variable in which we are not interested become infinite in the joint cumulative distribution. And in general, if $F(x_1, x_2, \cdots, x_k)$ is a k-variate cumulative distribution, the cumulative marginal distribution of x_1, x_2, x_4, for example, is

$$F_{124}(x_1, x_2, x_4) = F(x_1, x_2, \infty, x_4, \infty, \cdots, \infty) \quad (8)$$

In our specific example we may find the cumulative marginal distribution of x by integrating $f_1(x)$; thus

$$\begin{aligned} F_1(x) &= \int_{-\infty}^{x} f_1(t)dt \\ &= 0 & x < 0 \\ &= \tfrac{1}{8}x(6 - x) & 0 < x < 2 \\ &= 1 & x > 2 \end{aligned} \quad (9)$$

The same result is obtained by letting y become infinite in $F(x, y)$ given by equations (4.17).

4.6. Conditional Distributions. We shall consider first a bivariate density, say $f(x, y)$, which might be represented by the surface of Fig. 18, for example. Suppose a point (x, y) is drawn (a shot is fired

at a target, for example), and suppose the second variate y is observed but not the first. We seek a function, say $f(x|y)$, which will give the density of x when y is known; i.e., a function such that

$$P(a < x < b|y) = \int_a^b f(x|y)dx \tag{1}$$

for any arbitrarily chosen a and b.

If we change the above problem so that it concerns probabilities rather than distributions of continuous variates, we may use the definition of conditional probability given in Sec. 2.7. Thus we may compute (assuming $c > 0$)

$$P(a < s < b|y - c < t < y + c) = \frac{\int_a^b \int_{y-c}^{y+c} f(s, t)dt\, ds}{\int_{y-c}^{y+c} \int_{-\infty}^{\infty} f(s, t)ds\, dt} \tag{2}$$

The denominator may be written in terms of the marginal density of y, say $f_2(y)$, as

$$\int_{y-c}^{y+c} f_2(t)dt$$

and this is equal to $2cf_2(y')$, where y' is some value in the interval $y - c$ to $y + c$. Similarly the numerator of (2) is equal to

$$2c\int_a^b f(s, y'')ds$$

where y'' is some point in the interval $y - c$ to $y + c$. Hence the probability is

$$P(a < x < b|y - c < t < y + c) = \frac{\int_a^b f(s, y'')ds}{f_2(y')} \tag{3}$$

Now we shall let c approach zero. Since y', y'', and t are all in the interval $y - c$ to $y + c$ and must remain in the interval however small c becomes, it follows that they must all approach y. Hence the limit of (3) as c becomes zero is

$$P(a < x < b|y) = \frac{\int_a^b f(s, y)ds}{f_2(y)} \tag{4}$$

Since this relation holds true for any a and b, it follows that

$$f(x|y) = \frac{f(x, y)}{f_2(y)} \tag{5}$$

By similar reasoning, if $f_1(x)$ is the marginal density of x, the conditional density of y given x is

$$f(y|x) = \frac{f(x, y)}{f_1(x)} \tag{6}$$

The function $f(x|y)$ is a function of one variate x; y is simply a parameter and will have some numerical value in any specific conditional density. Thus $f_2(y)$ is to be regarded as a constant. The joint density $f(x, y)$ plots as a surface over the x, y plane. A plane perpendicular to the x, y plane which intersects the x, y plane on the line $y = c$ will intersect the surface in the curve $f(x, c)$. The area under this curve is

$$\int_{-\infty}^{\infty} f(x, c)dx = f_2(c)$$

hence if we divide $f(x, c)$ by $f_2(c)$, we obtain a density function which is precisely $f(x|c)$.

For the particular function

$$\begin{aligned} f(x, y) &= \tfrac{1}{8}(6 - x - y) && 0 < x < 2,\ 2 < y < 4 \\ &= 0 && \text{otherwise} \end{aligned}$$

we have found in the preceding section that the marginal density of x is

$$\begin{aligned} f_1(x) &= \tfrac{1}{4}(3 - x) && 0 < x < 2 \\ &= 0 && \text{otherwise} \end{aligned}$$

In view of (6) the conditional density of y for fixed x is therefore

$$f(y|x) = \frac{6 - x - y}{2(3 - x)} \qquad 2 < y < 4$$

Conditional distributions are defined analogously for multivariate distributions. Thus for five variates with a density $f(x_1, x_2, x_3, x_4, x_5)$, the conditional density of x_1, x_2, x_4, given specific values of x_3 and x_5, is

$$f(x_1, x_2, x_4|x_3, x_5) = \frac{f(x_1, x_2, x_3, x_4, x_5)}{f_{35}(x_3, x_5)}$$

where $f_{35}(x_3, x_5)$ represents the marginal density of x_3 and x_5.

4.7. Independence. If the conditional density $f(x|y)$ does not involve y and if the range of the conditional density does not depend on y, then x is independent of y in the probability sense. Suppose that this is the case and that we represent $f(x|y)$ by $g(x)$. Since, from Sec. 6,

$$f(x|y) = g(x) = \frac{f(x, y)}{f_2(y)} \tag{1}$$

it follows that

$$f(x, y) = g(x)f_2(y) \tag{2}$$

hence the joint density of x and y is the product of two functions, one involving x only and the other involving y only. If we integrate (2) with respect to y over the whole range of y, we find that $g(x)$ is simply the marginal density of x. Thus we may state:

If two variates x and y are independent in the probability sense, then their joint distribution is equal to the product of their marginal distributions.

The converse of this statement is also true. That is, if $f(x, y)$ can be factored into two functions, one involving x only and the other involving y only, and if the ranges of x and y do not depend on each other, then x and y are independent in the probability sense.

In general, if the conditional distribution of a subset of any set of variates is independent of the remaining fixed variables, then that subset is said to be independent of the remaining variables in the probability sense. The function defined in equation (3.6) provides an illustration:

$$\begin{aligned} f(x_1, x_2, x_3, x_4) &= 16x_1x_2x_3x_4 && 0 < x_i < 1 \text{ for all } i \\ &= 0 && \text{otherwise} \end{aligned}$$

The marginal density of, say, x_2 and x_4 is

$$\begin{aligned} f_{24}(x_2, x_4) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)dx_1\,dx_3 \\ &= 4x_2x_4 && 0 < x_2 < 1,\ 0 < x_4 < 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

Hence the conditional density of x_1 and x_3 is

$$\begin{aligned} f(x_1, x_3|x_2, x_4) &= 4x_1x_3 && 0 < x_1 < 1,\ 0 < x_3 < 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

This function and its range do not involve x_2 and x_4, so that the pair of variables (x_1, x_3) is independent of the pair (x_2, x_4) in the probability sense. In fact, all four variates of this distribution are mutually independent as may be deduced from the fact that the function may be factored into four functions each involving only one of the variates.

4.8. Problems

1. If $f(x) = 2x$ when $0 < x < 1$ and zero otherwise, find the probability that (*a*) $x < \frac{1}{2}$; (*b*) $\frac{1}{4} < x < \frac{1}{2}$; (*c*) $x > \frac{3}{4}$ given $x > \frac{1}{2}$.

2. Define a density function using the function $x(2 - x)$ over the range $0 < x < 2$. Find the probability that $a < x < b$ if

$$0 < a < b < 2$$

if $a < 0 < 2 < b$.

3. If $f(x) = 3x^2$ when $0 < x < 1$ and zero otherwise, find the number a such that x is equally likely to be greater than or less than a. Find the number b such that the probability that x will exceed b is equal to .05.

4. A variate x has the density $f(x) = x/2$ when $0 < x < 2$ and zero otherwise. If two values of x are drawn, what is the probability that both will be greater than one? If three are drawn, what is the probability that exactly two will be greater than one?

5. A variate x has the density $f(x) = 1$ when $0 < x < 1$ and zero otherwise. Determine the number a such that the probability will be .9 that at least one of four values of x drawn at random will exceed a.

6. Suppose the life in hours of a certain kind of radio tube has the density $f(x) = 100/x^2$ when $x > 100$ and zero when $x < 100$. What is the probability that none of three such tubes in a given radio set will have to be replaced during the first 150 hours of operation? What is the probability that all three of the original tubes will have been replaced during the first 150 hours?

7. A machine makes bolts with diameters distributed by the density $f(x) = K(x - .24)^2(x - .26)^2$ when $.24 < x < .26$ and zero otherwise. K is the number which makes $\int_{-\infty}^{\infty} f(x)dx = 1$. Bolts must be scrapped if their diameters deviate from .25 by more than .008. What proportion of the bolts may be expected to be scrap?

8. A bombing plane carrying three bombs flies directly above a railroad track. If a bomb falls within 40 feet of the track, the track will be sufficiently damaged to disrupt traffic. With a certain bombsight the density of points of impact of a bomb is

$$\begin{aligned} f(x) &= (100 + x)/10{,}000 && -100 < x < 0 \\ &= (100 - x)/10{,}000 && 0 < x < 100 \\ &= 0 && \text{elsewhere} \end{aligned}$$

x represents the vertical deviation from the aiming point, which is the track in this case. If all three bombs are used, what is the probability that the track will be damaged?

9. Referring to the above problem, the plane can carry eight bombs of a smaller size, but one of these must hit within 15 feet of the track

to damage it. Should the lighter or heavier bombs be used on this mission?

10. A country filling station is supplied with gasoline once a week. If its weekly volume x of sales in thousands of gallons is distributed by $f(x) = 5(1 - x)^4$, $0 < x < 1$, what must be the capacity of its tank in order that the probability that its supply will be exhausted in a given week shall be .01?

11. A batch of small-caliber ammunition is accepted as satisfactory if none of a sample of five shots falls more than 2 feet from the center of a target at a given range. If r, the distance from the target center of a given impact point, actually has the density

$$f(r) = \frac{2re^{-r^2}}{(1 - e^{-9})}$$

$0 < r < 3$, for a given batch, what is the probability that the batch will be accepted?

12. If $f(x, y) = 1$ when $0 < x < 1$, $0 < y < 1$, and zero otherwise, find the probability that (*a*) $x < \frac{1}{2}$, $y < \frac{1}{2}$; (*b*) $x + y < 1$; (*c*) $x + y > 1$; (*d*) $x > 2y$; (*e*) $x > \frac{1}{3}$; (*f*) $x^2 + y^2 < \frac{1}{4}$; (*g*) $x = y$; (*h*) $x > \frac{1}{2}$ given $y < \frac{1}{2}$; (*i*) $x > y$ given $y > \frac{1}{2}$.

13. If $f(x, y) = e^{-(x+y)}$ when $x > 0$, $y > 0$, and zero otherwise, find $P(x > 1)$; $P(a < x + y < b)$ if $0 < a < b$; $P(x < y | x < 2y)$.

14. Using the distribution of Prob. 13, find the number a such that $P(x + y < a) = \frac{1}{2}$.

15. If three points (x, y) are drawn at random where x and y are distributed by the function given in Prob. 13, what is the probability that at least one of them will fall in the square $0 < x < 1$, $0 < y < 1$?

16. A machine makes shafts with diameters x, and a second machine makes bushings with inside diameters y. Suppose the density of x and y is $f(x, y) = 2500$, $.49 < x < .51$, $.51 < y < .53$, and zero otherwise. A bushing fits a shaft satisfactorily if its diameter exceeds that of the shaft by at least .004 but not more than .036. What is the probability that a bushing and shaft chosen at random will fit?

17. Find and plot roughly the cumulative distribution for the distribution given in Prob. 6. Use the cumulative distribution to find $P(150 < x < 250)$.

18. Find and plot roughly the cumulative distribution for the function given in Prob. 13, and use it to find $P(1 < x < 2, 3 < y < 4)$.

19. Find the marginal density of x for the distribution of Prob. 13: (*a*) by integrating out y; (*b*) by using the result of Prob. 18 to get the cumulative marginal distribution, then differentiating the result.

20. Find the conditional density of x given y for the distribution of Prob. 13. What is the $P(0 < x < 1|y = 2)$?

21. If $f(x, y) = (n - 1)(n - 2)/(1 + x + y)^n$ when $x > 0$, $y > 0$, and zero elsewhere, find $F(x, y)$, $f_1(x)$, $F_1(x)$, $f(y|x)$.

22. If $f(x, y) = 24y(1 - x - y)$ over the triangle bounded by the axes and the line $x + y = 1$, find $f(x|y)$.

23. If $f(x, y) = 3x$, $0 < y < x$, $0 < x < 1$, find the conditional density of x.

24. If $f(x|y) = 3x^2/y^3$, $0 < x < y$, and $f_2(y) = 5y^4$, $0 < y < 1$, find $P(x > \frac{1}{2})$.

25. If $f(x, y, z) = 8xyz$, $0 < x < 1$, $0 < y < 1$, $0 < z < 1$, find $P(x < y < z)$.

26. If $f(x) = 1/(1 + x)^2$, $x > 0$, find the density of x given that $x > 1$.

27. If $f(x, y) = 1$, $0 < x < 1$, $0 < y < 1$, find the conditional density of x and y given that $y < x^n$, $n > 0$.

28. If $f(x) = 1$, $0 < x < 1$, find the density of $y = 3x + 1$. (Find first the cumulative distribution of y and then differentiate it.)

29. If $f(x) = 2xe^{-x^2}$, $x > 0$, find the density of $y = x^2$.

30. If $f(x, y) = 1$, $0 < x < 1$, $0 < y < 1$, find the density of $z = x + y$.

31. If $f(x, y) = e^{-(x+y)}$, $x > 0$, $y > 0$, find the density of

$$z = \frac{(x + y)}{2}$$

32. If $f(x, y) = 4xye^{-(x^2+y^2)}$, $x > 0$, $y > 0$, find the density of $z = \sqrt{x^2 + y^2}$.

33. If $f(x, y) = 4xy$, $0 < x < 1$, $0 < y < 1$, find the joint density of $u = x^2$, $v = y^2$.

34. If $f(x, y) = 3x$, $0 < y < x$, $0 < x < 1$, find the density of $z = x - y$.

35. If $f(x) = (1 + x)/2$, $-1 < x < 1$, find the density of $y = x^2$.

36. If $f(x, y) = 1$, $0 < x < 1$, $0 < y < 1$, find the density of z defined by: $z = x + y$ if $x + y < 1$, and $z = x + y - 1$ if $x + y > 1$.

37. If $f(x, y) = e^{-(x+y)}$, $x > 0$, $y > 0$, find the joint density of $u = x + y$ and $v = x$. What is the marginal density of v?

38. If $f(x, y, z) = e^{-(x+y+z)}$, $x > 0$, $y > 0$, $z > 0$, find the density of their average $u = (x + y + z)/3$.

39. If $f(x, y) = 4x(1 - y)$, $0 < x < 1$, $0 < y < 1$, find the density of x given that $y < \frac{1}{2}$.

40. If x is distributed by $f(x)$, $x > 0$, find the density of $y = ax^2 + b$, $a > 0$.

41. If x is distributed by $f(x)$, $-\infty < x < \infty$, and if $y = y(x)$ is any increasing function of x [i.e., $y(x_1) > y(x_0)$ when $x_1 > x_0$], find the density of y.

42. If $f(x, y) = g(x)g(y)$, $x > 0$, $y > 0$, find $P(x > y)$.

43. If $f(x, y, z) = g(x)g(y)g(z)$, $x > 0$, $y > 0$, $z > 0$, what is the probability that the coordinates of a randomly drawn point (x, y, z) will not satisfy either $x > y > z$ or $x < y < z$.

44. In which of the distributions defined in Probs. 21, 22, 23, 24, 31, 32, 33, and 34 are the variates independent in the probability sense?

CHAPTER 5

EXPECTED VALUES AND MOMENTS

5.1. Expected Values. The expected value of a random variable or any function of a random variable is obtained by finding the average value of the function over all possible values of the variable. To consider a specific example: If three coins are tossed, the distribution of the number of heads that appear is the binomial

$$f(x) = \binom{3}{x}\left(\frac{1}{2}\right)^3 \qquad x = 0, 1, 2, 3 \tag{1}$$

For a specific value of x, say $x = 2$, we think of $f(2) = \frac{3}{8}$ as the relative frequency with which two heads will appear in a large number of trials. Thus in 1000 trials we expect no heads to appear in about $1000 \times \frac{1}{8} = 125$ trials, one head to appear in $1000 \times \frac{3}{8} = 375$ trials, two heads in 375 trials, and three heads in 125 trials. Now let us find the average number of heads in the 1000 trials. The total number of heads is expected to be

$$125 \times 0 + 375 \times 1 + 375 \times 2 + 125 \times 3 = 1500$$

in the 1000 trials; thus the average is expected to be 1.5 heads per trial. This is the *expected value*, or *mean value*, of x. It is clear that the same result would have been obtained had we merely multiplied all possible values of x by their probabilities and added the results; thus,

$$0 \times \tfrac{1}{8} + 1 \times \tfrac{3}{8} + 2 \times \tfrac{3}{8} + 3 \times \tfrac{1}{8} = 1.5$$

The expected value is a theoretical or ideal average. We do not actually expect x to take on its expected value in a given trial; in fact that would be impossible in the present example. However, we might reasonably expect the average value of x in a great number of trials to be somewhere near the expected value of x.

These considerations lead us to define in general the expected value of a discrete variate as $\Sigma x f(x)$, where $f(x)$ is the distribution of x and the sum is taken over the whole range of x. The symbol $E(x)$ is used to denote the expected value of x. Thus in the illustrative example

$$E(x) = \sum_{x=0}^{3} xf(x) = 1.5$$

In general, we shall define the expected value of any function of x, say $h(x)$, as

$$E[h(x)] = \sum_{x} h(x)f(x) \tag{2}$$

Where the sum is taken over the whole range of x. Thus if

$$h(x) = x^2 + 1$$

and $f(x)$ is as defined in equation (1),

$$\begin{aligned} E(x^2 + 1) &= \sum_{x=0}^{3} (x^2 + 1)f(x) \\ &= \tfrac{1}{8} + 2 \times \tfrac{3}{8} + 5 \times \tfrac{3}{8} + 10 \times \tfrac{1}{8} = 4 \end{aligned}$$

Similarly for several discrete variates $x_1, x_2, \cdots, x_k$, with distribution $f(x_1, x_2, \cdots, x_k)$, the expected value of any function h of the variates is defined to be

$$\begin{aligned} &E[h(x_1, x_2, \cdots, x_k)] \\ &\quad = \sum_{x_1}\sum_{x_2} \cdots \sum_{x_k} h(x_1, x_2, \cdots, x_k)f(x_1, x_2, \cdots, x_k) \end{aligned} \tag{3}$$

where the sums are taken over the entire range of each variate.

For continuous variates we define expected values in terms of integrals rather than sums. If x has the distribution $f(x)$ and $h(x)$ is any function of x, then

$$E[h(x)] = \int_{-\infty}^{\infty} h(x)f(x)dx \tag{4}$$

This definition is suggested by the definition for discrete variates given in equation (2) together with the definition of a definite integral as the limit of a sum. Let the x axis be divided into intervals of length Δx_i ($i = 0, \pm 1, \pm 2, \cdots$) and let x_i' be a point in the interval Δx_i such that $f(x_i')\Delta x_i$ equals the area under $f(x)$ over Δx_i. Then an expected value of $h(x)$ may be computed by regarding x as a discrete variate which can take on only the values x_i' with the probabilities $f(x_i')\Delta x_i$. This expected value is

$$\sum_{i=-\infty}^{\infty} h(x_i')f(x_i')\Delta x_i$$

according to equation (2). The limit of this sum as all Δx_i approach zero will essentially remove the restriction that x be discrete, and the

limit is the integral given in (4). Similarly for several continuous variates, we define

$$E[h(x_1, x_2, \cdots, x_k)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, x_2, \cdots, x_k)f(x_1, x_2, \cdots, x_k)dx_1 \cdots dx_k \quad (5)$$

We shall avoid confusing the expected-value notation with the functional notation by never using the letter E to represent a function. $E(g)$ will always represent the expected value of g, never a function E of g. In the remainder of this chapter we shall not distinguish between discrete and continuous variates. Expected values will always be given in terms of integrals, but it is to be understood that the integrals are to be replaced by sums in specific problems which deal with discrete variates.

Two simple properties of E are worth noting. If x is distributed by $f(x)$, if c is any constant, and if $g(x)$ and $h(x)$ are any functions of x, then

$$E[cg(x)] = cE[g(x)] \quad (6)$$

$$E[g(x) + h(x)] = E[g(x)] + E[h(x)] \quad (7)$$

These two relations follow directly from the corresponding relations for integrals:

$$\int cg(x)f(x)dx = c\int g(x)f(x)dx$$

$$\int [g(x) + h(x)]f(x)dx = \int g(x)f(x)dx + \int h(x)f(x)dx$$

Of course (6) and (7) remain true if the single variate x is replaced by a set of variates $x_1, x_2, \cdots, x_k$.

5.2. Moments. The moments of a distribution are the expected values of the powers of the random variable which has the given distribution. The rth moment of x is usually denoted by μ_r' and is

$$\mu_r' = E(x^r) = \int_{-\infty}^{\infty} x^r f(x)dx \quad (1)$$

The first moment μ_1' is called the *mean* of x. The moments about any arbitrary point a are defined as

$$E[(x - a)^r] = \int_{-\infty}^{\infty} (x - a)^r f(x)dx \quad (2)$$

and when a is put equal to the mean, we have the moments about the mean, which are usually denoted by μ_r:

$$\mu_r = E[(x - \mu_1')^r] = \int_{-\infty}^{\infty} (x - \mu_1')^r f(x)dx \quad (3)$$

We have

$$\begin{aligned}\mu_1 &= \int_{-\infty}^{\infty} xf(x)dx - \mu_1' \int_{-\infty}^{\infty} f(x)dx \\ &= \mu_1' - \mu_1' = 0 \end{aligned} \tag{4}$$

and

$$\begin{aligned}\mu_2 &= \int_{-\infty}^{\infty} (x - \mu_1')^2 f(x)dx \\ &= \int_{-\infty}^{\infty} [x^2 - 2x\mu_1' + (\mu_1')^2] f(x)dx \\ &= \mu_2' - 2\mu_1'\mu_1' + (\mu_1')^2 \\ &= \mu_2' - (\mu_1')^2 \end{aligned} \tag{5}$$

This second moment about the mean is called the *variance* of x.

The mean value of a variate locates the center of its distribution in the following sense: If the x axis is thought of as a bar with variable density, the density at any point being given by $f(x)$, then it is shown in elementary calculus that the value $x = \mu_1'$ is the center of gravity of the bar. Thus the mean may be thought of as a central value of the variate. For this reason it is often referred to as a location parameter —it tells one where the center of the distribution (in the center-of-gravity sense) lies on the x axis. Other central values are sometimes used to indicate the location of a distribution. One is the *median*, which is defined as the point at which a vertical line bisects the area under the curve $f(x)$. The median is therefore the point μ'', say, such that

$$\int_{-\infty}^{\mu''} f(x)dx = \tfrac{1}{2} = \int_{\mu''}^{\infty} f(x)dx \tag{6}$$

Another central value for densities with one maximum is the *mode*, which is the point at which $f(x)$ attains its maximum. One could easily devise other central values; these are the ones commonly used, and of the three the mean is by far the most useful. We shall often employ the symbol μ without the prime or subscript to denote the mean.

The variance μ_2 of a distribution is a measure of its spread, or dispersion. If most of the area under the curve lies near the mean, the variance will be small; while if the area is spread out over a considerable range, the variance will be large. Distributions with different variances are plotted in Fig. 30 in the following chapter. The variance is necessarily positive, since it is the integral or sum of positive quantities. It will vanish only when the distribution is concentrated at one point, *i.e.*, when the distribution is discrete and there is only one possible outcome. The symbol σ^2 is commonly used to denote the

variance; the positive square root of the variance, σ, is called the *standard deviation*.

We shall look a little further into the manner in which the variance characterizes the distribution. Suppose $f_1(x)$ and $f_2(x)$ are two densities with the same mean such that

$$\int_{\mu-a}^{\mu+a} [f_1(x) - f_2(x)]dx \geq 0 \tag{7}$$

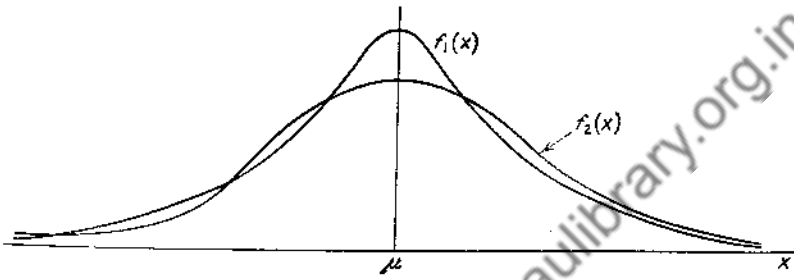


FIG. 24.

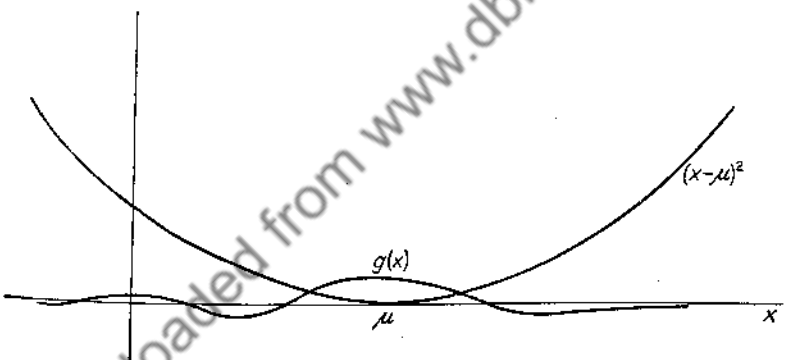


FIG. 25.

for every value of a. Two such densities are illustrated in Fig. 24. It can be shown that in this case the variance σ_1^2 of the first density is smaller than the variance σ_2^2 of the second density. We shall not take the time to prove this in detail, but the argument is roughly this: Let

$$g(x) = f_1(x) - f_2(x)$$

where $f_1(x)$ and $f_2(x)$ satisfy (7). Since $\int_{-\infty}^{\infty} g(x)dx = 0$, the positive area between $g(x)$ and the x axis is equal to the negative area. Furthermore, in view of (7), every positive element of area $g(x')dx'$ may be balanced by a negative element $g(x'')dx''$ in such a way that x''

is farther from μ than x'. When these elements of area are multiplied by $(x - \mu)^2$, the negative elements will be multiplied by larger factors than their corresponding positive elements; hence

$$\int_{-\infty}^{\infty} (x - \mu)^2 g(x) dx < 0$$

unless $f_1(x)$ and $f_2(x)$ are equal. Thus it follows that $\sigma_1^2 < \sigma_2^2$.

The converse of these statements is not true. That is, if one is told that $\sigma_1^2 < \sigma_2^2$, he cannot conclude that the corresponding densities satisfy (7) for all values of a, though it can be shown that (7) must be true for certain values of a. Thus the condition $\sigma_1^2 < \sigma_2^2$ does not give one any precise information about the nature of the corresponding

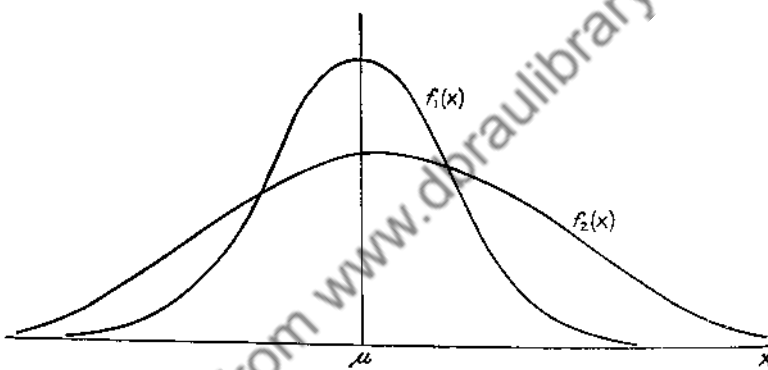


FIG. 26.

distributions, but it is evidence that $f_1(x)$ has more area near the mean than $f_2(x)$, at least for certain intervals about the mean. The two densities in Fig. 26, for example, might have about equal variances, and one could alter either one slightly so as to make it have a smaller or larger variance than the other.

The third moment μ_3 about the mean is sometimes called a measure of asymmetry, or *skewness*. Symmetric distributions like those in Figs. 26 and 30 can be shown to have $\mu_3 = 0$. A curve shaped like $f_1(x)$ in Fig. 27 is said to be skewed to the left and can be shown to have a negative third moment about the mean; one shaped like $f_2(x)$ is called skewed to the right and can be shown to have a positive third moment about the mean. Actually, however, knowledge of the third moment gives almost no clue as to the shape of the distribution, and we mention it at all mainly to point out the fact. Thus, for example, the density $f_3(x)$ in Fig. 27 has $\mu_3 = 0$, but it is far from symmetric.

By changing the curve slightly we could give it either a positive or negative third moment as we pleased.

While a particular moment or a few of the moments give little information about a distribution, the whole set of moments (μ_1', μ_2', μ_3', $\cdots$) will ordinarily determine the distribution exactly, and for this reason we shall have occasion to use the moments in theoretical work.

In applied statistics, the first two moments are of great importance, as we shall see, but the third and higher moments are rarely useful. Ordinarily one does not know what distribution function he is working with in a practical problem, and often it makes little difference what the actual shape of the distribution is. But it is usually necessary to know at least the location of the distribution and to have some idea of its dispersion. These characteristics can be estimated by examining

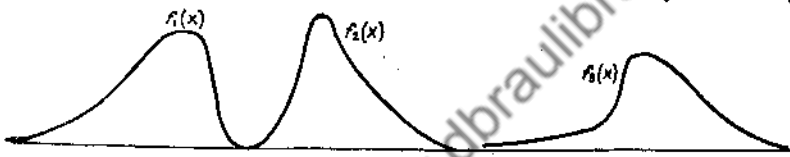


FIG. 27.

a sample drawn from a set of objects known to have the distribution in question. This estimation problem is probably the most important problem in applied statistics, and a large part of this course will be devoted to a study of it.

Illustrative example: Find the mean and variance of the hypergeometrical distribution

$$f(x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \qquad x = 0, 1, 2, \cdots, k \tag{8}$$

This problem will illustrate a technique that may be used to find the moments of a great many discrete distributions. The first step is to use the distribution to determine an identity in the parameters. Since $\Sigma f(x) = 1$, it follows that

$$\sum_{x=0}^{k} \binom{m}{x}\binom{n}{k-x} = \binom{m+n}{k} \tag{9}$$

for any positive integral values of m, n, and k. [Actually, as we have seen before, the range depends on the relative sizes of m, n, and k, but we can avoid dealing with these details by defining the binomial

coefficient $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ to be zero when either b or $a - b$ is negative.]

The mean of the distribution is

$$\mu_1' = E(x) = \sum_{x=0}^{k} xf(x)$$

$$= \frac{\sum_{x=0}^{k} x\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \tag{10}$$

In this expression x may be canceled with the x in the denominator of $\binom{m}{x}$ to get

$$x\binom{m}{x} = m\binom{m-1}{x-1}$$

and we have

$$\mu = \frac{\sum_{x=1}^{k} m\binom{m-1}{x-1}\binom{n}{k-x}}{\binom{m+n}{k}} \tag{11}$$

where we have written the sum to range from 1 to k because the first term in (10) vanishes and may be omitted. Actually, since we have defined a binomial coefficient to be zero when its lower index is negative, there would be no objection to leaving the limits 0 to k. Now in this last expression let us substitute y for $x - 1$ and factor out factors which do not involve the summation index. We get

$$\mu = \frac{m}{\binom{m+n}{k}} \sum_{y=0}^{k-1} \binom{m-1}{y}\binom{n}{k-1-y} \tag{12}$$

This sum may be evaluated by means of the identity (9); we simply replace m by $m - 1$ and k by $k - 1$ in the right-hand side of (9) to get

$$\mu = \frac{m}{\binom{m+n}{k}} \binom{m-1+n}{k-1}$$

$$= \frac{mk}{m+n} \tag{13}$$

To get the variance, we shall need the second moment

$$\mu_2' = \sum_{x=0}^{k} x^2 f(x)$$

If we substitute directly for $f(x)$, we shall be able to cancel only one of the x's, and the other x will remain to prevent our using the identity to evaluate the sum. The trick here is to write x^2 in the form

$$x(x-1) + x$$

to get

$$\mu_2' = \Sigma x(x-1)f(x) + \Sigma x f(x) \tag{14}$$

We have already evaluated the second sum in obtaining the mean, and the same procedure used on the first sum gives

$$\begin{aligned}
E[x(x-1)] &= \frac{\sum_{x=0}^{k} x(x-1)\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \\
&= \frac{\sum_{x=2}^{k} m(m-1)\binom{m-2}{x-2}\binom{n}{k-x}}{\binom{m+n}{k}} \\
&= \frac{m(m-1)}{\binom{m+n}{k}} \sum_{y=0}^{k-2} \binom{m-2}{y}\binom{n}{k-2-y} \\
&= \frac{m(m-1)}{\binom{m+n}{k}} \binom{m-2+n}{k-2} \\
&= \frac{m(m-1)k(k-1)}{(m+n)(m+n-1)}
\end{aligned} \tag{15}$$

Adding (13) to this, we get μ_2' in accordance with (14); the variance is then obtained by subtracting the square of (13) from μ_2' in accordance with (5). Thus the variance is

$$\begin{aligned}
\sigma^2 &= \frac{m(m-1)k(k-1)}{(m+n)(m+n-1)} + \frac{mk}{m+n} - \left(\frac{mk}{m+n}\right)^2 \\
&= \frac{mnk(m+n-k)}{(m+n)^2(m+n-1)}
\end{aligned} \tag{16}$$

The general method for higher moments is now evident. To get the third moment, we would find the expected value of

$$x(x-1)(x-2)$$

since this is equal to $x^3 - 3x^2 + 2x$, we have

$$\mu_3' - 3\mu_2' + 2\mu_1' = E[x(x-1)(x-2)]$$

and having evaluated the right-hand side of this expression, we could solve for μ_3', since μ_2' and μ_1' have already been determined. Having the third moment, we could obtain the fourth by finding the expected value of $x(x-1)(x-2)(x-3)$, then solving for μ_4' in

$$\mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1' = E[x(x-1)(x-2)(x-3)]$$

The right-hand side of this last expression is called the fourth *factorial moment* of the distribution. The rth factorial moment is

$$E[x(x-1)(x-2)\cdots(x-r+1)]$$

Illustrative example: Find the mean and standard deviation of the continuous distribution $f(x) = 2(1-x)$, $0 < x < 1$. The rth moment is

$$\begin{aligned}\mu_r' = E(x^r) &= \int_0^1 x^r 2(1-x)dx \\ &= 2\int_0^1 (x^r - x^{r+1})dx \\ &= \frac{2}{(r+1)(r+2)}\end{aligned}$$

The mean is

$$\mu = \mu_1' = \frac{2}{2\times 3} = \frac{1}{3}$$

and the variance is

$$\sigma^2 = \mu_2' - \mu^2 = \frac{2}{3\times 4} - \frac{1}{9} = \frac{1}{18}$$

hence

$$\sigma = \sqrt{\frac{1}{18}} = \frac{1}{3\sqrt{2}}$$

5.3. Moment Generating Functions. When all the moments of a distribution exist (i.e., when all moments are finite), it is possible to associate a moment generating function with the distribution. This is defined as $E(e^{xt})$, where x is the random variable and t is a continuous variable; the expected value of e^{xt} will be a function of t which we

shall denote by

$$m(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} f(x)dx \tag{1}$$

If we differentiate the members of this relation r times with respect to t, we have

$$\frac{d^r}{dt^r} m(t) = \int_{-\infty}^{\infty} x^r e^{xt} f(x)dx \tag{2}$$

and on putting $t = 0$, we find

$$\frac{d^r}{dt^r} m(0) = E(x^r) = \mu'_r \tag{3}$$

where the symbol on the left is to be interpreted to mean the rth derivative of $m(t)$ evaluated at $t = 0$. Thus the moments of a distribution may be obtained from the moment generating function by differentiation.

If in equation (1) we replace e^{xt} by its series expansion, we obtain the series expansion of $m(t)$ in terms of the moments of $f(x)$; thus

$$\begin{aligned} m(t) &= E\left(1 + xt + \frac{1}{2!}(xt)^2 + \frac{1}{3!}(xt)^3 + \cdots\right) \\ &= 1 + \mu'_1 t + \frac{1}{2!}\mu'_2 t^2 + \cdots \\ &= \sum_{i=0}^{\infty} \frac{1}{i!}\mu'_i t^i \end{aligned} \tag{4}$$

from which it is again evident that μ'_r may be obtained by differentiating $m(t)$ r times and then putting $t = 0$.

We may illustrate this technique for finding moments by obtaining the mean and variance of the Poisson density:

$$f(x) = \frac{e^{-a}a^x}{x!} \qquad x = 0, 1, 2, \cdots$$

We find

$$\begin{aligned} m(t) = E(e^{xt}) &= \sum_{x=0}^{\infty} \frac{e^{xt}e^{-a}a^x}{x!} \\ &= e^{-a} \sum_{x=0}^{\infty} \frac{(ae^t)^x}{x!} \\ &= e^{-a}e^{ae^t} \end{aligned}$$

The first two derivatives are

$$m'(t) = e^{-a}ae^te^{ae^t}$$
$$m''(t) = e^{-a}ae^te^{ae^t}(1 + ae^t)$$

whence

$$\mu = m'(0) = a$$
$$\mu_2' = m''(0) = a(1 + a)$$
$$\sigma^2 = a(1 + a) - a^2 = a$$

The *factorial moment generating function* is defined as $E(t^x)$, and the factorial moments are obtained from this function in the same way as the ordinary moments are obtained from $E(e^{xt})$ except that t is put equal to one instead of zero. This function sometimes simplifies the problem of finding moments of discrete distributions. It is, however, of no help in the example used in the preceding section, because the sum $\Sigma t^x f(x)$ has no simple expression. For the Poisson distribution:

$$E(t^x) = e^{a(t-1)}$$

whence

$$E(x) = ae^{a(t-1)}\Big]_{t=1} = a$$
$$E[x(x-1)] = a^2e^{a(t-1)}\Big]_{t=1} = a^2$$

giving the same moments as before.

Sometimes we shall have occasion to speak of the moments of a function of a random variable. Thus we may want the moments of $h(x)$, where x has the distribution $f(x)$. The rth moment of $h(x)$ is

$$E[h(x)]^r = \int_{-\infty}^{\infty} [h(x)]^r f(x)dx \tag{5}$$

and a function which will generate the moments is obviously

$$E(e^{th(x)}) = \int_{-\infty}^{\infty} e^{th(x)}f(x)dx \tag{6}$$

5.4. Moments for Multivariate Distributions. The preceding ideas are readily extended to distributions of several variates. Suppose, for example, that we have three variates (x, y, z) with density $f(x, y, z)$. The rth moment of y, for example, is

$$E(y^r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^r f(x, y, z)dz\,dy\,dx \tag{1}$$

Besides the moments of the individual variates, there are various *joint moments* defined in general by

$$E(x^q y^r z^s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^q y^r z^s f(x, y, z)dz\,dy\,dx \tag{2}$$

where q, r, and s are any positive integers including zero. The most important joint moment is the *covariance*, which is the joint moment about the means of the product of two variates. Thus the covariance between x and z is

$$\sigma_{xz} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [x - E(x)][z - E(z)]f(x, y, z)dz\,dy\,dx \tag{3}$$

and there are two other covariances σ_{xy} and σ_{yz} defined analogously. The *correlation* between two variates, say x and z, is denoted by ρ_{xz} and is defined by

$$\rho_{xz} = \frac{\sigma_{xz}}{\sigma_x\sigma_z} \tag{4}$$

where σ_x and σ_z are the standard deviations of x and z.

Also one can define a *joint moment generating function:*

$$m(t_1, t_2, t_3) = E(e^{t_1x+t_2y+t_3z}) \tag{5}$$

It is clear that the rth moment of y, for example, may be obtained by differentiating the moment generating function r times with respect to t_2 and then putting all the t's equal to zero. Similarly the joint moment (2) would be obtained by differentiating the function q times with respect to t_1, r times with respect to t_2, s times with respect to t_3, and then putting all the t's equal to zero.

5.5. The Moment Problem. We have seen that a distribution $f(x)$ determines a set of moments $(\mu_1', \mu_2', \mu_3', \cdots)$. One of the important problems of theoretical statistics is to find $f(x)$ when the moments are given. A study of this problem requires advanced mathematical techniques, and we shall have to omit it. However we shall prove the following theorem which will be required in our later work:

If two continuous densities have the same set of moments and if the difference of the densities has a series expansion about the origin, then the two densities are equivalent.

Suppose the two densities are represented by $f(x)$ and $g(x)$ and suppose the series expansion of their difference is

$$f(x) - g(x) = c_0 + c_1x + c_2x^2 + \cdots$$

Now let us consider the integral

$$\begin{aligned}\int_{-\infty}^{\infty} [f(x) - g(x)]^2\,dx &= \int_{-\infty}^{\infty} (c_0 + c_1x + c_2x^2 + \cdots)[f(x) - g(x)]dx \\ &= c_0(1 - 1) + c_1(\mu_1' - \mu_1') + \cdots \\ &= 0\end{aligned}$$

since the two densities are assumed to have the same moments. The function $[f(x) - g(x)]^2$ is necessarily positive or zero, and as we have found the area under the function to be zero, we must conclude that the function is zero and hence that

$$f(x) = g(x)$$

Under the conditions of this theorem it follows that

If two random variables have the same moment generating function, then they have the same density function.

For if the variables have the same moment generating function, they necessarily have the same moments.

5.6. Problems

1. If 5000 lottery tickets are sold at \$1 each on a \$2000 car, what is the expected gain of a person who buys three tickets?

2. A coin is tossed until a head appears; what is the expected number of tosses?

3. A bowl contains n chips numbered from 1 to n; m are drawn without replacement; what is the expected value of the sum of the numbers drawn?

4. An event occurs with probability p and fails to occur with probability $q = 1 - p$. In a single trial, what are the mean and variance of x, the number of successes?

5. If n trials are made of the event described in Prob. 4, and if x is the total number of successes, what are the mean and variance of x?

6. Find the mean of the continuous variate x distributed by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-a)^2/2}, \qquad -\infty < x < \infty$$

7. Find the mean and variance of x if $f(x) = 1$, $0 < x < 1$.

8. Find the mean and variance of $2x^2$ if $f(x) = 1$, $0 < x < 1$.

9. Find the mean and variance of x if

$$f(x) = 1/(x+1)^2 \qquad 0 < x < \infty$$

10. Show that $E(xy) = E(x)E(y)$ when x and y are independently distributed.

11. Show that

$$\mu_r = \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i} \mu_i' (\mu_1')^{r-i}$$

12. What is the median of x if $f(x) = 2(1 - x)$, $0 < x < 1$?

13. Find the moment generating function associated with the density $f(x) = ae^{-ax}$, $x > 0$, and use it to obtain the mean and variance of x.

14. Find the factorial moment generating function for the binomial distribution, and use it to obtain the third moment μ_3'.

15. If x has the density $f(x) = x/2$, $0 < x < 2$, find the rth moment of x^2. Then show that $y = x^2$ has the distribution

$$g(y) = \tfrac{1}{4} \qquad 0 < y < 4$$

by showing that y has the same moments as x^2.

16. If $f(x, y) = a^2e^{-a(x+y)}$, $x > 0$, $y > 0$, find the generating function for the moments of $u = x + y$. Deduce the distribution of u from the form of this generating function.

17. Show that if a density function $f(x)$ is symmetric about a point, say b, [i.e., $f(b + c) = f(b - c)$ for every value of c], then that point must be the mean of x. Show also in this case that all odd moments about the mean must be zero.

18. Given the moment generating function $m(t)$ for the moments μ_r' about the origin, how would one obtain the moment generating function for the moments μ_r about the mean?

19. In place of the moments μ_r', another infinite set of constants γ_r called the *cumulants* of a distribution is often useful for characterizing the distribution function. The cumulants are defined by the generating function $c(t) = \log m(t)$, where $m(t)$ is the generating function for the μ_r', i.e., $\gamma_r = \dfrac{d^r c(t)}{dt^r}$ evaluated at $t = 0$. Show that $\gamma_1 = \mu_1'$ and $\gamma_2 = \sigma^2$.

20. Find the rth cumulant γ_r for the density $f(x) = ae^{-ax}$, $x > 0$.

21. Show that if $M(t)$ generates the moments about an arbitrary point b, i.e.,

$$M(t) = \int_{-\infty}^{\infty} e^{t(x-b)} f(x)dx$$

then $C(t) = \log M(t)$ will correctly generate all the cumulants except the first. The cumulants of a distribution beyond γ_1 are thus said to be invariant under translations of the variate.

22. If x has cumulants γ_r, show that $y = kx$ has cumulants $k^r\gamma_r$.

23. Show that the correlation between two variates is zero if they are independently distributed. (The converse of this statement is not true, as the following problem shows.)

24. Let x have the marginal density $f_1(x) = 1$, $-\frac{1}{2} < x < \frac{1}{2}$, and let the conditional density of y be

$$\begin{aligned} f(y|x) &= 1 \qquad x < y < x + 1,\ -\tfrac{1}{2} < x < 0 \\ &= 1 \qquad -x < y < 1 - x,\ 0 < x < \tfrac{1}{2} \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Find the correlation between x and y.

25. Could the function $E[1/(1 + tx)]$ be used to generate the moments of a variate x?

CHAPTER 6

SPECIAL CONTINUOUS DISTRIBUTIONS

6.1. Uniform Distribution. The simplest distribution for a continuous variate is the uniform density:

$$f(x) = \frac{1}{\beta - \alpha} \qquad \alpha < x < \beta \tag{1}$$
$$= 0 \qquad \text{otherwise}$$

which is plotted in Fig. 28. The probability that an observation will fall in any interval within $\alpha < x < \beta$ is equal to $1/(\beta - \alpha)$ times the

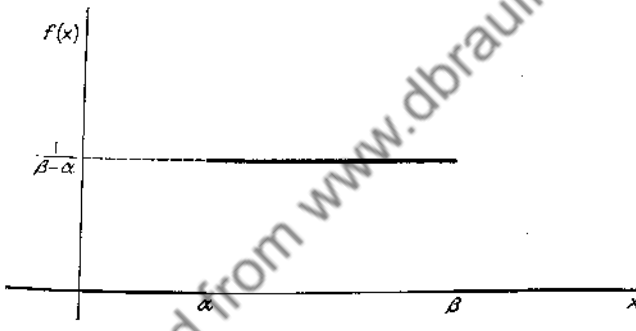


FIG. 28.

length of the interval. The distribution is particularly useful in theoretical statistics because it is convenient to deal with mathematically.

We are enabled to deal only with this simple distribution when discussing certain properties of distributions in general by the following theorem:

Any density for a continuous variate x may be transformed to the uniform density

$$f(y) = 1 \qquad 0 < y < 1 \tag{2}$$

by letting $y = G(x)$, where $G(x)$ is the cumulative distribution of x.

It is clear that y must have range zero to one since a cumulative distribution must vary between zero and one. We need only show that the density of y is $f(y) = 1$ over that range. Now a value of y is determined by drawing a value of x, say x_0, and substituting in $G(x)$

to get a corresponding $y_0 = G(x_0)$. The transformation $y = G(x)$ sets up a correspondence between points of the x axis and points on the interval (0, 1) on the y axis. To find the probability that y lies in an interval, say $a < y < b$, we find the values, say a' and b', on the x axis which correspond to a and b, as in Fig. 29, and compute the probability for that interval (a', b') in terms of x. Thus,

$$P(a < y < b) = G(b') - G(a')$$

but by definition $G(b') = b$ and $G(a') = a$; hence

$$P(a < y < b) = b - a \qquad 0 < a < b < 1$$

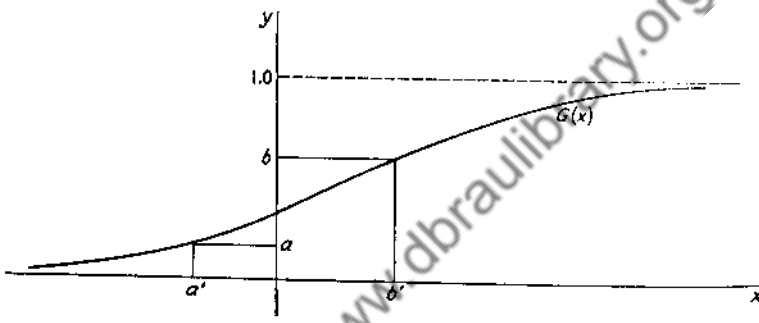


FIG. 29.

Suppose we denote the cumulative distribution of y by $F(y)$; then

$$F(b) - F(a) = b - a$$

and replacing b by $y + \Delta y$ and a by y, we get

$$\frac{F(y + \Delta y) - F(y)}{\Delta y} = 1$$

The limit of the expression on the left as Δy approaches zero gives the derivative of the cumulative distribution, which is the density we seek:

$$f(y) = \frac{d}{dy} F(y) = \lim_{\Delta y \to 0} \frac{F(y + \Delta y) - F(y)}{\Delta y} = 1 \qquad 0 < y < 1$$

which proves that y has the density (2). The transformation $y = G(x)$ is called the *probability transformation.*

By means of this theorem it is possible to demonstrate many properties of continuous distributions in general by proving them merely for the uniform distribution over the unit interval.

6.2. The Normal Distribution. A great many of the techniques used in applied statistics are based upon the normal distribution, and

much of the remainder of this course will be devoted to a study of this distribution. The density is

$$n(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty \tag{1}$$

and the function is plotted in Fig. 30 for several values of σ. Changing μ merely shifts the curves to the right or left without changing their shape. The function given actually represents a two-parameter family of distributions, the parameters being μ and σ^2. We have used the symbols μ and σ^2 to represent the parameters because the parameters turn out, as we shall see, to be the mean and variance, respectively, of the distribution.

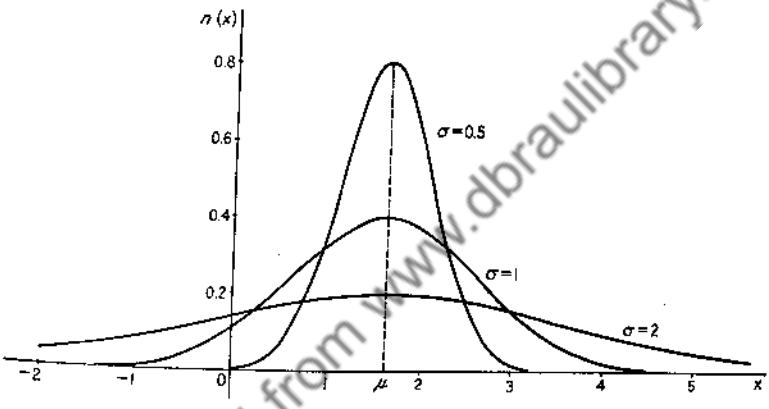


FIG. 30.

Since $n(x)$ is given to be a density function, it is implied that

$$\int_{-\infty}^{\infty} n(x)dx = 1$$

but we should satisfy ourselves that this is true. The verification is somewhat troublesome because this particular function does not integrate into a simple closed expression. Suppose we represent the area under the curve by A; then

$$A = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}\, dx$$

and on making the substitution

$$y = \frac{x-\mu}{\sigma}$$

we find

$$A = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\, dy$$

We wish to show that $A = 1$, and this is most easily done by showing A^2 is one and then reasoning that $A = 1$, since $f(x)$ is positive. We may put

$$\begin{aligned} A^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\, dy \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^2+z^2)}\, dy\, dz \end{aligned}$$

writing the product of two integrals as a double integral. In this integral we change the variables to polar coordinates by the substitution

$$\begin{aligned} y &= r \sin \theta \\ z &= r \cos \theta \end{aligned}$$

and the integral becomes

$$\begin{aligned} A^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} r e^{-\frac{1}{2}r^2}\, dr\, d\theta \\ &= \int_0^{\infty} r e^{-\frac{1}{2}r^2}\, dr \\ &= 1 \end{aligned}$$

Since the integral of $n(x)$ does not have a simple functional form, we can only exhibit the cumulative distribution formally as

$$N(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{-[(t-\mu)^2/2\sigma^2]}\, dt \tag{2}$$

and if we let

$$y = \frac{t - \mu}{\sigma}$$

we find

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-\frac{1}{2}y^2}\, dy \tag{3}$$

and given a specific value for $(x - \mu)/\sigma$, the integral can be computed by numerical methods. A tabulation of this function may be found in Table II. Since the density is symmetric about μ, i.e., since

$$n(\mu - a) = n(\mu + a)$$

it follows that $N(x)$ for $(x - \mu)/\sigma$ negative is equal to $1 - N(x')$, where $(x' - \mu)/\sigma = -(x - \mu)/\sigma$. The graph of $N(x)$ is given in Fig. 31.

To illustrate the use of the table, we shall find $P(-1 < x < 4)$ when x has the density:

$$n(x) = \frac{1}{4\sqrt{2\pi}} e^{-[(x-2)^2/32]} \tag{4}$$

We note that

$$\mu = 2 \qquad \sigma = 4$$

and thus that the values of $(x - \mu)/\sigma$ corresponding to -1 and 4 are

$$\frac{-1-2}{4} = -\frac{3}{4} \qquad \frac{4-2}{4} = \frac{1}{2}$$

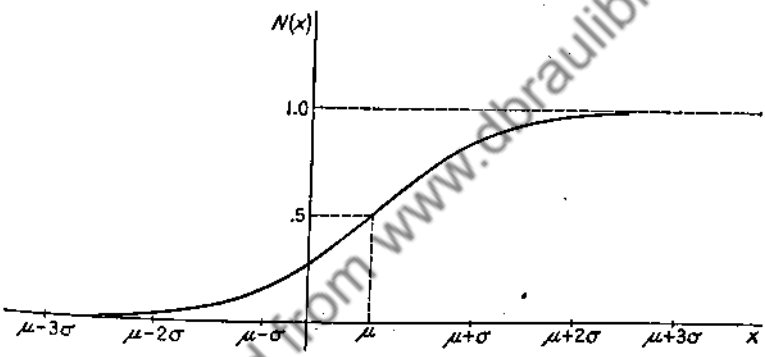


FIG. 31.

hence:

$$\begin{aligned} P(-1 < x < 4) &= N(4) - N(-1) \\ &= .6915 - (1 - .7734) \\ &= .4649 \end{aligned}$$

It is a great convenience that $N(x)$ is of such a form that it need not be tabulated for various combinations of values of μ and σ. The transformation $y = (x - \mu)/\sigma$ brings all normal distributions to the same form, called the *standard* or *normalized* form. We shall reserve the letters n and N henceforth to indicate the normal density and its cumulative form. Often we shall wish to indicate the parameters, and this will be done by writing the functions as $n(x; \mu, \sigma^2)$ and $N(x; \mu, \sigma^2)$, separating the parameters from the variate by a semicolon. In this notation the distribution (4) would be symbolized by $n(x; 2, 16)$. The

standard normal distribution is then

$$n(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{5}$$

and its cumulative form is

$$N(x; 0, 1) = \int_{-\infty}^{x} n(t; 0, 1)dt \tag{6}$$

We shall now find the moments of $n(x; \mu, \sigma^2)$ by finding first the moment generating function. The computation is as follows:

$$\begin{aligned} m(t) &= E(e^{tx}) = e^{t\mu}E(e^{t(x-\mu)}) \\ &= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{t(x-\mu)} e^{-(1/2\sigma^2)(x-\mu)^2} dx \\ &= e^{t\mu} \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} e^{-(1/2\sigma^2)[(x-\mu)^2-2\sigma^2t(x-\mu)]} dx \end{aligned}$$

On completing the square inside the bracket, it becomes

$$\begin{aligned} (x-\mu)^2 - 2\sigma^2t(x-\mu) &= (x-\mu)^2 - 2\sigma^2t(x-\mu) + \sigma^4t^2 - \sigma^4t^2 \\ &= (x-\mu-\sigma^2t)^2 - \sigma^4t^2 \end{aligned}$$

and we have

$$m(t) = e^{t\mu}e^{\sigma^2t^2/2} \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu-\sigma^2t)^2/2\sigma^2} dx$$

The integral together with the factor $1/\sqrt{2\pi}\,\sigma$ is necessarily one, since it is the area under a normal distribution with mean $\mu + \sigma^2t$ and variance σ^2. Hence,

$$m(t) = e^{t\mu+(\sigma^2t^2/2)} \tag{7}$$

On differentiating this function twice and substituting $t = 0$ in the results, we find

$$\begin{aligned} \mu_1' &= \mu \\ \mu_2' &= \sigma^2 + \mu^2 \\ \text{Variance} &= \mu_2' - (\mu_1')^2 = \sigma^2 \end{aligned}$$

thus justifying our use of the moment symbols for the parameters.

6.3. The Gamma Distribution. The function

$$\begin{aligned} f(x) &= \frac{1}{\alpha!\beta^{\alpha+1}} x^\alpha e^{-x/\beta} \qquad x > 0 \\ &= 0 \qquad x < 0 \end{aligned} \tag{1}$$

is called the gamma distribution. This is a two-parameter family of distributions, the parameters being α and β. β must be positive, and

α must be greater than minus one. The function is plotted in Fig. 32 for $\beta = 1$ and several values of α. Changing β merely changes the scale on the two axes, as is evident on examining the form of the function.

To show that the function represents a density (has unit area), we shall evaluate the integral

$$A = \int_0^\infty \frac{1}{\beta^{\alpha+1}} x^\alpha e^{-x/\beta}\, dx$$
$$= \int_0^\infty y^\alpha e^{-y}\, dy$$

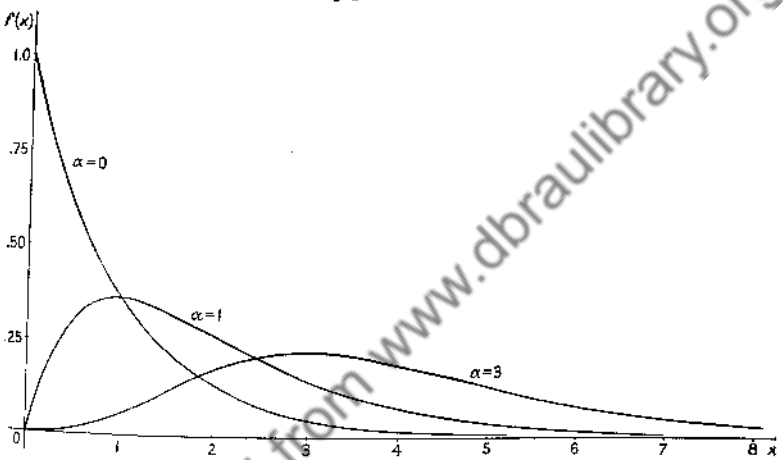


FIG. 32.

on substituting y for x/β; hence A is necessarily a function of α only. If $\alpha > 0$, we may integrate at once by parts to obtain

$$A(\alpha) = -y^\alpha e^{-y}\Big]_0^\infty + \int_0^\infty \alpha y^{\alpha-1} e^{-y}\, dy$$
$$= \alpha \int_0^\infty y^{\alpha-1} e^{-y}\, dy$$

Whence it follows that

$$A(\alpha) = \alpha A(\alpha - 1) \tag{2}$$

If α is a positive integer, we may apply this recurrence formula (2) successively to obtain

$$A(\alpha) = \alpha(\alpha - 1)(\alpha - 2) \cdots (2)(1)A(0)$$

and since

$$A(0) = \int_0^\infty e^{-y}\, dy = 1$$

we have

$$A(\alpha) = \alpha!$$

when α is an integer. The function $A(\alpha)$ is often denoted by $\Gamma(\alpha + 1)$ in mathematical literature, but we shall use the symbol $\alpha!$ whether or not α is an integer.

In practically all applications of the distribution, α is either an integer or a multiple of one-half. Hence for our purposes we need only to evaluate $(\frac{1}{2})!$ in order to be able to compute $\alpha!$ for any value of α we may encounter.

$$\begin{aligned}(\tfrac{1}{2})! &= \tfrac{1}{2}(-\tfrac{1}{2})! \\ &= \tfrac{1}{2}\int_0^\infty y^{-\frac{1}{2}}e^{-y}\,dy\end{aligned}$$

and if we let $y = z^2/2$, we have

$$\begin{aligned}(\tfrac{1}{2})! &= \tfrac{1}{2}\int_0^\infty \sqrt{2}\,e^{-(z^2/2)}\,dz \\ &= \sqrt{\pi}\int_0^\infty \frac{1}{\sqrt{2\pi}}e^{-(z^2/2)}\,dz \\ &= \frac{\sqrt{\pi}}{2}\end{aligned}$$

since the integral is half the area under a normal density function and is therefore one-half. Knowing this number, we can evaluate $\alpha!$ for any multiple of one-half by using the relation (2); thus

$$\begin{aligned}(\tfrac{5}{2})! &= \tfrac{5}{2}(\tfrac{3}{2})! = \tfrac{5}{2} \times \tfrac{3}{2}(\tfrac{1}{2})! \\ &= \frac{15\sqrt{\pi}}{8}\end{aligned}$$

The cumulative distribution is

$$F(x) = \int_0^x \frac{1}{\alpha!\beta^{\alpha+1}}t^\alpha e^{-t/\beta}\,dt \qquad x > 0 \tag{3}$$

and is, of course, zero when $x < 0$. It must be evaluated by numerical methods unless α is a positive integer, in which case the function can be found by successive integrations by parts to be

$$F(x) = 1 - \left[1 + \frac{x}{\beta} + \frac{1}{2!}\left(\frac{x}{\beta}\right)^2 + \frac{1}{3!}\left(\frac{x}{\beta}\right)^3 + \cdots + \frac{1}{\alpha!}\left(\frac{x}{\beta}\right)^\alpha\right]e^{-x/\beta} \qquad x > 0 \tag{4}$$

But in any case it is usually simpler to refer to tables of the function in dealing with specific problems. The function $F(x)$ is called the *incomplete gamma function* and has been extensively tabulated by Karl Pearson ("Tables of the Incomplete Gamma Function," Cambridge University Press, London, 1922).

The moment generating function for this distribution is

$$m(t) = \int_0^\infty e^{tx} \frac{1}{\alpha!\beta^{\alpha+1}} x^\alpha e^{-x/\beta}\, dx$$
$$= \int_0^\infty e^{\beta t y} \frac{1}{\alpha!} y^\alpha e^{-y}\, dy$$

on substituting y for x/β. This may then be put in the form:

$$m(t) = \frac{1}{\alpha!} \int_0^\infty y^\alpha e^{-y(1-\beta t)}\, dy$$
$$= \frac{1}{(1-\beta t)^{\alpha+1}} \int_0^\infty \frac{(1-\beta t)^{\alpha+1}}{\alpha!} y^\alpha e^{-y(1-\beta t)}\, dy$$
$$= \frac{1}{(1-\beta t)^{\alpha+1}} \tag{5}$$

provided $t < 1/\beta$, since the last integral represents the area under a gamma distribution with parameters α and $\beta' = 1/(1-\beta t)$, and is therefore one. On differentiating $m(t)$ twice and putting $t = 0$ in the results, we find

$$\mu = \beta(\alpha + 1) \tag{6}$$
$$\mu_2' = \beta^2(\alpha + 1)(\alpha + 2) \tag{7}$$
$$\sigma^2 = \beta^2(\alpha + 1) \tag{8}$$

6.4. The Beta Distribution. The density

$$f(x) = \frac{(\alpha+\beta+1)!}{\alpha!\beta!} x^\alpha (1-x)^\beta \qquad 0 < x < 1 \tag{1}$$
$$= 0 \qquad \text{elsewhere}$$

is called the beta density. The function represents a two-parameter family of distributions, and a few examples are plotted in Fig. 33. The parameters α and β must both be greater than minus one. The distribution becomes the uniform distribution over the unit interval when $\alpha = \beta = 0$.

To show that the area under $f(x)$ is one, we shall compute the integral

$$A(\alpha, \beta) = \int_0^1 x^\alpha (1-x)^\beta\, dx \tag{2}$$

Clearly A will be a function of α and β; we wish to show that it is the reciprocal of the constant multiplier in (1). Referring back to the gamma distribution, we may write

$$\alpha!\beta! = \left(\int_0^\infty x^\alpha e^{-x}\,dx\right)\left(\int_0^\infty y^\beta e^{-y}\,dy\right)$$
$$= \int_0^\infty \int_0^\infty x^\alpha y^\beta e^{-(x+y)}\,dx\,dy$$

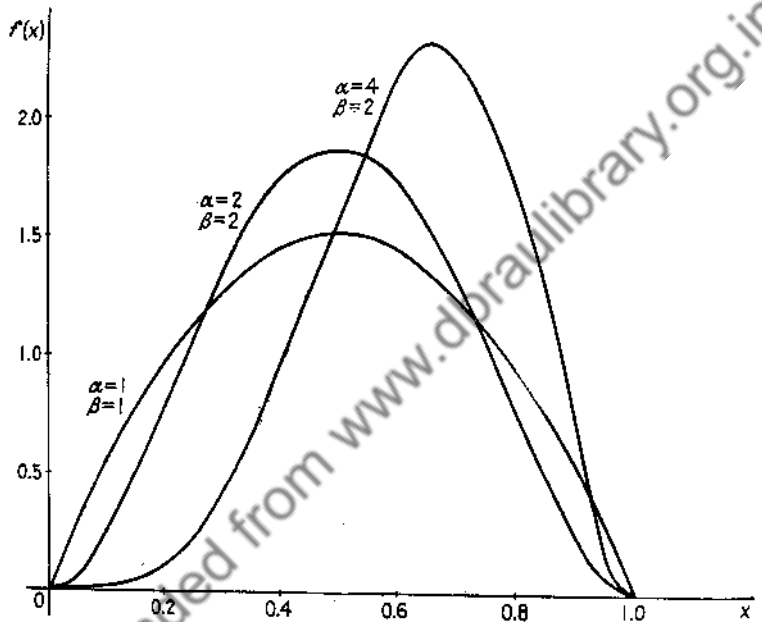


FIG. 33.

and in this last integral we shall change the variable x to u by the substitution

$$u = \frac{x}{x+y}$$

or

$$x = \frac{uy}{1-u} \qquad dx = \frac{y\,du}{(1-u)^2}$$

Since u obviously has the range zero to one, the integral becomes

$$\alpha!\beta! = \int_0^\infty \int_0^1 \left(\frac{uy}{1-u}\right)^\alpha y^\beta e^{-y/(1-u)} \frac{y}{(1-u)^2}\,du\,dy$$

In this integral we change y to v by the substitution

$$y = (1 - u)v \qquad dy = (1 - u)dv$$

to get

$$\begin{aligned}\alpha!\beta! &= \int_0^\infty \int_0^1 u^\alpha(1 - u)^\beta v^{\alpha+\beta+1}e^{-v}\,du\,dv \\ &= \left(\int_0^\infty v^{\alpha+\beta+1}e^{-v}\,dv\right)\left(\int_0^1 u^\alpha(1 - u)^\beta\,du\right) \\ &= (\alpha + \beta + 1)!\int_0^1 u^\alpha(1 - u)^\beta\,du\end{aligned}$$

which shows that $A(\alpha, \beta)$ has the stated value. $A(\alpha - 1, \beta - 1)$ is called the *beta function* of α and β in the literature and is usually denoted by $B(\alpha, \beta)$.

The cumulative distribution, often called the *incomplete beta function*, is

$$\begin{aligned}F(x) &= 0 & x < 0 \\ &= \int_0^x \frac{(\alpha + \beta + 1)!}{\alpha!\beta!} t^\alpha(1 - t)^\beta\,dt & 0 < x < 1 \qquad (3) \\ &= 1 & x > 1\end{aligned}$$

and has also been extensively tabulated by Karl Pearson ("Tables of the Incomplete Beta Function," Cambridge University Press, London, 1932).

The moment generating function for this distribution does not have a simple form, but the moments are readily found directly:

$$\begin{aligned}\mu_r' = E(x^r) &= \frac{(\alpha + \beta + 1)!}{\alpha!\beta!}\int_0^1 x^{r+\alpha}(1 - x)^\beta\,dx \\ &= \frac{(\alpha + \beta + 1)!(\alpha + r)!}{(\alpha + \beta + r + 1)!\alpha!}\int_0^1 \frac{(\alpha + \beta + r + 1)!}{(\alpha + r)!\beta!} x^{r+\alpha}(1 - x)^\beta\,dx \\ &= \frac{(\alpha + \beta + 1)!(\alpha + r)!}{(\alpha + \beta + r + 1)!\alpha!} \qquad (4)\end{aligned}$$

since the integral must be one.

6.5. Other Distribution Functions. A distribution which we shall find useful for illustrative purposes is the Cauchy density

$$f(x) = \frac{1}{\pi}\frac{1}{1 + (x - \mu)^2} \qquad -\infty < x < \infty \qquad (1)$$

which has a mean only in a restricted sense and no higher moments. The cumulative distribution is

$$F(x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{dt}{1 + (t - \mu)^2}$$
$$= \frac{1}{\pi} \text{arc tan } (t - \mu)\Big]_{-\infty}^{x}$$
$$= \frac{1}{2} + \frac{1}{\pi} \text{arc tan } (x - \mu) \tag{2}$$

Pearson's Distributions. A general class of distribution functions is given by the families of solutions of the differential equation

$$\frac{dy}{dx} = \frac{(x + a)y}{bx^2 + cx + d} \tag{3}$$

The equation was obtained by Karl Pearson by putting dy/dx equal to the slope of a straight line joining two successive points of the discrete hypergeometric distribution. The solutions of this equation were classified by Pearson into twelve families of curves, those of one family being called Type I curves, those of a second Type II, and so on. The gamma distributions are essentially the Type III curves of Pearson; the normal distributions are his Type VII curves; the beta distributions represent his Type I curves, while with $\alpha = \beta$ they represent his Type II curves.

The different families of curves arise when different relations are assumed between the constants a, b, c, d in the differential equation. Thus, for example, when b and c are zero, the equation becomes

$$\frac{dy}{y} = \frac{1}{d}(x + a)\,dx$$

and its solution is

$$\log y = \frac{1}{2d}(x + a)^2 + K$$

or

$$y = ke^{(x+a)^2/2d}$$

which becomes the normal density when d is taken to be negative and k is determined so as to make the area under the curve equal to one. By considering various other conditions on the constants in (3), we could derive all twelve of Pearson's types of curves, but we shall not develop these because most of them have not proved to be of great importance in statistics.

The Gram-Charlier Series. A wide class of density functions may be represented by an infinite series called the Gram-Charlier series. Suppose $f(x)$ is a density function and suppose its mean and variance

are μ and σ^2. Let

$$y = \frac{x - \mu}{\sigma}$$

then y has zero mean and unit variance. The Gram-Charlier series is a series in the derivatives of the normal distribution of y. Let $n_i(y)$ represent the ith derivative of the standard normal density $n(y; 0, 1)$ Thus

$$\begin{aligned} n_0(y) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \\ n_1(y) &= -(y) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} = -yn_0(y) \\ n_2(y) &= (y^2 - 1)n_0(y) \\ n_3(y) &= -(y^3 - 3y)n_0(y) \end{aligned}$$

and in general

$$n_i(y) = H_i(y)n_0(y)$$

where $H_i(y)$ is a polynomial of degree i in y called the ith Hermite polynomial. The Gram-Charlier theorem states that under rather general conditions $f(x)$ may be put in the form

$$\begin{aligned} f(x) &= \alpha_0 n_0(y) + \alpha_1 n_1(y) + \alpha_2 n_2(y) + \cdots \\ &= \sum_{i=0}^{\infty} \alpha_i n_i(y) \\ &= n_0(y) \sum_{i=0}^{\infty} \alpha_i H_i(y) \end{aligned} \tag{4}$$

where the α_i are constants and $y = (x - \mu)/\sigma$. It can be shown that

$$H_i(y) = (-1)^i \left[y^i - \frac{i(i-1)}{2} y^{i-2} + \frac{i(i-1)(i-2)(i-3)}{2 \times 4} y^{i-4} - \cdots \right] \tag{5}$$

$$\begin{aligned} \int_{-\infty}^{\infty} H_i(y)H_j(y)n_0(y)dy &= 0 \qquad \text{if } i \neq j \\ &= i! \qquad \text{if } i = j \end{aligned} \tag{6}$$

We shall not prove these relations. By means of the second one we may determine the coefficients α_i when $f(x)$ is known and can be expressed by (4). Let equation (4) be multiplied by $H_j(y)$ and then integrated on both sides with respect to x after putting $y = (x - \mu)/\sigma$. We find

$$\int_{-\infty}^{\infty} H_i\left(\frac{x - \mu}{\sigma}\right) f(x)dx = \alpha_i i!$$

on applying (6), and hence that

$$\alpha_i = \frac{1}{i!}\int_{-\infty}^{\infty} H_i\left(\frac{x-\mu}{\sigma}\right) f(x)dx \tag{7}$$

Since the $H_i[(x - \mu)/\sigma]$ are polynomials in $(x - \mu)$, the α_i will be linear functions of the moments of x about the mean.

The Pearson curves and the Gram-Charlier series were devised to meet the following practical problem: In general $f(x)$ is unknown, and all that is available is a sample of values of x. By means of the sample, the moments of $f(x)$ can be estimated. A Pearson curve which is intended to approximate $f(x)$ may be fitted to the sample by equating the sample moments to the theoretical moments and solving for the parameters which appear in the theoretical moments. These values of the parameters are then substituted in the function to obtain a specific function which is meant to approximate $f(x)$. Similarly, having estimated the moments, they may be used to determine a set of values of α_i which, when substituted in (4), gives an approximation to $f(x)$; in this method only the first few terms of the infinite series are used.

Actually the process of fitting a smooth curve to a sample does not add anything to our information about $f(x)$ that is not contained in the sample. The fitted curve may, in fact, give one an entirely misleading impression of the real density function. However, when the sample is quite large, it is sometimes convenient to replace the data by some sort of fitted curve in order to simplify further computations. Insurance companies and certain government agencies which deal with large masses of data find the technique convenient.

6.6. Problems

1. Find and plot the cumulative form for the uniform distribution.

2. What transformation will change the variate x to one which will have the uniform distribution over the unit interval if

$$f(x) = \frac{(x-1)}{2}$$

$1 < x < 3$? What interval for the new variate corresponds to $1.1 < x < 2.9$?

3. Plot $n(x; 0, .25)$, $n(x; 1, .25)$, and $n(x; 1, 9)$ on the same graph. What would be the appearance of the distribution if σ were very small? (Use Table I.)

4. If x is normally distributed with unit mean and $\sigma = .4$, find $P(x > 0)$ and $P(.2 < x < 1.8)$.

5. Find the number k such that for a normally distributed variate, $P(\mu - k\sigma < x < \mu + k\sigma) = .95$. What would k be if $P = .90$? .99? For what value of k is $P(x > \mu - k\sigma) = .95$?

6. Find the generating function $E(e^{t(x-\mu)})$ for the moments about the mean for a normal distribution.

7. Find μ_r in terms of σ for a normal distribution for r even and r odd. (Expand the above generating function in an infinite series.)

8. What constant multiplier will change the function e^{-x^2+x} into a density function? What are the mean and the variance of the resulting distribution?

9. Evaluate $\int_1^2 e^{-x^2}\,dx$.

10. Evaluate $\int_0^\infty x^{3/2}e^{-x/2}\,dx$.

11. Plot the gamma density for $\alpha = 1, \beta = 1$; $\alpha = 1, \beta = 2$; $\alpha = 2, \beta = 1$; $\alpha = 4, \beta = 1$.

12. Find the third moment, μ_3', of the gamma distribution.

13. If in the gamma distribution β is put equal to 2 and α is put equal to $(n - 2)/2$, the resulting distribution is called the chi-square distribution with n degrees of freedom. Find its moment generating function and its mean and variance.

14. Find k such that $P(x > k) = .05$ for the chi-square distribution with two degrees of freedom.

15. Find the rth moment of the gamma distribution without using the moment generating function.

16. Find the rth moment of the gamma distribution using the generating function.

17. Plot the beta density for $\alpha = 0, \beta = 0$; $\alpha = 1, \beta = 1$; $\alpha = 3, \beta = 3$; $\alpha = 2, \beta = 3$; $\alpha = 3, \beta = 2$. What would be the appearance of the function if both α and β were large?

18. Find the mean and variance of the beta distribution.

19. Show that the beta density is symmetric about the point $x = \frac{1}{2}$ when $\alpha = \beta$.

20. Find the mean of the Cauchy distribution if $\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x\,dx}{1 + (x - \mu)^2}$ is defined to be

$$\lim_{A \to \infty} \int_{\mu - A}^{\mu + A} \frac{1}{\pi} \frac{x\,dx}{1 + (x - \mu)^2}$$

Show that the distribution does not have any higher moments.

21. Integrate Pearson's differential equation when c and d equal zero. What family of distributions does the result represent?

22. Show that any Gram-Charlier expansion must have $\alpha_0 = 1$, $\alpha_1 = 0$, $\alpha_2 = 0$.

23. Evaluate α_4 for the Gram-Charlier expansion of $f(x) = 1$, $0 < x < 1$. Plot $f(x)$ and plot

$$f_1(x) = n_0(y) \sum_{i=0}^{4} \alpha_i H_i(y) \qquad y = \frac{x - \mu}{\sigma}$$

in order to see how the sum of first few terms of the Gram-Charlier series begins to approximate $f(x)$.

24. Compare the Cauchy density and the normal density with $\sigma = 2$ by plotting them on the same graph both with mean zero. Notice that the variance is a poor criterion for comparing two distributions unless it is known that they have the same functional form.

25. What are the cumulants of the normal distribution?

26. Let x have the gamma distribution with parameters $\alpha = 10$, $\beta = 1$. How many moments does $y = 1/x$ have?

27. If x has the gamma distribution, find the moment generating function of $y = \log x$.

28. A variate x has the density

$$f(x) = 2\sqrt{\frac{2}{\pi}}\, x e^{-x^2/2} \qquad x > 0$$

Find its mean and variance.

29. A variate has moments $\mu'_r = r!$. Find its moment generating function and then deduce its distribution.

30. A variate x has the uniform distribution over the unit interval; what function of x has the gamma distribution with $\alpha = 0$, $\beta = 1$?

31. A variate x has the beta distribution with $\alpha = 0$, $\beta = 1$. What function of x has the gamma distribution with $\alpha = 0$, $\beta = 1$?

32. A variate has moments $\mu'_r = \dfrac{r!}{(r/2)!}$ when r is even and $\mu'_r = 0$ when r is odd. Deduce the distribution of the variate from its moment generating function.

33. Show how tables of the incomplete gamma function $F(x; \alpha, \beta)$ may be used to evaluate the cumulative Poisson distribution, say,

$$\sum_{y=0}^{n} \frac{e^{-m} m^y}{y!}$$

34. If $\log x$ is normally distributed with $\mu = 1$, $\sigma^2 = 4$, find

$$P(1\tfrac{1}{2} < x < 2)$$

($\log 2 = .693$)

35. A variate x has the density

$$f(x) = 2\sqrt{\frac{2}{\pi}}\, x e^{-\frac{1}{2}x^4} \qquad x > 0$$

Find $P(x < 4)$.

36. Determine the mean and variance of the normal distribution by differentiating the identity

$$\int_{-\infty}^{\infty} n(x;\, \mu,\, \sigma^2)dx = 1$$

with respect to μ and with respect to σ^2.

37. A variate x is said to be transformed to standard scale if it is divided by its standard deviation. Show that the cumulants of x/σ are equal to $\gamma_r/\gamma_2^{r/2}$, where γ_r is the rth cumulant of x.

38. Show that the gamma distribution is nearly normal when α is large, by comparing the cumulants of the two distributions on standard scale.

39. A variate x is normally distributed with mean μ and variance σ^2. Show that the mean of the conditional distribution of x, given

$$a < x < b$$

is

$$\mu + \frac{n(a) - n(b)}{N(b) - N(a)}\,\sigma^2$$

40. A variate x has density $f(x)$. How might one determine a function $u(x)$ such that u is distributed by $g(u)$?

CHAPTER 7

SAMPLING

7.1. Inductive Inference. Up to now we have been concerned with certain aspects of the theory of probability. The subject of sampling brings us to the theory of statistics proper, and we shall consider briefly here one important area of the theory of statistics and its relation to sampling.

Progress in science is ascribed to experimentation. The research worker performs an experiment and obtains some data. On the basis of the data certain conclusions are drawn. The conclusions usually go beyond the materials and operations of the particular experiment. In other words, the scientist may generalize from a particular experiment to the class of all similar experiments. This sort of extension from the particular to the general is called *inductive inference*. It is the way in which new knowledge is found.

Inductive inference is well known to be a hazardous process. In fact, it is a theorem of logic that exact inductive inference is impossible. One simply cannot make a perfectly valid generalization. However, uncertain inferences can be made, and the degree of uncertainty can be measured if the experiment has been performed in accordance with certain principles. One function of statistics is the provision of techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences. Uncertainty is measured in terms of probability, and that is the reason we have devoted so much time to the theory of probability.

Let us consider a particular experiment to make the above ideas somewhat more concrete. Suppose a nutritionist studying a vitamin deficiency wishes to discover the effect of a certain diet. He selects, say, ten individuals and gives them the diet for a number of days or weeks. And let us suppose that the diet plainly affected all the individuals as reflected by some measurable criterion, say loss of weight or decreased metabolism. The nutritionist is not interested in confining his conclusions to this particular group of individuals. He would like to conclude that all or at least a large proportion of all individuals would react similarly to the diet.

It is clear that no certain generalization is possible. It is conceivable, for example, that the nutritionist was unfortunate enough to have selected individuals who happened to be physically on the downgrade at the time, so that the apparent results of the experiment were not in fact a consequence of the diet. Or the individuals may have been exposed to some minor malady which was not recognized. Some item of food in the diet may have been tainted. In fact, one could list a great many accidental circumstances which could have produced the observed results quite independently of the diet. Whatever generalization is made must be an uncertain one.

To complete the discussion, we shall consider one very simple kind of inference that may be made. Let us suppose that the individuals were selected from some large group of individuals, say the inhabitants of a county or state. We may envisage the possibility that there is some proportion p of the individuals in the large group which will be adversely affected by the diet and that the remaining proportion $q = 1 - p$ will be favorably affected or unaffected by the diet. Of course it is possible that q may be zero. If the ten individuals were drawn at random (with replacement) from the large group, then the probability that all ten would be adversely affected is p^{10}. Suppose we consider a few specific values for p. If $p = \frac{1}{2}$, then $p^{10} = \frac{1}{1024}$. If in fact p is one-half for the large group, then the experimenter has been most unlucky in his selection, for then a 1 in 1024 chance has occurred. If we try $p = .7$, we find $p^{10} = .03$, which would still make the sample rather improbable. We may reasonably suppose that $p > .7$. In fact, we may say, "Taking account of sampling fluctuations only, p is greater than .7 unless a chance with probability less than three in one hundred has occurred in the experiment."

The last statement is an inductive inference. Somewhat more useful inferences could be made by taking account of the actual measurements of, say, the losses in weight, but this simple one will illustrate the points we wish to make here. While we say that $p > .7$, we admit the possibility that we may be wrong, and we give a measure, .03, of the maximum probability that we may be in error. By increasing the maximum probability of error we could narrow the range for p. Thus we might say $p > .9$ unless a chance with probability less than .103 has occurred. The size of the probability of error is a matter of taste to a large extent. Some investigators commonly use .05 while others wish to be more conservative and use .01 or .001.

It is to be observed that the probability of error measures only the error due to random sampling fluctuations. We have not said any-

thing about the possible accidents that were mentioned earlier. And in fact it is impossible to say what the probability of such accidents may be. The nutritionist can only say something like this: "Barring accidents, $p > .7$ for the group of individuals from which the ten were selected, unless a chance with probability less than .03 has occurred in the experiment."

We may mention one other point here. Referring to the same experiment, is it possible to conclude without error that $p > 0$? The answer to this is "Yes" in theory but generally "No" in practice. The accidents that may have occurred rule out an inference of this kind. An experimenter willingly assumes that he performs his experiments with such care that the probability of accidents is negligible in comparison with the probability of his sampling errors, but he cannot assume that accidents are impossible.

The theory of statistics thus has a part in any inductive inference based on experimental data. Its role is to provide a measure, in terms of probability, of the uncertainty of the inference. The measure will be based entirely on sampling errors. It is up to the experimenter to guard against accidents which may invalidate his results, and the theory of statistics makes no attempt to deal with this aspect of the problem of inference.

7.2. Populations and Samples. The word *population* in statistics is used to refer to any collection of objects or results of operations. Thus we may speak of the population of dairy cattle in Wisconsin, the population of prices of bread in the City of New York, the population of mileages of automobile tires, the hypothetical population of heads and tails obtained by tossing a coin an infinite number of times, the hypothetical population of an infinite number of measurements of the velocity of light, and so forth.

The problem of inductive inference is regarded as follows from the point of view of statistics: The object of an experiment is to find out something about some specified population. It is impossible or impracticable to examine the entire population, but one may examine a part or sample of it, and on the basis of this limited investigation make inferences regarding the whole population.

It is important that the sample be chosen from the population it is desired to study. This obvious principle is violated surprisingly often. Thus in the nutrition example mentioned above, if the nutritionist wishes to make an inference about the population of the United States, his ten subjects must be randomly selected from that population. If, in fact, the ten subjects were chosen from among thirty students in

one of his classes in home economics, then he has studied a very limited population indeed. He can make a rigorous inference only concerning the thirty students. Actually, of course, he would probably extend his results to cover a larger population with considerable justification. He could argue that the mere fact that the ten subjects happened to be taking a particular course in home economics could not conceivably influence the experiment and that the results could certainly be taken as representative of all women students in the college. And from other experiments he may assume that sex has no effect on reactions to diets and claim his results apply to men students as well. He may generalize further and say the results reasonably represent all people of college age in the region from which the college draws most of its students. But here he might be getting on shaky ground, because it is well known that college students come from the wealthier and hence better nourished families in the region. It is even more doubtful if the results could be taken as representative of the whole adult population of the region. And it would be completely unjustifiable to claim that the results are valid for the adult population of the whole nation, because reactions to a given diet depend on the normal diet, which is quite different in different regions.

Extension of the population originally studied to a larger population increases the probability of error by an unknown amount and thus destroys the measure of confidence to be placed in the inference. The careful investigator does not indulge in this practice, but chooses his sample from the entire population he wishes to study if it is at all practicable. For example, the nutritionist, if he wishes to make an inference about the adult population of the nation, might actually select, by some device or other, a random sample of individuals from the whole adult population and then enlist the aid of colleagues who happen to live near the individuals selected.

We have implied that a sample must be random. It is this property of a sample that enables one to compute the probability of error of his inference. The theory of probability cannot be applied to a nonrandom sample, so that there is no way to measure the degree of confidence to be placed in any inference from such a sample. The word random refers to the manner in which the sample is selected rather than to the particular sample. Any possible sample is a random sample. Thus a person may shuffle a deck of cards thoroughly and then blindly draw four cards from it, thus obtaining a random sample of four cards. If, in fact, it turned out that he drew the four aces, then he obtained a very unrepresentative sample of denominations, but still it was a

random sample by virtue of the method by which it was obtained. Similarly, the nutritionist may have carefully drawn a random sample of ten adults and obtained unfavorable reactions to his diet in all cases. It may be, in fact, that only a small proportion of individuals in the population would have such a reaction and that the nutritionist was particularly unlucky in his sample. The margin of error given in his inference measures the probability of such a contingency.

An investigator hopes, by drawing a random sample, to get a fairly representative portion of the population he wishes to study. Often it is possible to introduce a certain amount of nonrandomness in the sampling procedure to obtain partial assurance of a representative sample. This can be done when something is known about the population. Thus a public-opinion agency may wish to take a preelection poll of the United States. It knows the populations of the various states and can assure itself a degree of representativeness by allocating its sample to states according to the populations of the states. Thus, if 1 per cent of population is in a given state, 1 per cent of the sample will be taken in that state. Within the state further allocations may be made. The sample may be evenly divided between the sexes. The proportions of urban and rural dwellers may be forced to agree with the actual known proportions within the state. The effect here is to divide the population into a great many smaller populations. But somewhere along the line random samples of the subpopulations must be taken, if inferences with measurable uncertainty are to be made.

7.3. Sample Distributions. Suppose a variate x has density $f(x)$ in some population. And suppose a sample of two values of x, say x_1 and x_2, are drawn at random. The pair of numbers (x_1, x_2) determine a point in a plane, and the population of all such pairs of numbers that might have been drawn forms a bivariate population. We are interested in finding the distribution of this bivariate population in terms of the original distribution $f(x)$.

The joint density function for x_1 and x_2 must be some function, say $f(x_1, x_2)$, such that for any a_1, a_2, b_1, b_2 we have

$$P(a_1 < x_1 < b_1, a_2 < x_2 < b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_2\, dx_1 \qquad (1)$$

Now by a random sample we shall mean that the value of the first observation x_1 has no effect whatever on the value of the second observation. In other words, for a random sample, x_1 and x_2 are independent in the probability sense. When the two variates of a bivariate

distribution are independent in the probability sense, we have seen that the joint distribution is the product of the marginal distributions. In the present instance, the marginal distributions are simply $f(x_1)$ and $f(x_2)$, so that we have, by definition of randomness,

$$f(x_1, x_2) = f(x_1)f(x_2) \tag{2}$$

or, what is the same thing,

$$P(a_1 < x_1 < b_1, a_2 < x_2 < b_2) = P(a_1 < x_1 < b_1)P(a_2 < x_2 < b_2) \tag{3}$$

As a simple example, suppose x can have only two values, zero and one, with probabilities q and p, respectively. That is, x is a discrete variate which has the binomial distribution

$$f(x) = \binom{1}{x} p^x q^{1-x} \qquad x = 0, 1 \tag{4}$$

and since $\binom{1}{0} = \binom{1}{1} = 1$, we may write it as

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

The joint density for samples of two values of x is

$$f(x_1, x_2) = p^{x_1+x_2} q^{2-x_1-x_2} \qquad x_1 = 0, 1, x_2 = 0, 1 \tag{5}$$

which is defined at the four points (0, 0)(0, 1)(1, 0)(1, 1) in the x_1, x_2 plane. It is to be observed that this density is not what we should have obtained by drawing two elements from a binomial population and counting the number of successes, say y; that density is

$$f(y) = \binom{2}{y} p^y q^{2-y} \qquad y = 0, 1, 2, \tag{6}$$

and it differs from (5) in that it is the distribution of the single variate $x_1 + x_2$. Equation (5) gives us the joint distribution of the two random variates x_1 and x_2.

It is to be noted that $f(x_1, x_2)$ gives us the distribution of the sample in *the order drawn*. Thus in (5), $f(0, 1) = pq$ not $2pq$. $f(0, 1)$ refers to the probability of drawing first a zero, then a one. And in general, (1) represents the probability that the first observation drawn falls in the interval (a_1, b_1) and the second falls in (a_2, b_2). The opposite occurrence does not satisfy the specification unless, of course, the two intervals happen to be the same.

By reasoning exactly as before, we find that the joint density for a random sample of size n, $x_1, x_2, \cdots, x_n$, from a population with

distribution $f(x)$ is

$$f(x_1, x_2, \cdots, x_n) = f(x_1)f(x_2) \cdots f(x_n) \tag{7}$$

and this again gives the distribution of the sample in the order drawn.

Our definition of random sampling has automatically ruled out sampling without replacement from a finite population. If, for example, we draw two balls from an urn containing, say, two white and three black balls, the result of the first draw certainly affects the probability of the result of the second. The two drawings are not independent in the probability sense. In this case, another definition of random sampling must be adopted (Probs. 26 and 32). Our present discussion in this and in the following chapters is thus concerned with sampling from continuous populations (where the question of drawing with or without replacement does not arise) and to sampling with replacement from finite populations.

7.4. Sample Moments. If $x_1, x_2, \cdots, x_n$ are a sample of n values drawn from a population with density $f(x)$, the rth sample moment is defined to be

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \qquad r = 1, 2, \cdots \tag{1}$$

m'_1 is called the sample mean and is more often designated by $\bar{x}$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

We shall show that m'_r may be taken to be an estimate of the population moment μ'_r.

Suppose $g(x)$ is any function of x; then the expected value of the function is

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{3}$$

We shall see that for a large sample, $x'_1, x'_2, \cdots, x'_n$, the expression

$$\frac{1}{n}\sum_{i=1}^{n} g(x'_i)$$

may be expected to approximate $E[g(x)]$. Let the area under $f(x)$ be divided into strips of width Δx_j, and let n_j be the number of sample elements which fall in Δx_j with $\Sigma n_j = n$. Let x_j be the mid-point of the interval Δx_j. Then if the Δx_j are small, all the x_i which fall in Δx_j will not differ much from x_j, and we may write

$$\frac{1}{n}\sum_i g(x_i') \cong \frac{1}{n}\sum_j n_j g(x_j) \tag{4}$$

Now the area over any Δx_j is approximately $f(x_j)\Delta x_j$, and it is the probability, say p_j, that any randomly drawn value of x will fall in Δx_j. If a sample of n values of x is drawn, we expect np_j of the sample values to fall in Δx_j. It follows then that n_j/n is an estimate of p_j, and (4) may be written

$$\frac{1}{n}\sum g(x_i') \cong \sum \frac{n_j}{n} g(x_j)$$
$$\cong \Sigma g(x_j) f(x_j)\Delta x_j$$

This last sum approximates the integral in (3).

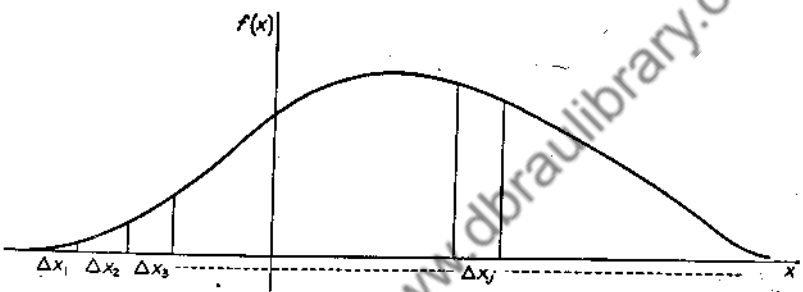


FIG. 34.

The above argument is merely heuristic and does not prove anything. It does give some insight, however, into the way in which samples provide information about distributions. We can prove directly that the expected value of $(1/n)\Sigma g(x_i)$ is $E[g(x)]$. (We now drop the primes from the x_i; they were used above to distinguish the sample values from the mid-points of the intervals.) The joint density of the $x_1, x_2, \cdots, x_n$ is

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) \tag{5}$$

hence the expected value of the sum is

$$E\left[\frac{1}{n}\sum g(x_i)\right] = \int\int \cdots \int \frac{1}{n}\sum g(x_i) \prod_{i=1}^{n} f(x_i) \prod_{i=1}^{n} dx_i \tag{6}$$

This integral may be written as the sum of n integrals of the form

$$\frac{1}{n}\int\int \cdots \int g(x_i) \prod_j [f(x_j)dx_j]$$

which in turn may be written as the product of n integrals, all but one of which are of the form

$$\int f(x_j)dx_j = 1$$

and the remaining one is

$$\frac{1}{n}\int g(x_i)f(x_i)dx_i = \frac{1}{n}E[g(x)] \tag{7}$$

Since (6) is the sum of n such integrals, we have

$$E\left[\frac{1}{n}\sum g(x_i)\right] = E[g(x)] \tag{8}$$

On choosing $g(x)$ to be x^r, we find that the expected value of the rth sample moment is the rth population moment,

$$\begin{aligned} E(m_r') &= E\left(\frac{1}{n}\sum x_i^r\right) \\ &= E(x^r) \\ &= \mu_r' \end{aligned} \tag{9}$$

We may review the meaning of this result. The sample moment m_r' is a function of n random variables and is therefore a random variable itself. As such, it has some probability distribution, and equation (9) shows that the mean value of that distribution is μ_r'. We do not therefore suppose that m_r' is in any sense equal to μ_r' for a given sample; it is simply a random variable whose mean is μ_r'. We shall speak of m_r' as being an estimate of μ_r'. Whether or not it will be an accurate estimate depends on how closely the distribution of m_r' is concentrated about its mean.

Corresponding to the population moments μ_r about the mean, we may define sample moments about the sample mean as follows:

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$$

We have $m_1 = 0$ just as $\mu_1 = 0$, since

$$\begin{aligned} m_1 &= \frac{1}{n}\sum(x_i - \bar{x}) \\ &= \frac{1}{n}\sum x_i - \frac{1}{n}\sum \bar{x} \\ &= \bar{x} - \bar{x} = 0 \end{aligned}$$

The m_r may be regarded as estimates of the μ_r in the same sense that m_r' estimate μ_r'; however, they are *biased* estimates. That is, it is not

true that

$$E(m_r) = \mu_r$$

except when $r = 1$. We shall illustrate this fact for $r = 2$ in Probs. 12 and 13.

7.5. The Law of Large Numbers. We have seen that the expected value of a sample mean is the population mean.

$$E(\bar{x}) = \mu \tag{1}$$

Let us find the variance of the sample mean

$$\begin{aligned}\sigma_{\bar{x}}^2 &= E(\bar{x} - \mu)^2 \\ &= E\left(\frac{1}{n}\sum x_i - \mu\right)^2 \\ &= E\left[\frac{1}{n}\sum (x_i - \mu)\right]^2 \\ &= \frac{1}{n^2} E\left[\sum (x_i - \mu)^2\right]\end{aligned} \tag{2}$$

On squaring the sum, we get n terms of the form $(x_i - \mu)^2$ and $\binom{n}{2}$ terms of the form $2(x_i - \mu)(x_j - \mu)$ with $i \neq j$. The expected value of $(x_i - \mu)^2$ depends only on the marginal distribution of x_i, since in the integral

$$\int\int \cdots \int (x_i - \mu)^2 \prod_j [f(x_j)dx_j]$$

all factors not involving x_i become one and we are left with

$$\int (x_i - \mu)^2 f(x_i)dx_i = \sigma^2 \tag{3}$$

where σ^2 is the population variance. Similarly,

$$\begin{aligned}E[(x_i - \mu)(x_j - \mu)] &= \int\int (x_i - \mu)(x_j - \mu)f(x_i)f(x_j)dx_i\, dx_j \\ &= \int (x_i - \mu)f(x_i)dx_i \int (x_j - \mu)f(x_j)dx_j \\ &= 0\end{aligned} \tag{4}$$

Equation (2) then becomes

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{1}{n^2}\sum_{i=1}^{n} E(x_i - \mu)^2 \\ &= \frac{1}{n^2}\sum \sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned} \tag{5}$$

Thus the variance of the sample mean is equal to the population variance divided by the sample size; this is true for any population with a finite variance.

This fact is of extreme importance in applied statistics. It implies that whatever the population distribution (provided it has a finite variance), the distribution of the sample mean becomes more and more concentrated near the population mean as the sample size increases. It follows that the larger the sample, the more certain we can be that the sample mean will be a good estimate of the population mean. This is essentially the law of large numbers. We shall obtain a more precise statement of it below.

Suppose the density of the sample mean is $g(\bar{x})$, where $\bar{x}$ is the mean of a sample of size n from a population with density $f(x)$. We have

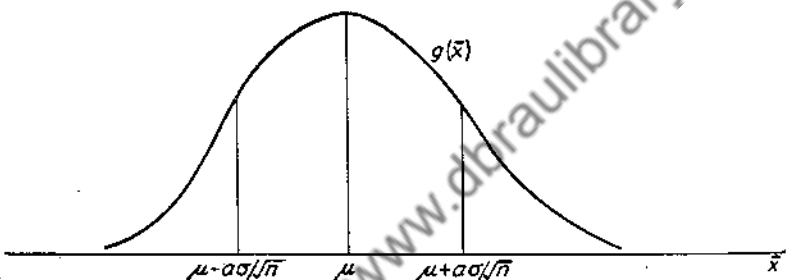


FIG. 35.

found that the mean and variance of $g(\bar{x})$ are μ and σ^2/n, where μ and σ^2 are the mean and variance of $f(x)$. It follows from the definition of the variance that

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \int_{-\infty}^{\infty} (\bar{x} - \mu)^2 g(\bar{x}) d\bar{x} \tag{6}$$

Now let us break up the range of integration into three parts, as illustrated in Fig. 35:

$$\frac{\sigma^2}{n} = \int_{-\infty}^{\mu-(a\sigma/\sqrt{n})} (\bar{x} - \mu)^2 g(\bar{x}) d\bar{x} + \int_{\mu-(a\sigma/\sqrt{n})}^{\mu+(a\sigma/\sqrt{n})} (\bar{x} - \mu)^2 g(\bar{x}) d\bar{x} + \int_{\mu+(a\sigma/\sqrt{n})}^{\infty} (\bar{x} - \mu)^2 g(\bar{x}) d\bar{x} \tag{7}$$

where a is any arbitrarily chosen positive number. We are going to obtain an inequality by reducing the right-hand side of equation (7). We shall discard the second integral, and since it is positive, the right-hand side will be decreased. Also in the first integral we shall replace

the factor $(\bar{x} - \mu)^2$ by $a^2\sigma^2/n$. This will clearly reduce the value of the integral, since in the range of integration

$$|\bar{x} - \mu| \geq \frac{a\sigma}{\sqrt{n}}$$

The same substitution will reduce the third integral also. We shall have then

$$\frac{\sigma^2}{n} > \frac{a^2\sigma^2}{n} \int_{-\infty}^{\mu-(a\sigma/\sqrt{n})} g(\bar{x})d\bar{x} + \frac{a^2\sigma^2}{n} \int_{\mu+(a\sigma/\sqrt{n})}^{\infty} g(\bar{x})d\bar{x} \tag{8}$$

or, what is the same thing,

$$\frac{1}{a^2} > P\left(|\bar{x} - \mu| > \frac{a\sigma}{\sqrt{n}}\right) \tag{9}$$

since the two integrals in (8) give exactly the probability that $\bar{x}$ lies outside the interval $\mu - \left(\frac{a\sigma}{\sqrt{n}}\right)$ to $\mu + (a\sigma/\sqrt{n})$.

Now in (9) let $a\sigma/\sqrt{n} = b$; then $1/a^2 = \sigma^2/nb^2$, and (9) becomes

$$P\left(|\bar{x} - \mu| > b\right) < \frac{\sigma^2}{nb^2} \tag{10}$$

This relation is known as Tchebysheff's inequality. It may be written in the alternative form

$$P(-b < \bar{x} - \mu < b) > 1 - \frac{\sigma^2}{nb^2} \tag{11}$$

Tchebysheff's inequality gives a precise formulation of the law of large numbers. Referring to (11), we may choose any small number b and determine a small interval about the population mean; having done this, we may choose n large enough to give a value as near one as we please for the probability that the sample mean will lie within the small interval containing the population mean.

To consider an example, suppose some distribution with an unknown mean has a variance equal to one. How large a sample must be taken in order that the probability will be at least .95 that the sample mean will lie within .5 of the population mean? We have $\sigma^2 = 1$, $b = .5$, and we wish to choose n so that $1 - \sigma^2/nb^2$ will be .95.

$$1 - \frac{\sigma^2}{nb^2} = .95$$

whence

$$n = \frac{\sigma^2}{.05b^2} = \frac{1}{.05(.5)^2} = 80$$

The example is not realistic because the variance is assumed to be known. Later we shall have to consider ways of circumventing this difficulty. The important thing here is the indication of the possibility of making very accurate and reliable inferences provided large samples can be obtained.

7.6. The Central-limit Theorem. The central-limit theorem gives a still more precise statement of the law of large numbers. It is the most important theorem in statistics from both the theoretical and applied points of view. And it is one of the most remarkable theorems in the whole of mathematics. A great many eminent mathematicians (De Moivre, Laplace, Gauss, Tchebysheff, Liapounoff, Levy, Cramer, and others) have contributed to its development. The theorem is this:

If a population has a finite variance σ^2 and mean μ, then the distribution of the sample mean approaches the normal distribution with variance σ^2/n and mean μ as the sample size n increases.

The astonishing thing about the theorem is the fact that nothing is said about the form of the population distribution function. Whatever the distribution function, provided only that it have a finite variance, the sample mean will have approximately the normal distribution for large samples. The condition that the variance be finite is not a critical restriction so far as applied statistics is concerned, because in almost any practical situation the range of the variate will be finite, in which case the variance must necessarily be finite.

We shall not be able to prove this theorem, because it requires rather advanced mathematical techniques. However, in order to make the theorem plausible, we shall consider an argument for the more restricted situation in which the distribution has a moment generating function. The argument will be essentially a matter of showing that the moment generating function for the sample mean approaches the moment generating function for the normal distribution. We shall first obtain the moment generating function for

$$y = \frac{x' - \mu'}{\sigma'}$$

when x is normally distributed. The generating function is

$$m_1(t) = \int_{-\infty}^{\infty} e^{ty} n(x'; \mu', \sigma'^2) dx' \tag{1}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{t(x'-\mu')/\sigma'} e^{-\frac{1}{2}(x'-\mu')^2/\sigma'^2} dx' \tag{2}$$

and as in section 6.2 we find

$$m_1(t) = e^{\frac{1}{2}t^2} \tag{3}$$

Now suppose x has some arbitrary density function $f(x)$ with mean μ and variance σ^2 which has a moment generating function. The moment generating function of $(x - \mu)/\sigma$, say $m_2(t)$, is defined as

$$m_2(t) = \int_{-\infty}^{\infty} e^{t(x-\mu)/\sigma} f(x)dx \tag{4}$$

A sample of size n will have a mean with some distribution, say $g(\bar{x})$, which we have seen must have mean μ and variance σ^2/n. The moment generating function for

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{5}$$

say $m_3(t)$, is defined as

$$m_3(t) = \int_{-\infty}^{\infty} e^{t\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}} g(\bar{x})d\bar{x} \tag{6}$$

It is our purpose to show that $m_3(t)$ must approach $m_1(t)$ when n, the sample size, becomes large.

We can determine $m_3(t)$ in terms of $m_2(t)$. $m_3(t)$ is the expected value,

$$E\left(e^{t\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}\right) = E\left(e^{\frac{t}{n}\sum\frac{x_i-\mu}{\sigma/\sqrt{n}}}\right)$$

and since we know that the joint distribution of the $x_1, x_2, \cdots, x_n$ is $\prod_{i=1}^{n} f(x_i)$, we may write

$$m_3(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{\frac{t}{\sqrt{n}}\sum\frac{x_i-\mu}{\sigma}} \prod_{i=1}^{n} f(x_i)dx_i$$

$$= \prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} e^{\frac{t}{\sqrt{n}}\cdot\frac{x_i-\mu}{\sigma}} f(x_i)dx_i\right] \tag{7}$$

and by virtue of (4), each factor in this product is simply $m_2(t/\sqrt{n})$; hence

$$m_3(t) = \left[m_2\left(\frac{t}{\sqrt{n}}\right)\right]^n \tag{8}$$

The rth derivative of $m_2(t/\sqrt{n})$ evaluated at $t = 0$ obviously gives us the rth moment about the mean divided by $(\sigma\sqrt{n})^r$. And we have seen in Sec. 5.3 that we may write

$$m_2\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{\mu_1}{\sigma}\frac{t}{\sqrt{n}} + \frac{1}{2!}\frac{\mu_2}{\sigma^2}\left(\frac{t}{\sqrt{n}}\right)^2 + \frac{1}{3!}\frac{\mu_3}{\sigma^3}\left(\frac{t}{\sqrt{n}}\right)^3 + \cdots \tag{9}$$

and since $\mu_1 = 0$, $\mu_2 = \sigma^2$, this may be written

$$m_2\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{1}{n}\left(\frac{1}{2}t^2 + \frac{1}{3!\sqrt{n}}\frac{\mu_3}{\sigma^3}t^3 + \frac{1}{4!n}\frac{\mu_4}{\sigma^4}t^4 + \cdots\right) \quad (10)$$

If we recall that the definition of e^u is

$$e^u = \lim_{n\to\infty}\left(1 + \frac{u}{n}\right)^n$$

we see that $m_3(t)$, as n becomes infinite, becomes of exactly this form, where u represents the parenthesis in (10), and when n becomes infinite, all terms in u vanish except the first, so we have

$$\lim_{n\to\infty} m_3(t) = e^{\frac{1}{2}t^2} \quad (11)$$

Hence in the limit z has the same moment generating function as y and, by virtue of the statement at the end of Sec. 5.4, has the same

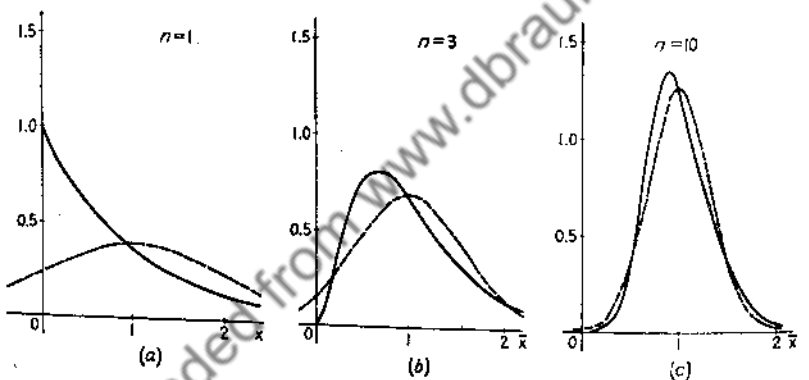


FIG. 36.

distribution. Thus in the limit the sample mean must have the normal distribution whatever the distribution $f(x)$, provided that $f(x)$ has a moment generating function, or more generally, provided that $f(x)$ has a second moment. And for large n, we may say that the sample mean is approximately normally distributed.

The degree of approximation depends, of course, on the sample size and on the particular density function $f(x)$. The approach to normality is illustrated in Fig. 36 for the particular function $f(x) = e^{-x}$, $x > 0$. The solid curves give the actual distributions, while the dashed curves give the normal approximations. (*a*) gives the original distribution which corresponds to samples of one; (*b*) shows the distribution of sample means for $n = 3$; (*c*) gives the distribution of

sample means for $n = 10$. The curves rather exaggerate the approach to normality because they cannot show what happens on the tails of the distribution. Ordinarily distributions of sample means approach normality fairly rapidly with the sample size in the region of the mean, but more slowly at points distant from the mean; usually the greater the distance of a point from the mean, the more slowly the normal approximation approaches the actual distribution.

The central-limit theorem applies to discrete as well as to continuous distributions. The moment generating functions used in this section could have been moment generating functions for discrete distributions, and the argument would have been just the same except that the integrals would have been replaced by sums. We shall investigate the nature of this approximation in the next section for a particular discrete distribution.

7.7. Normal Approximation to the Binomial Distribution. We shall consider the density

$$f(x) = p^x q^{1-x} \qquad x = 0, 1 \tag{1}$$

which has

$$\mu = p \qquad \sigma^2 = pq \tag{2}$$

and suppose a sample, $x_1, x_2, \cdots, x_n$, of size n is drawn. The sample will simply be a sequence of zeros and ones in this instance, one denoting a success, say, and zero a failure. And

$$\bar{x} = \frac{1}{n} \sum x_i$$

is the proportion of successes in the sample. We have seen that the mean and variance of $\bar{x}$ are

$$E(\bar{x}) = \mu = p \tag{3}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{pq}{n} \tag{4}$$

The distribution of $\bar{x}$ is discrete; in fact $\bar{x}$ can take on only the values

$$0, \frac{1}{n}, \frac{2}{n}, \cdots, \frac{j}{n}, \cdots, 1$$

and we know that the density of j is

$$f(j) = \binom{n}{j} p^j q^{n-j} \qquad j = 0, 1, 2, \cdots, n \tag{5}$$

Thus since $j = n\bar{x}$, the density of $\bar{x}$ is

$$h(\bar{x}) = \binom{n}{n\bar{x}} p^{n\bar{x}} q^{n(1-\bar{x})} \qquad \bar{x} = 0, \frac{1}{n}, \frac{2}{n}, \cdots, 1 \tag{6}$$

The way in which this discrete density is approximated by a continuous density function is illustrated in Fig. 37.

Suppose we construct rectangles of heights $h(\bar{x})$ and widths $1/n$ with mid-points of the bases at j/n, $j = 0, 1, 2, \cdots, n$. The tops of

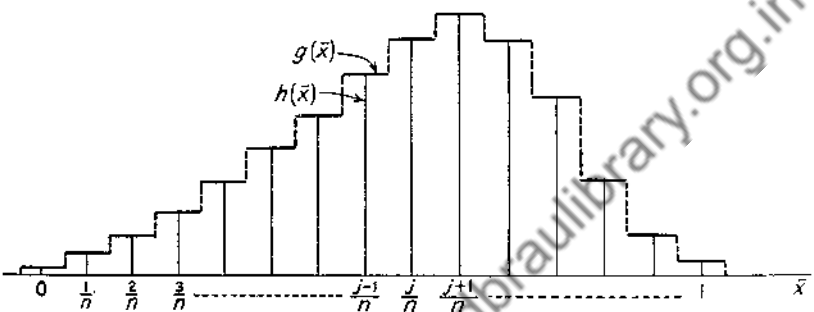


FIG. 37.

these rectangles form a broken curve which we may represent by $g(\bar{x})$. Since $\Sigma h(\bar{x}) = 1$, the area under $g(\bar{x})$ will be $1/n$. It is clear that

$$P\left(\frac{a}{n} \leq \bar{x} \leq \frac{b}{n}\right) = n \int_{(a-\frac{1}{2})/n}^{(b+\frac{1}{2})/n} g(\bar{x}) d\bar{x} \tag{7}$$

for any integers a and b $(b \geq a)$ in the range of j, since the integral is simply the area under the tops of the rectangles over the points a to b and is therefore

$$\sum_{\bar{x}=a/n}^{b/n} h(\bar{x}) \frac{1}{n} = \frac{1}{n} \sum_{j=a}^{b} \binom{n}{j} p^j q^{n-j} \tag{8}$$

As n becomes large, the width of the rectangles decreases and the steps in the function $ng(\bar{x})$ become closer together so that it has the appearance, say, of the function in Fig. 38. The normal approximation to the binomial distribution may be regarded as the limiting form of this broken curve as n becomes infinite.

This normal approximation is of particular interest because it provides a method of computing easily the approximate value of sums of the binomial distribution. As an illustration, let us suppose a true die is cast and a one or a two counted as a success. Then $p = \frac{1}{3}$,

$q = \frac{2}{3}$. For a sample of 300 trials, the total number of successes, j, has the density

$$f(j) = \binom{300}{j}\left(\frac{1}{3}\right)^j\left(\frac{2}{3}\right)^{300-j} \qquad j = 0, 1, \cdots, 300$$

Suppose we wanted the probability that the number of successes will not deviate from 100 by more than 15; we should have to sum $f(j)$

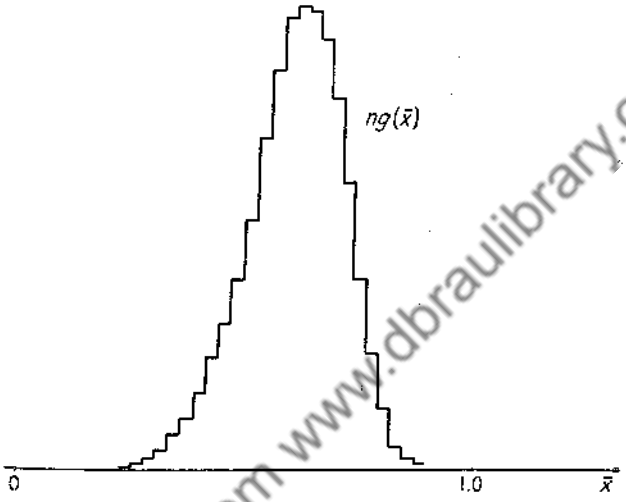


FIG. 38.

over the range 85 to 115, a very tedious calculation. We can approximate the sum by using the fact that

$$P(85 \le j \le 115) = P\left(\frac{85}{300} \le \frac{j}{300} \le \frac{115}{300}\right)$$

and since $\bar{x} = j/300$ is approximately normally distributed with mean $\frac{1}{3}$ and variance $\frac{1}{3} \times \frac{2}{3} \times \frac{1}{300}$, we have

$$P(\tfrac{85}{300} \le \bar{x} \le \tfrac{115}{300}) \cong \int_{\frac{85}{300}}^{\frac{115}{300}} n(\bar{x}; \tfrac{1}{3}, \tfrac{2}{2700})d\bar{x}$$

$$\cong \int_{\frac{85}{300}}^{\frac{115}{300}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{2}{2700}}} e^{-\frac{1}{2}(\bar{x}-\frac{1}{3})^2/\frac{2}{2700}}\, d\bar{x}$$

and letting $t = (\bar{x} - \frac{1}{3})/\sqrt{\frac{2}{2700}}$, we have

$$P\left(\frac{85}{300} \le \bar{x} \le \frac{115}{300}\right) \cong \int_{-1.84}^{1.84} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt$$

since

$$\frac{(85/300 - 1/3)}{\sqrt{2/2700}} \cong -1.84 \qquad \frac{(115/300 - 1/3)}{\sqrt{2/2700}} \cong 1.84$$

Using tables of the normal distribution, we find

$$P(85 \leq j \leq 115) \cong .934 \tag{9}$$

The approximation could be slightly improved by using $85 - \frac{1}{2}$ and $115 + \frac{1}{2}$ in computing limits on the integral as indicated by (7).

In general, for the binomial distribution, it is now evident that

$$P(a \leq j \leq b) = \sum_{j=a}^{b} \binom{n}{j} p^j q^{n-j} \tag{10}$$

$$= \int_{a'}^{b'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt \tag{11}$$

where

$$a' = \frac{[(a - \frac{1}{2})/n] - p}{\sqrt{pq/n}} \qquad b' = \frac{[(b + \frac{1}{2})/n] - p}{\sqrt{pq/n}} \tag{12}$$

A more detailed investigation would show that the error in this approximation is less than

$$\frac{.15}{\sqrt{npq}} \tag{13}$$

provided $npq > 25$. Thus in the above example our maximum error is measured by

$$\frac{.15}{\sqrt{300 \times 1/3 \times 2/3}} = .018$$

so that the approximation (9) does not quite have two-place accuracy in so far as we can judge by (13). More accurate approximations are provided by Uspensky ("Introduction to Mathematical Probability," Chap. VII, McGraw-Hill Book Company, Inc., New York, 1937).

7.8. Role of the Normal Distribution in Statistics. It will be found in the ensuing chapters that the normal distribution plays a very predominant part. Of course, the central-limit theorem alone ensures that this will be the case, but there are other almost equally important reasons.

In the first place, many populations encountered in the course of research in many fields seem to have a normal distribution to a good degree of approximation. It has often been argued that this phenomenon is quite reasonable in view of the central-limit theorem. We may consider the firing of a shot at a target as an illustration. The

course of the projectile is affected by a great many factors all admittedly with small effect. The net deviation is the net effect of all these factors. Suppose the effect of each factor is an observation from some population; then the total effect is essentially the mean of a set of observations from a set of populations. Being of the nature of means, the actually observed deviations might therefore be expected to be approximately normally distributed. We do not intend to imply here that most distributions encountered in practice are normal, for such is not the case at all, but nearly normal distributions are encountered quite frequently.

Another consideration which favors the normal distribution is the fact that sampling distributions based on a parent normal distribution are fairly manageable analytically. In making inferences about populations from samples it is necessary to have the distributions for various functions of the sample observations. The mathematical problem of obtaining these distributions is often easier for samples from a normal population than from any other.

Because all these auxiliary distributions are required in statistical inference, the economical thing to do is obtain them for one kind of population distribution only. When another kind of population is under examination, the observations may be transformed so that they follow the distribution first chosen. The normal distribution is the logical candidate for this choice. Thus if a complete theory of statistical inference is developed based on the normal distribution alone, then one has in reality a system which may be employed quite generally, because other distributions can be transformed to the normal form.

In applying statistical methods based on the normal distribution, the experimenter must know, at least approximately, the general form of the distribution function which his data follow. If it is normal, he may use the methods directly; if it is not, he may transform his data so that the transformed observations follow a normal distribution. When the experimenter does not know the form of his population distribution, then he must use other more general but usually less powerful methods of analysis called *distribution-free* methods. Some of these methods will be presented in the final chapter of the book.

7.9. Problems

1. In the joint distribution $p^{x_1+x_2}q^{2-x_1-x_2}$, for a sample of two from a binomial population, let $x_1 = y - x_2$ and find the joint distribution of y and x_2.

2. Find the marginal distribution of y from the results of the above problem.

3. What is the probability that the two observations of a sample of two from a population with a rectangular distribution over the unit interval will not differ by more than one-half?

4. What is the probability that the mean of a sample of two observations from a rectangular distribution (over the unit interval) will be between $\frac{1}{4}$ and $\frac{3}{4}$?

5. What is the probability that the larger of two random observations from any continuous distribution will exceed the median?

6. If x_1 and x_2 are a sample of two from a population with density $f(x)$, and if the smaller of these values is denoted by y_1 and the larger by y_2, what is the joint density of y_1 and y_2?

7. Generalize the result of Prob. 6 to samples of size n, letting y_1 be the smallest and y_2 the largest of the n observations.

8. What is the marginal density of the smallest observation for samples of size n?

9. Considering random samples of size n from a population with density $f(x)$, what is the expected value of the area under $f(x)$ to the left of the smallest sample observation?

10. Balls are drawn with replacement from an urn containing one white and two black balls. Let $x = 0$ for a white ball and $x = 1$ for a black ball. For samples $x_1, x_2, \cdots, x_9$ of size nine, what is the joint distribution of the observations? The distribution of the sum of the observations?

11. Referring to Prob. 10, find the expected values of the sample mean and sample variance.

12. For samples of size two from a population with variance σ^2, show that the expected value of the sample variance is $\sigma^2/2$.

13. Generalize the result of Prob. 12 to samples of size n.

14. What value of y minimizes $\sum_1^n (x_i - y)^2$?

15. If $\bar{x} = (1/n) \sum_1^n x_i$, show that

$$\sum_1^n (x_i - \mu)^2 = \sum_1^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

Using this result and that of Prob. 14, explain why the sample variance gives a biased estimate of the population variance.

16. Find $E(m_3)$ for samples of size two from a population with a finite third moment.

17. Show that $E[(1/n)\Sigma(x_i - \mu)^r] = \mu_r$ for samples of size n from a population with mean μ and rth moment μ_r.

18. Use Tchebysheff's inequality to find how many times a coin must be tossed in order that the probability will be at least .90 that $\bar{x}$ will lie between .4 and .6. (Assume the coin is true.)

19. How could one determine the number of tosses required in Prob. 18 more accurately, i.e., make the probability very nearly equal to .90? What is the number of tosses?

20. If a population has $\sigma = 2$ and $\bar{x}$ is the mean of samples of size 100, find limits between which $\bar{x} - \mu$ will lie with probability .90. Use both Tchebysheff's inequality and the central-limit theorem. Why do the two results differ?

21. Suppose x_1 and x_2 are means of two samples of size n from a population with variance σ^2. Determine n so that the probability will be about .01 that the two sample means will differ by more than σ. (Consider the variate $y = \bar{x}_1 - \bar{x}_2$.)

22. Suppose light bulbs made by a standard process have an average life of 2000 hours with a standard deviation of 250 hours. And suppose it is considered worth while to replace the process if the mean life can be increased by at least 10 per cent. An engineer wishes to test a proposed new process, and he is willing to assume that the standard deviation of the distribution of lives is about the same as for the standard process. How large a sample should he examine if he wishes the probability to be about .01 that he will fail to adopt the new process if in fact it produces bulbs with a mean life of 2250 hours?

23. A research worker wishes to estimate the mean of a population using a sample large enough that the probability will be .95 that the sample mean will not differ from the population mean by more than 25 per cent of the standard deviation. How large a sample should he take?

24. A polling agency wishes to take a sample of voters in a given state large enough that the probability is only .01 that they will find the proportion favoring a certain candidate to be less than 50 per cent when in fact it is 52 per cent. How large a sample should be taken?

25. A standard drug is known to be effective in about 80 per cent of cases in which it is used to treat infections. A new drug has been found effective in 85 of the first 100 cases tried. Is the superiority of the new drug well established? (If the new drug were equally effective as the old, what would be the probability of obtaining 85 or more successes in a sample of 100?)

26. A bowl contains five chips numbered from one to five. A sample of two drawn without replacement from this finite population is said

to be random if all possible pairs of the five chips have an equal chance to be drawn. What is the expected value of the sample mean? What is the variance of the sample mean?

27. Suppose the two chips of Prob. 26 were drawn with replacement, what would be the variance of the sample mean? Why might one guess that this variance would be larger than the one obtained before?

28. If a density $f(x)$ has a moment generating function $m(t)$, show that the mean of samples of size n has the moment generating function $[m(t/n)]^n$.

29. Use the result of Prob. 28 to show that the mean and variance of the sample mean are μ and σ^2/n.

30. Find the third moment of the sample mean for samples of size n from a binomial population. Show that it approaches zero as n becomes large (as it must if the normal approximation is to be valid).

31. Suppose the life of a certain part of a machine is distributed by $.01e^{-.01t}$ where t is measured in days. The machine comes supplied with one spare. What is the density of the combined life of the part and its spare?

32. Generalize Prob. 26, considering N chips and samples of size n. The variance of the sample mean is

$$\frac{\sigma^2}{n}\frac{N-n}{N-1}$$

where σ^2 is the population variance,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left(i - \frac{N+1}{2}\right)^2$$

CHAPTER 8

POINT ESTIMATION

8.1. Estimation of Parameters. The estimation of parameters is a primary purpose of all scientific experimentation, and before formulating the problem precisely, it may be worth while to consider briefly its practical implications.

Suppose a plant breeder wishes to determine the general yielding ability of a new hybrid line of corn in some agricultural region. To do this, he selects a number of farms in the region and obtains the yields, say in pounds, of small plots planted on each of several farms. He thus obtains a set of observations, say 45, 27, 36, 34, 59, 40, $\cdots$. The average of these numbers gives a measure of the yielding ability. This average is an estimate of the mean μ of some population with a density $f(x)$. Of course the population needs to be carefully specified. Were the farms selected at random? Did the farmers cultivate the plot along with the rest of the crop, or did the plots have special treatment? What were the weather conditions in that season? And so on. But leaving aside these questions and assuming randomness, we regard the experiment as a drawing of a sample from a population with density $f(x)$ for the purpose of estimating the mean of the distribution.

Since the observations were obtained only to the nearest pound, the distribution is, in fact, discrete. However, for measurements (as opposed to countings) it is customary to think of a continuous distribution. The observations could have been obtained more accurately, but any effort in that direction would have been wasted because the sampling error of the estimate would well exceed errors of rounding to the nearest pound. In this connection, however, it is not always possible to reduce errors of measurement well below the magnitude of sampling errors. Thus a metallurgist studying thermal expansion of some alloy might require a very accurate measurement of the length of a rod and make several observations in inches, say 8.562, 8.564, 8.563, 8.563 $\cdots$, with precision equipment which can measure to within about .001 inch. His distribution is discrete (defined at intervals of .001) and cannot be refined; this discreteness may be the major source of the error of his estimate.

In general, the estimation problem may be stated as follows: One is investigating a population with a density function $f(x; \theta_1, \theta_2, \cdots, \theta_k)$, where x is the variate and $\theta_1, \theta_2, \cdots, \theta_k$ are parameters in the distribution. Thus in the case of the gamma distribution there are two parameters which we have called α and β, and in the present notation we might exhibit the parameters by writing the gamma density as $f(x; \alpha, \beta)$. On the basis of a random sample of observations, say $x_1, x_2, \cdots, x_n$, one wishes to estimate one or more of the parameters $\theta_1, \theta_2, \cdots, \theta_k$. The problem here is to find functions of the observations which we may represent by $\hat{\theta}_1(x_1, x_2, \cdots, x_n), \hat{\theta}_2(x_1, x_2, \cdots, x_n), \cdots$, such that the distribution of these functions will be concentrated as closely as possible near the true values of the parameters. We shall call such functions *estimators*. We have already seen, for example, that if the parameter to be estimated is the population mean μ, then the function

$$\hat{\mu}(x_1, x_2, \cdots, x_n) = \bar{x} = \frac{1}{n}\sum x_i \tag{1}$$

is an estimator for μ and that the distribution of $\hat{\mu}$ actually does become closely concentrated near the true mean μ for large samples when the population variance exists.

In speaking of the estimation of parameters, the moments of a distribution are usually intended to be included by the term "parameters" even though they may not enter explicitly in the distribution function. The moments will ordinarily be functions of the parameters which do enter into the functional expression of the distribution, and once those parameters are estimated, corresponding functions of those estimates will estimate the moments. Of course, the moments can also be estimated by means of the sample moments as indicated in the preceding chapter.

Any estimate of a parameter is naturally subject to the errors of sampling, and it is important to make some statement about the possible size of the error when giving an estimate. We shall defer the study of errors, however, until a later chapter and consider here only point estimates, i.e., single-valued estimates, as opposed to more general estimates which merely specify the parameter to be within a given interval.

8.2. Properties of Good Estimators. To consider the case of a single parameter for simplicity, suppose we have a random sample of size n drawn from a population with a distribution $f(x; \theta)$. There are infinitely many ways of choosing an estimating function $\hat{\theta}(x_1, x_2, \cdots, x_n)$, and our problem is to choose a good one. Intuitively it is clear

what is meant by "good"—the distribution of the estimator should be concentrated near the true parameter value θ. Thus if $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ are different estimators of θ with densities $g_1(\hat{\theta}_1)$, $g_2(\hat{\theta}_2)$, $g_3(\hat{\theta}_3)$ as illustrated in Fig. 39, then $\hat{\theta}_2$ is clearly a better estimator than either $\hat{\theta}_1$ or $\hat{\theta}_3$, and $\hat{\theta}_3$ is better than $\hat{\theta}_1$ even though it is biased to the right.

One method of comparing two estimators is by their *relative efficiency*. If an estimator $\hat{\theta}_1(x_1, x_2, \cdots, x_n)$ has $E(\hat{\theta}_1 - \theta)^2 = A_1$, and if a second estimator $\hat{\theta}_2(x_1, x_2, \cdots, x_n)$ has $E(\hat{\theta}_2 - \theta)^2 = A_2$, then the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$ is defined to be A_1/A_2; the ratio is usually expressed as a percentage. If the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$ is greater than 100 per cent, then $\hat{\theta}_2$ may reasonably be regarded as a

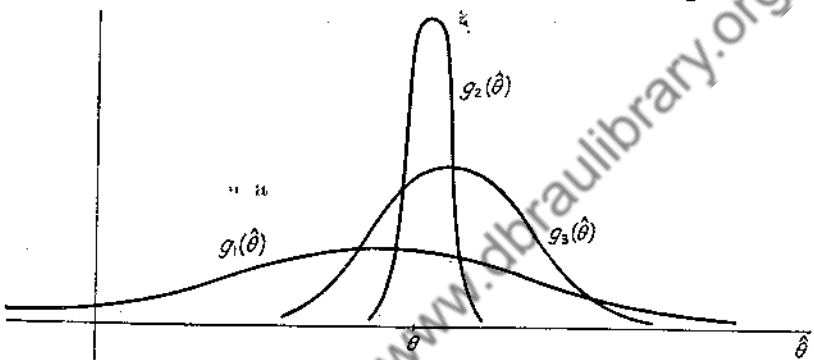


FIG. 39.

better estimator of θ than $\hat{\theta}_1$. It is to be noted that A_1 and A_2 will not be the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$ unless $E(\hat{\theta}_1) = \theta$ and $E(\hat{\theta}_2) = \theta$.

Several terms have come to be commonly used to describe estimators, and we shall define them now.

Unbiased. If an estimator $\hat{\theta}(x_1, x_2, \cdots, x_n)$ for a parameter θ is such that

$$E(\hat{\theta}) = \theta \tag{1}$$

then $\hat{\theta}$ is said to be unbiased. If $E(\hat{\theta}) > \theta$, the estimator is said to be positively biased; while if $E(\hat{\theta}) < \theta$, the estimator is said to be negatively biased. In constructing estimators, it is obviously of some advantage to construct an unbiased estimator, but this is not a very crucial requirement. If the mean of an estimator differs but little from the parameter value relative to the standard deviation of the estimator, the estimator may be quite saitsfactory.

Consistent. If an estimator $\hat{\theta}(x_1, x_2, \cdots, x_n)$ for a parameter θ is such that

$$P(\hat{\theta} \to \theta) \to 1 \qquad \text{as } n \to \infty \tag{2}$$

then $\hat{\theta}$ is said to be a consistent estimate of θ. The symbolic criterion is a way of stating that the estimate becomes near the true parameter value with probability approaching one as the sample size increases without limit. The sample mean $\bar{x}$ is an example of a consistent estimator when the population variance is finite, for $\bar{x}$ has a variance σ^2/n, and as $n \to \infty$, the variance of $\bar{x}$ approaches zero. Since

$$E(\bar{x}) = \mu$$

for any n, it follows that the distribution of $\bar{x}$ must become concentrated at μ when $\sigma^2/n \to 0$.

A consistent estimator is obviously unbiased in the limit, but for finite sample sizes it may be biased though in such a way that the bias approaches zero as n becomes large. An unbiased estimator may or may not be consistent depending on whether or not its distribution becomes concentrated near its mean as the sample size increases. In estimating the mean, for example, we might define an estimator $\hat{\theta} = x_1$, where x_1 is the first observation of the sample; this estimate is unbiased but not consistent.

Efficient. In a great many estimation problems it is possible to construct estimators $\hat{\theta}(x_1, x_2, \cdots, x_n)$, such that $\sqrt{n}(\hat{\theta} - \theta)$ has a normal distribution with zero mean in the limit as the sample size n increases. Confining our attention to this class of estimators (and assuming such a class exists), there may be one or more estimators which will have a limiting variance which is smaller than the limiting variances of the other estimators. These estimators which have the smallest limiting variance are called *efficient* estimators of θ.

It can be shown, for example, that for samples drawn from a normal population with mean μ and variance σ^2, $\hat{\theta}_1 = \bar{x}$ is an efficient estimator of μ. In fact, the limiting distribution of $\sqrt{n}(\bar{x} - \mu)$ is normal with zero mean and variance σ^2. No other estimator can have a smaller limiting variance. However, there are many other efficient estimators, i.e., estimators with the same limiting normal distribution. For example,

$$\hat{\theta}_2 = \frac{1}{n+1} \sum_{i=1}^{n} x_i$$

is efficient since it can be shown that $\sqrt{n}(\hat{\theta}_2 - \mu)$ has a normal distribution with zero mean and variance σ^2 in the limit as n becomes large. It is to be observed that $\hat{\theta}_2$ is biased, since

$$E(\hat{\theta}_2) = \frac{n\mu}{(n+1)}$$

and in general efficient estimators need not be unbiased for finite samples though they are clearly unbiased in the limit. Efficient estimators are necessarily consistent.

Sufficient. An estimator is said to be sufficient if it contains all the information in the sample regarding the parameter. More precisely, if $x_1, x_2, \cdots, x_n$ is a sample from a population with density $f(x; \theta)$ and if $\hat{\theta}(x_1, x_2, \cdots, x_n)$ is an estimator such that the conditional distribution of $x_1, x_2, \cdots, x_n$ given $\hat{\theta}$ does not depend on θ, then $\hat{\theta}$ is a sufficient estimator. This means that the joint density of the sample may be put in the form

$$\prod_{i=1}^{n} f(x_i; \theta) = g(x_1, x_2, \cdots, x_n|\hat{\theta})h(\hat{\theta}; \theta) \tag{3}$$

where the function g does not involve θ. In this form it is clear that no other function of the x_i can provide any information about θ. For consider any other function of the x_i, say $u(x_1, x_2, \cdots, x_n)$. The distribution of u for a fixed $\hat{\theta}$ will be determined by the conditional density $g(x_1, x_2, \cdots, x_n|\hat{\theta})$ and will have $\hat{\theta}$ but not θ as a parameter. Hence u can only provide information about $\hat{\theta}$. But $\hat{\theta}$ is known in any given problem, so that any information provided by u is of no use.

Sufficient estimators are obviously the most desirable kind of estimators to have, but unfortunately they do not exist except in rather special cases. Ordinarily we shall have to be content with less satisfactory estimators.

We have defined all these concepts in terms of one parameter, but the extension to several parameters is straightforward. Thus if x is distributed by $f(x; \theta_1, \theta_2, \cdots, \theta_k)$, a set of estimators $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k$ is unbiased if, for every i,

$$E(\hat{\theta}_i) = \theta_i$$

The set is consistent if, for every i,

$$P(\hat{\theta}_i \to \theta_i) \to 1 \qquad \text{as } n \to \infty$$

where n is the sample size. The set is sufficient if

$$\prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \cdots, \theta_k) = g(x_1, x_2, \cdots, x_n|\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$$
$$h(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k; \theta_1, \theta_2, \cdots, \theta_k)$$

The generalization of the meaning of efficient requires some knowledge of the multivariate normal distribution, a distribution which we shall study in the next chapter. If k variables $u_1, u_2, \cdots, u_k$ have a

multivariate normal distribution, it can be shown that there are linear functions $V_1, V_2, \cdots, V_k$ of the u_i which are independent in the probability sense and each of which has the simple normal distribution, so that the multivariate normal distribution of the u_i may be written as the product of k single-variate normal distributions of the V_i. (This is illustrated in Prob. 25 of Chap. 9.) A set of estimators is efficient if $\sqrt{n}\,(\hat{\theta}_i - \theta)$ have the multivariate normal distribution in the limit as the sample size increases, and if the linear functions V_i of the $\sqrt{n}\,(\hat{\theta}_i - \theta)$ which are independent in the probability sense are such that the product of their variances is a minimum.

8.3. Principle of Maximum Likelihood. To introduce the idea, we shall consider a very simple estimation problem. Suppose an urn contains a number of black and white balls, and suppose it is known that the ratio of the numbers is three to one but that it is not known whether the black or the white balls are the more numerous. That is, the probability of drawing a black ball is either $\frac{1}{4}$ or $\frac{3}{4}$. If n balls are drawn with replacement from the urn, the distribution of the number of black balls is given by the binomial

$$f(x; p) = \binom{n}{x} p^x q^{n-x} \tag{1}$$

where $q = 1 - p$ and p is the probability of drawing a black ball.

We shall draw a sample of three balls with replacement and attempt to estimate the unknown parameter p of the distribution. The estimation problem is particularly simple in this case because we have only to choose between the two numbers .25 and .75. Let us anticipate the result of the drawing of the sample. The possible outcomes and their probabilities under the two possibilities are given below:

x	0	1	2	3
$f\left(x; \frac{3}{4}\right)$	$\frac{1}{64}$	$\frac{9}{64}$	$\frac{27}{64}$	$\frac{27}{64}$
$f\left(x; \frac{1}{4}\right)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

The principle of maximum likelihood essentially assumes that the sample is representative of the population. We shall state it more precisely later. In the present example, if we found $x = 0$ in a sample of three, the estimate .25 for p would be preferred over .75 because

the probability $27/64$ is greater than $1/64$, i.e., because a sample with $x = 0$ is more likely to arise from a population with $p = 1/4$ than from one with $p = 3/4$. And in general we should estimate p by .25 when $x = 0$ or 1, and by .75 when $x = 2$ or 3. The estimator may be defined as

$$\begin{aligned} \hat{p}(x) &= .25 \qquad x = 0, 1 \\ &= .75 \qquad x = 2, 3 \end{aligned} \tag{2}$$

The estimator thus selects for every x the value of p such that

$$f(x; \hat{p}) > f(x; p')$$

where p' is the alternative value of p.

More generally, if several alternative values of p were possible, say $p = 0.1, 0.2, 0.3, \cdots, 1.0$, we might reasonably proceed in the same manner. Thus if we found $x = 6$ in a sample of 25 from a binomial population, we should substitute all possible values of p in the expression

$$f(6; p) = \binom{25}{6} p^6(1 - p)^{19} \tag{3}$$

and choose as our estimate that value of p which maximized $f(6, p)$. For the given possible values of p we should find our estimate to be $\hat{p}(6) = .2$. If there were no restriction on p except that $0 \leq p \leq 1$, then $f(6, p)$ would be regarded as a continuous function of p over the given interval and the position of its maximum value would be found by putting its derivative with respect to p equal to zero and solving the resulting equation for p. Thus,

$$\frac{d}{dp} f(6; p) = \binom{25}{6} p^5(1 - p)^{18}[6(1 - p) - 19p] \tag{4}$$

and on putting this equal to zero and solving for p, we find $p = 0, 1, 6/25$ are the roots. The first two roots are impossible as far as the given sample is concerned, and our estimate is therefore $\hat{p} = 6/25$. This estimate has the property that

$$f(6; \hat{p}) > f(6; p') \tag{5}$$

where p' is any other value of p in the interval $0 < p < 1$.

The principle of maximum-likelihood estimation is simply this: *If $f(x_1, x_2, \cdots, x_n; \theta)$ is the density for a random sample of size n drawn from a population with an unknown parameter θ, then the maxi-*

mum-likelihood estimate of θ is the number $\hat{\theta}$, if it exists, such that

$$f(x_1, x_2, \cdots, x_n; \hat{\theta}) > f(x_1, x_2, \cdots, x_n; \theta')$$

where θ' is any other possible value of θ.

While we have been discussing a discrete distribution in particular, the principle is the same for a continuous distribution. Suppose x is continuous and has the density $f(x; \theta)$. The probability that x will lie in a small interval Δx is approximately $f(x; \theta)\Delta x$. Given a sample of one observation, x_1, we may choose arbitrarily a small interval Δx about x_1 and maximize the probability $f(x_1, \theta)\Delta x$ as a function of θ. However, since Δx is arbitrary, it is not a function of θ and behaves as a constant in so far as variations in θ are concerned. Hence in the maximization we may disregard Δx and deal only with $f(x_1, \theta)$. The conclusion is obviously the same for samples of more than one observation.

The function $\prod_1^n f(x_i; \theta)$, which gives the sample distribution when regarded as a function of the x_i, is regarded as a function of θ for fixed values of the x_i in determining the maximum-likelihood estimate of θ. When regarded as a function of θ, the expression is often referred to as the likelihood function of θ. The maximum-likelihood estimate of θ is therefore the point at which the likelihood function has a maximum.

When more than one parameter is involved, the maximum-likelihood estimates of the parameters are defined similarly. Thus if a sample of size n has the density

$$\prod_{i=1}^n f(x_i; \theta_1, \theta_2, \cdots, \theta_k)$$

then the maximum-likelihood estimates of the parameters are the numbers $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k$, if such a set exists, which maximize the given expression as a function of the θ_i. It often happens in practice that one wishes to estimate some but not all of the unknown parameters of a distribution. Usually it turns out that the maximizing values for the desired set of parameters depend on the remaining parameters, so that it is necessary actually to estimate all the unknown parameters.

8.4. Some Maximum-likelihood Estimators. We shall obtain in this section maximum-likelihood estimators for parameters of some of the common distribution functions. Ordinarily the parameters may be regarded as continuous variables, and the maximizing value may be obtained by putting the derivative of the likelihood function equal to zero and solving for the parameter in the resulting equation.

Since likelihood functions are products, and since sums are usually more convenient to deal with than products, it is customary to maximize the logarithm of the likelihood rather than the likelihood itself, i.e., to maximize

$$L = \log \prod_{i=1}^{n} f(x_i; \theta) = \sum_{i=1}^{n} \log f(x_i; \theta) \tag{1}$$

Of course the logarithm of the likelihood has its maximum at the same point as does the likelihood.

Binomial. Suppose samples of size n are drawn from the binomial distribution

$$f(x; p) = p^x q^{1-x} \qquad x = 0, 1 \tag{2}$$

The sample values, $x_1, x_2, \cdots, x_n$, will be a sequence of zeros and ones, and the likelihood is

$$\prod_{i=1}^{n} p^{x_i} q^{1-x_i} = p^{\Sigma x_i} q^{n-\Sigma x_i} \tag{3}$$

and letting $y = \Sigma x_i$, we have

$$L = y \log p + (n - y) \log q \tag{4}$$

$$\frac{dL}{dp} = \frac{y}{p} - \frac{n-y}{q} \tag{5}$$

remembering that $q = 1 - p$. On putting this last expression equal to zero and solving for p, we find the estimator

$$\hat{p} = \frac{y}{n} = \frac{1}{n} \sum x_i = \bar{x} \tag{6}$$

which is, of course, the obvious estimator for this parameter.

We can show that this estimator is sufficient and therefore that it would be fruitless to search for a better estimator for the parameter. We need to show that the conditional distribution of the x_i given $\bar{x}$ is independent of p. Since the marginal distribution of $n\bar{x} = y$ is given by

$$\binom{n}{y} p^y q^{n-y} \tag{7}$$

the conditional distribution of the x_i given y is obtained by dividing (3) by (7) to get, say,

$$g(x_1, x_2, \cdots, x_n | \hat{p}) = \frac{1}{\binom{n}{n\hat{p}}} \qquad x_i = 0, 1; \ \Sigma x_i = n\hat{p} \tag{8}$$

a distribution which is independent of the parameter p.

Normal. Samples of size n from the normal distribution have the density

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2\sigma^2)(x_i-\mu)^2} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-(1/2\sigma^2)\Sigma(x_i-\mu)^2} \tag{9}$$

The logarithm of the likelihood is

$$L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum (x_i - \mu)^2 \tag{10}$$

To find the location of its maximum, we compute

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum (x_i - \mu) \tag{11}$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum (x_i - \mu)^2 \tag{12}$$

and on putting these derivatives equal to zero and solving the resulting equations for μ and σ^2, we find the estimators

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x} \tag{13}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 \tag{14}$$

which turn out to be the sample moments corresponding to μ and σ^2. The estimator $\hat{\mu}$ is unbiased, but $\hat{\sigma}^2$ is not, since

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \tag{15}$$

We shall see later that this pair of estimators is a sufficient pair for estimating the parameters; the sample distribution for given values of $\hat{\mu}$ and $\hat{\sigma}^2$ does not involve μ and σ^2. We note in this case that it is possible to estimate μ without estimating σ^2, but not possible to estimate σ^2 without first estimating μ.

Uniform. The density for samples of size n from the uniform distribution over the range $\alpha < x < \beta$ is

$$\frac{1}{(\beta - \alpha)^n} \tag{16}$$

so that

$$L = -n \log (\beta - \alpha) \tag{17}$$

If we put the derivatives of this expression with respect to α and β equal to zero and attempt to solve for α and β, we find that at least one

of α, β must be infinite, a nonsensical result. The trouble here is that the likelihood does not have zero slope at its maximum value, so that we must locate its maximum by other means. It is evident from (16) that the likelihood will be made as large as possible when $\beta - \alpha$ is made as small as possible. Given a sample of n observations $x_1, x_2, \cdots, x_n$, suppose we denote the smallest of the observations by x' and the largest by x''. Clearly α can be no larger than x' and β can be no smaller than x''; hence the smallest possible value for $\beta - \alpha$ is $x'' - x'$. The maximum-likelihood estimators are obviously

$$\begin{aligned} \hat{\alpha} &= x' \\ \hat{\beta} &= x'' \end{aligned} \tag{18}$$

a somewhat curious result because no use is made of the intervening observations.

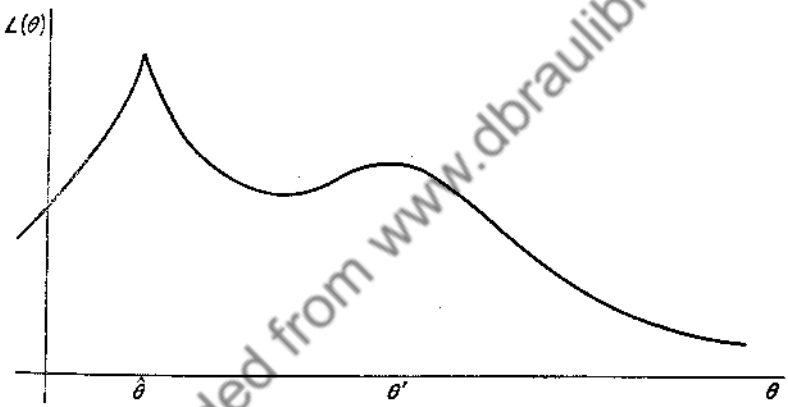


FIG. 40.

These three examples are sufficient to illustrate the application of the method of maximum likelihood. The last example shows that one must not rely on the differentiation process to locate the maximum. The function $L(\theta)$ may, for example, be represented by the curve in Fig. 40, where the actual maximum is at $\hat{\theta}$, but the differentiation process would locate θ' as the maximum. One must also remember that the equation $\partial L/\partial\theta = 0$ locates minima as well as maxima, and hence one must avoid using a root of the equation which actually locates a minimum.

We have not illustrated the estimation of a parameter which appears as a factorial in the distribution function. This may be done in any given problem with the aid of tables of the derivative of the factorial

function. However, such a problem arises so rarely that it is not worth while to study it here. The parameters—n in the binomial distribution, α in the gamma distribution, and α and β in the beta distribution—are usually determined by the sample size and need not be estimated since the sample size is ordinarily known.

8.5. Properties of Maximum-likelihood Estimators. There is no general argument which will show that maximum-likelihood estimators are the best possible estimators. There is, in fact, no way of dealing with the estimation problem (or any other problem requiring inductive inference) completely within the framework of the theory of probability. The theory of probability as a branch of mathematics is a deductive science—given certain axioms, certain conclusions necessarily follow. Uncertain conclusions are outside the realm of the theory. It is precisely here that statistics departs from that theory and becomes an independent discipline. New axioms are required to deal with the problems of statistics; one such axiom might be the principle of maximum likelihood. Whether the new axiom is good or not from the practical viewpoint is, of course, of no interest from the strictly logical viewpoint. When a new axiom is added to a given set of axioms, a new theory involving additional theorems arises, and from the logical viewpoint the only requirement of the new axiom is that it be consistent with the other axioms.

We cannot, therefore, hope to prove that a new axiom or principle is right or wrong. From the practical viewpoint, we naturally want an axiom that will give rise to a useful theory of estimation. In framing such a principle, one would first consider what he wanted the theory to do in practice in terms of certain intuitively desirable criteria (unbiasedness, consistency, for example) and then try to formulate a principle which would lead to such a theory. The principle of maximum likelihood, which is due to R. A. Fisher, forms one basis for a theory of estimation. Other principles would lead to different theories. A choice between principles is, in the last analysis, a matter of opinion as to what is a good theory. After examining the properties of maximum-likelihood estimates, it will become apparent that Fisher's principle leads to a very useful theory, and that for general purposes of estimation there is little if any room for improvement in the theory.

Bias. Maximum-likelihood estimators are not, in general, unbiased, as we have already seen in the case of the variance of a normal population where

$$E(\hat{\sigma}^2) = E\left[\frac{1}{n}\sum (x_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \tag{1}$$

In this case the estimator could be made unbiased by multiplying it by $n/(n-1)$ to obtain the estimator

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \tag{2}$$

which is an unbiased estimator of σ^2. And in general, when maximum-likelihood estimators are biased, it is possible to modify them slightly so that they will be unbiased.

If one requires his estimators to be unbiased, he is using an additional principle which is somewhat in conflict with the principle of maximum likelihood. While there is no particular harm in this (aside from a minor logical inconsistency), there is really nothing to be gained by it. The only claim for unbiasedness as a good criterion is that it forces the distribution of the estimator to be centered (in the center-of-gravity sense) at the true parameter value. But one could just as well require the median, for example, of the distribution to be the true parameter value. Or some other central value might be used. The point is that all one can ask is that the true parameter value be somewhere near the center of the distribution of the estimator. He may choose to define the center however he pleases [mean, median, point such that $E(\hat{\theta} - \theta)^2$ is minimized], but as between reasonable definitions of "center" there is not much choice.

Maximum-likelihood estimators do, in fact, have the true parameter values near the centers of their distributions; we shall not be concerned if the parameter does not happen to be at the exact center of gravity of the distribution.

Invariance. A particularly convenient property of maximum-likelihood estimators is the fact that if $\hat{\theta}$ is the maximum-likelihood estimator for θ, and if $u(\theta)$ is any single-valued function of θ, then $u(\hat{\theta})$ is the maximum-likelihood estimator for $u(\theta)$. This is easily seen to be the case. Let

$$L(\theta) = \sum_1^n \log f(x_i; \theta)$$

Instead of estimating θ we wish to estimate $u(\theta)$. The function $u(\theta)$ defines an inverse function $\theta = v(u)$. The estimator $\hat{u}$ for u is the value of u which maximized $L[v(u)]$. Since the largest value of L occurs at $\theta = \hat{\theta}$, it follows that $v(u)$ must equal $\hat{\theta}$, and hence that $\hat{u} = u(\hat{\theta})$, since u is the inverse function of v.

On the basis of this argument we can conclude directly, for example, that the maximum-likelihood estimator of the standard deviation of a

normal distribution is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

Similarly since the fourth moment about the mean for a normal population is $\mu_4 = 3\sigma^4$, it follows that the maximum-likelihood estimator for μ_4 is

$$\hat{\mu}_4 = 3(\hat{\sigma}^2)^2 = 3\left[\frac{1}{n}\sum (x_i - \bar{x})^2\right]^2$$

not the fourth sample moment, m_4, about the mean as might have been anticipated. Of course, m_4 could be used as an estimator for μ_4, but an examination of the sampling distribution of $\hat{\mu}_4$ and m_4 would show that the former has a distribution which is more closely concentrated about μ_4.

In general, since the moments of a population are ordinarily functions of the parameters that appear in the distribution function, it follows that the maximum-likelihood estimators of the moments are the same functions of the estimators of the parameters. Thus the rth moment of a population with density $f(x; \theta)$ will be some function, say $\mu'_r(\theta)$, of θ. The maximum-likelihood estimator of the parameter will therefore be $\mu'_r(\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator of θ.

Sufficiency. Not all parameters have sufficient estimators, but if a parameter does have sufficient estimators, it can be shown that the maximum-likelihood estimator will be a sufficient estimator. The proof of this statement is of a somewhat advanced mathematical character and will be omitted.

Efficiency. When we examine the large-sample distribution of maximum-likelihood estimators in a later chapter, we shall see that under fairly general conditions the quantity $\sqrt{n}\,(\hat{\theta} - \theta)$ is asymptotically normally distributed with a finite variance; furthermore no other asymptotically normally distributed estimator can have a smaller variance. It follows then that maximum-likelihood estimators are efficient and incidentally are consistent estimators.

All these properties show that the principle of maximum likelihood leads to a very satisfactory theory of estimation. However, perhaps the most important character of the theory from a practical standpoint is of a different kind. It is easy enough to set up in theory a system of estimation by specifying certain properties the estimators should have, but to find the actual functional forms of the estimators may be a very difficult matter. The theory of maximum likelihood does not have

any difficulty of this kind. The estimating functions are determined directly by the maximization process. Thus the theory is eminently satisfactory on two counts: it gives estimators which have desirable properties, and the estimators are easy to find.

8.6. Notes and References. Fisher's paper in which the principle of maximum likelihood was first expounded is cited below. Before the publication of this paper, the customary method for estimating parameters was the *method of moments*. If a distribution function involved r parameters $\theta_1, \theta_2, \cdots, \theta_r$, this technique called for finding the first r population moments as functions of the parameters:

$$\mu_i'(\theta_1, \theta_2, \cdots, \theta_r) = \int_{-\infty}^{\infty} x^i f(x; \theta_1, \theta_2, \cdots, \theta_r) dx$$

then equating the sample moments to these functions, and solving the resulting equations for the parameters. In a few instances this method gives the same estimators as does the maximum likelihood method, but generally the estimators are different.

Fisher was able to demonstrate that his maximum-likelihood estimators were usually far superior to those obtained by the older method. In the second paper cited below he further showed that maximum-likelihood estimators could not be essentially improved. Thus Fisher virtually solved the whole problem of point estimation in these two remarkable papers.

1. R. A. Fisher: "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society*, Series A, Vol. 222 (1922).
2. R. A. Fisher: "Theory of statistical estimation," *Proceedings of the Cambridge Philosophical Society*, Vol. 22 (1925).

8.7. Problems

1. Is the sample mean necessarily an efficient estimator of the population mean for every population?

2. If an estimator is unbiased, can it be expected, for repeated samplings, to underestimate the true parameter half the time and overestimate it half the time?

3. For samples of size 20 find the efficiency of $\bar{x}_1 = \sum_1^{20} x_i$ relative to $\bar{x}_2 = \frac{1}{10} \sum_1^{10} x_i$ as estimators of the population mean.

4. If $\hat{\theta}$ is a sufficient estimator of θ and if $\mu(\theta)$ is a function of θ, is $\mu(\hat{\theta})$ a sufficient estimator of μ?

5. Find the maximum-likelihood estimator for β given a sample of size n from a population with $f(x) = 1/\beta$, $0 < x < \beta$.

6. The sample 1.3, 0.6, 1.7, 2.2, 0.3, 1.1, was drawn from a population with the density $f(x) = 1/\beta$, $0 < x < \beta$. What are the maximum-likelihood estimates of the mean and variance of the population?

7. What is the maximum-likelihood estimator for α in the density $f(x) = (\alpha + 1)x^\alpha$, $0 < x < 1$?

8. Assuming α known, find the maximum-likelihood estimator for β in the gamma distribution.

9. Find the maximum-likelihood estimator for the parameter of the Poisson distribution.

10. Find the maximum-likelihood estimator for the variance of a normal population, assuming the mean is known.

11. Find the maximum-likelihood estimator for the variance of the gamma distribution, assuming α is known.

12. If x is distributed by $f(x) = 1/\beta$, $0 < x < \beta$, and one considers samples consisting of only one observation x, then since $E(x) = \beta/2$, a reasonable estimator for β might be $\hat{\beta}_1 = 2x$. On the other hand, the maximum-likelihood estimator for β is $\hat{\beta}_2 = x$. Is there any choice between these two estimators on grounds of relative efficiency?

13. If x is normally distributed with mean μ and variance σ^2, find, for samples of size k, the maximum-likelihood estimator of the point A such that $\int_A^\infty n(x; \mu, \sigma^2)dx = .05$.

14. It is shown in Chap. 10 that the mean of a sample from a normal population is exactly normally distributed. Use this fact to show that the sample mean is a sufficient estimator of the population mean.

15. In genetic investigations one frequently samples from a binomial $f(x) = \binom{m}{x} p^x q^{m-x}$ except that observations of $x = 0$ are impossible, so that in fact the sampling is from the conditional distribution

$$f(x) = \binom{m}{x} \frac{p^x q^{m-x}}{1 - q^m} \qquad x = 1, 2, \cdots, m$$

Find the maximum-likelihood estimator of p in the case $m = 2$ for samples of size n.

16. Find the estimator for α in the density

$$f(x; \alpha) = \frac{2}{\alpha^2}(\alpha - x) \qquad 0 < x < \alpha$$

for samples of size 2.

17. Referring to Prob. 16, what is the maximum-likelihood estimator of the population mean?

18. An urn contains black and white balls. A sample of size n is drawn with replacement. What is the maximum-likelihood estimator of the ratio R of black to white balls in the urn?

19. Referring to Prob. 18, suppose one draws balls one by one with replacement until a black ball appears. Let x be the number of draws required (not counting the last draw). This operation is repeated n times to obtain a sample $x_1, x_2, \cdots, x_n$. What is the maximum-likelihood estimator of R on the basis of this sample?

20. Suppose n cylindrical shafts made by a machine are selected at random from the production of the machine and their diameters and lengths measured. It is found that n_{11} have both measurements within the tolerance limits, n_{12} have satisfactory lengths but unsatisfactory diameters, n_{21} have satisfactory diameters but unsatisfactory lengths, and n_{22} are unsatisfactory as to both measurements. $\Sigma n_{ij} = m$. Each shaft may be regarded as a drawing from a multinomial population with density

$$p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}} (1 - p_{11} - p_{12} - p_{21})^{x_{22}} \qquad x_{ij} = 0, 1, \Sigma x_{ij} = 1$$

having three parameters. What are the maximum-likelihood estimates of the parameters if $n_{11} = 90$, $n_{12} = 6$, $n_{21} = 3$, $n_{22} = 1$?

21. Referring to the above problem, suppose there is no reason to believe that defective diameters can in any way be related to defective lengths. Then the distribution of the x_{ij} can be set up in terms of two parameters: p_1, the probability of a satisfactory length, and q_1, the probability of a satisfactory diameter. The density of the x_{ij} is then

$$(p_1 q_1)^{x_{11}} [p_1(1 - q_1)]^{x_{12}} [(1 - p_1) q_1]^{x_{21}} [(1 - p_1)(1 - q_1)]^{x_{22}}$$
$$x_{ij} = 0, 1, \Sigma x_{ij} = 1$$

What are the maximum-likelihood estimates for these parameters? Are the probabilities for the four classes different under this model from those obtained in the above problem?

22. A sample of size n_1 is to be drawn from a normal population with mean μ_1 and variance σ_1^2. A second sample of size n_2 is to be drawn from a normal population with mean μ_2 and variance σ_2^2. What is the maximum-likelihood estimator of $\alpha = \mu_1 - \mu_2$? Assuming the total sample size $n = n_1 + n_2$ is fixed, how should the n observations be divided between the two populations in order to minimize the variance of $\hat{\alpha}$.

23. Suppose intelligence quotients for students in a particular age group are normally distributed about a mean of 100 with standard deviation 15. The I.Q., say x_1, of a particular student is to be estimated by a test on which he scores 130. It is further given that test scores are normally distributed about the true I.Q. as a mean with standard deviation 5. What is the maximum-likelihood estimate of the student's I.Q.? (The answer is not 130.)

24. A sample of size n is drawn from each of four normal populations, all of which have the same variance σ^2. The means of the four populations are $a + b + c$, $a + b - c$, $a - b + c$, $a - b - c$. What are the maximum-likelihood estimators of a, b, c, and σ^2? (The sample observations may be denoted by x_{ij}, $i = 1, 2, 3, 4$, and $j = 1, 2, \cdots, n$.)

25. Observations $x_1, x_2, \cdots, x_n$ are drawn from normal populations with the same mean μ but with different variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$. Is it possible to estimate all the parameters? Assuming the σ_i^2 are known, what is the maximum-likelihood estimator of μ?

26. Is $\hat{\sigma}_1$, the square root of the expression on the right of equation (5.2), an unbiased estimate of σ?

CHAPTER 9

THE MULTIVARIATE NORMAL DISTRIBUTION

9.1. The Bivariate Normal Distribution. The bivariate normal distribution is a generalization of the normal distribution for a single variate. The density has the form

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]} \tag{1}$$

and may be represented by a bell-shaped surface $z = f(x, y)$ as in Fig. 41. Any plane parallel to the x, y plane which cuts the surface

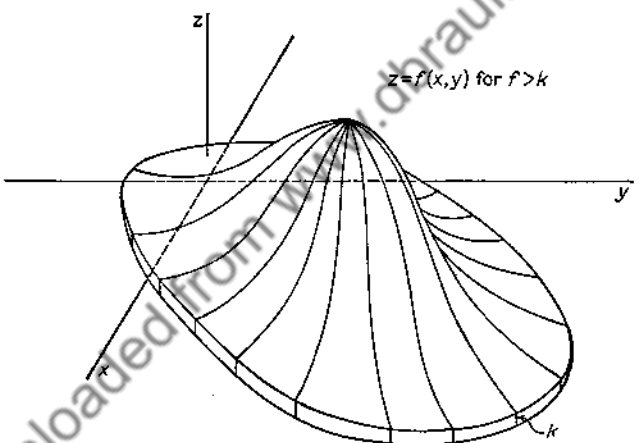


FIG. 41.

will intersect it in an elliptical curve, while any plane perpendicular to the x, y plane will cut the surface in a curve of the normal form. The probability that a point (x, y) drawn at random will lie in any region R of the x, y plane is obtained by integrating the function over that region,

$$P[(x, y) \text{ is in } R] = \iint_R f(x, y)dy\, dx \tag{2}$$

The function might, for example, represent the distribution of hits on a vertical target (Chap. 4) where x and y represent the horizontal

and vertical deviations from the central lines. And in fact the distribution closely approximates the distribution of this as well as many other bivariate populations encountered in practice.

We must first show that the function actually represents a distribution by showing that its integral over the whole plane is one, i.e.,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dy\, dx = 1 \tag{3}$$

The function will, of course, be positive if $-1 < \rho < 1$. To simplify the integral, we shall substitute

$$u = \frac{x - \mu_x}{\sigma_x} \qquad v = \frac{y - \mu_y}{\sigma_y} \tag{4}$$

so that it becomes

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-[1/2(1-\rho^2)](u^2-2\rho uv+v^2)}\, dv\, du$$

On completing the square on u in the exponent, we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-[1/2(1-\rho^2)][(u-\rho v)^2+(1-\rho^2)v^2]}\, dv\, du$$

and on substituting

$$w = \frac{u - \rho v}{\sqrt{1-\rho^2}} \qquad dw = \frac{du}{\sqrt{1-\rho^2}}$$

the integral may be written as the product of two simple integrals,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(w^2/2)}\, dw \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(v^2/2)}\, dv \tag{5}$$

both of which are one, as we have seen in studying the univariate normal distribution. Equation (3) is thus verified.

To obtain the moments of x and y, we shall find their joint moment generating function, say,

$$m(t_1, t_2) = E(e^{t_1x+t_2y}) \tag{6}$$

$$= \iint e^{t_1x+t_2y} f(x, y)dy\, dx \tag{7}$$

Let us again substitute for x and y in terms of u and v to obtain

$m(t_1, t_2) =$

$$e^{t_1\mu_x+t_2\mu_y} \iint e^{t_1\sigma_x u+t_2\sigma_y v} \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-[1/2(1-\rho^2)](u^2-2\rho uv+v^2)}\, dv\, du \tag{8}$$

The combined exponents in the integrand may be written

$$-\frac{1}{2(1-\rho^2)}[u^2 - 2\rho uv + v^2 - 2(1-\rho^2)t_1\sigma_x u - 2(1-\rho^2)t_2\sigma_y v]$$

and on completing the square first on u and then on v, we find this expression becomes

$$-\frac{1}{2(1-\rho^2)}\{[u - \rho v - (1-\rho^2)t_1\sigma_x]^2 + (1-\rho^2)(v - \rho t_1\sigma_x - t_2\sigma_y)^2 - (1-\rho^2)(t_1^2\sigma_x^2 + 2\rho t_1 t_2\sigma_x\sigma_y + t_2^2\sigma_y^2)\}$$

which, on substituting

$$w = \frac{u - \rho v - (1-\rho^2)t_1\sigma_x}{\sqrt{1-\rho^2}}$$

$$z = v - \rho t_1\sigma_x - t_2\sigma_y$$

becomes

$$-\tfrac{1}{2}w^2 - \tfrac{1}{2}z^2 + \tfrac{1}{2}(t_1^2\sigma_x^2 + 2\rho t_1 t_2\sigma_x\sigma_y + t_2^2\sigma_y^2)$$

and the integral in (8) may be written

$$\begin{aligned} m(t_1, t_2) &= e^{t_1\mu_x + t_2\mu_y} e^{\frac{1}{2}(t_1^2\sigma_x^2 + 2\rho t_1 t_2\sigma_x\sigma_y + t_2^2\sigma_y^2)} \int_{-\infty}^{\infty}\int \frac{1}{2\pi} e^{-(w^2/2)-(z^2/2)}\, dw\, dz \\ &= e^{t_1\mu_x + t_2\mu_y + \frac{1}{2}(t_1^2\sigma_x^2 + 2\rho t_1 t_2\sigma_x\sigma_y + t_2^2\sigma_y^2)} \end{aligned} \tag{9}$$

since the integral is obviously one.

The moments may be obtained by evaluating the derivatives of $m(t_1, t_2)$ at $t_1 = 0$, $t_2 = 0$. Thus,

$$E(x) = \left.\frac{\partial m}{\partial t_1}\right]_{t_1, t_2 = 0} = \mu_x \tag{10}$$

$$E(x^2) = \left.\frac{\partial^2 m}{\partial t_1^2}\right]_{t_1, t_2 = 0} = \mu_x^2 + \sigma_x^2 \tag{11}$$

hence the variance of x is

$$E(x - \mu_x)^2 = E(x^2) - \mu_x^2 = \sigma_x^2 \tag{12}$$

Similarly, on differentiating with respect to t_2, one finds the mean and variance of y to be μ_y and σ_y^2. We can also obtain joint moments

$$E(x^r y^s)$$

by differentiating $m(t_1, t_2)$ r times with respect to t_1 and s times with respect to t_2, then putting t_1 and t_2 equal to zero. The *covariance* of x and y is

$$\begin{aligned} E[(x - \mu_x)(y - \mu_y)] &= E(xy - x\mu_y - y\mu_x + \mu_x\mu_y) \\ &= E(xy) - \mu_x\mu_y \\ &= \rho\sigma_x\sigma_y \end{aligned} \tag{13}$$

as may be verified by differentiating $m(t_1, t_2)$ once with respect to each variable, then putting the variables equal to zero. The parameter ρ is called the *correlation* between x and y. When the correlation is zero, it will be observed in (1) that $f(x, y)$ becomes the product of two univariate normal distributions; hence in this case ($\rho = 0$), x and y will be independent in the probability sense.

The marginal density of one of the variables, x, for example, is by definition

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)dy \tag{14}$$

and again substituting

$$v = \frac{y - \mu_y}{\sigma_y}$$

and completing the square on v, one finds

$$f_1(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_x\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{1}{2(1-\rho^2)}\left(v-\rho\frac{x-\mu_x}{\sigma_x}\right)^2} dv$$

Then the substitution

$$w = \frac{v - \rho[(x - \mu_x)/\sigma_x]}{\sqrt{1-\rho^2}} \qquad dw = \frac{dv}{\sqrt{1-\rho^2}}$$

shows at once that

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} \tag{15}$$

the univariate normal density. Similarly the marginal density of y may be found to be

$$f_2(y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2} \tag{16}$$

Having the marginal distributions, it is possible to determine the conditional distributions. Thus the conditional density of x for fixed values of y is

$$f(x|y) = \frac{f(x, y)}{f_2(y)}$$

and after substituting for the functions on the right, the expression may be put in the form

$$f(x|y) = \frac{1}{\sqrt{2\pi}\,\sigma_x\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_x^2(1-\rho^2)}\left[x-\mu_x-\frac{\rho\sigma_x}{\sigma_y}(y-\mu_y)\right]^2} \tag{17}$$

which is a univariate normal density with mean, $\mu_x + (\rho\sigma_x/\sigma_y)(y - \mu_y)$,

and with variance, $\sigma_x^2(1 - \rho^2)$. The conditional distribution of y may be obtained by interchanging x and y throughout (17) to get

$$f(y|x) = \frac{1}{\sqrt{2\pi}\,\sigma_y\sqrt{1-\rho^2}}\,e^{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left[y-\mu_y-\frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)\right]^2} \tag{18}$$

The mean value of a variate in a conditional distribution is called the *regression* function when regarded as a function of the fixed variates in the conditional distribution. Thus the regression function for x in (17) is $\mu_x + (\rho\sigma_x/\sigma_y)(y - \mu_y)$, which is a linear function of y in the present case. For bivariate distributions in general, the mean of x

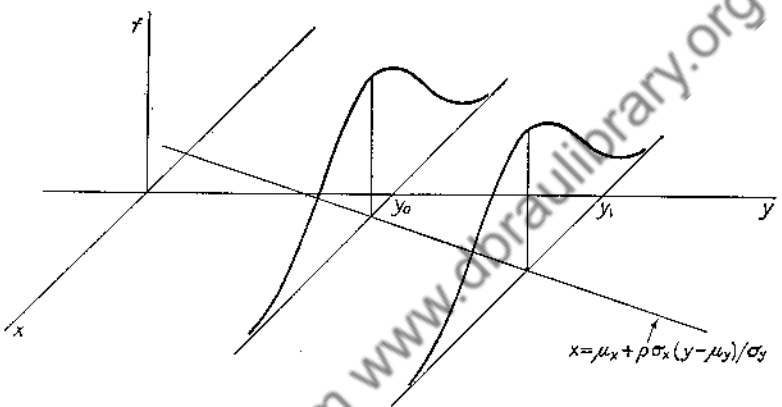


FIG. 42.

in the conditional distribution of x will be some function, say $g(y)$, and the equation

$$x = g(y)$$

when plotted in the x, y plane gives the *regression curve* for x. It is simply a curve which gives the location of the mean of x for various values of y.

For the bivariate normal distribution, the regression curve is the straight line obtained by plotting

$$x = \mu_x + \frac{\rho\sigma_x}{\sigma_y}(y - \mu_y) \tag{19}$$

as shown in Fig. 42. The conditional density of x, $f(x|y)$, is also plotted in the figure for two particular values, y_0 and y_1, of y.

The cumulative bivariate normal distribution

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s, t)\,dt\,ds$$

may be reduced to a form involving only the parameter ρ by the substitution (4). Thus,

$$F(x, y) = F_0(u, v) = \int_{-\infty}^{u}\int_{-\infty}^{v} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-[1/2(1-\rho^2)](s^2-2\rho st+t^2)}\, dt\, ds$$

The function $F_0(u, v)$ is tabulated for $\rho = 0, .05, .10, \cdots, .95$ in Karl Pearson's "Tables for Statisticians and Biometricians" (Part I, Cambridge University Press, London, 1914).

9.2. Matrices and Determinants. It is apparent, from our study of the bivariate normal distribution, that an investigation of the k-variate normal distribution may involve some very unwieldly algebraic expressions. In order to simplify such expressions, it is worth while to develop briefly the algebra of matrices.

A matrix is any rectangular array of quantities. For example,

$$\begin{Vmatrix} 3 & 0 & \log x \\ e^{kx} & a & f(y) \end{Vmatrix}$$

is a matrix with two rows and three columns. The matrix is nothing more than the set of quantities; no operation on the quantities is implied by writing them in such an array. The coordinates (x, y) of a point in a plane may be regarded as a matrix $\|x, y\|$ with one row and two columns. A sample of n observations $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ from a bivariate population may be regarded as a matrix

$$\begin{Vmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_n & y_n \end{Vmatrix}$$

with n rows and two columns, or alternatively as a matrix

$$\begin{Vmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{Vmatrix}$$

with two rows and n columns. The individual quantities which make up the matrix are called *elements* of the matrix.

We shall be concerned with *square* matrices, which have the same number of rows as columns. A general expression for a square matrix is

$$\begin{Vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1k} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2k} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ a_{k1} & a_{k2} & a_{k3} & \cdots & a_{kk} \end{Vmatrix} \tag{1}$$

where the elements are represented by a_{ij}. The subscripts i and j give the position of the element in the array. The first subscript designates the row, and the second one the column. Thus the element represented by a_{57} lies in the fifth row and the seventh column. The top row is generally taken to be the first row, and the left-hand column the first column. The *order* of a square matrix is its number of rows or columns; the matrix in (1) is of order k. The set of elements a_{11}, a_{22}, a_{33}, $\cdots$, a_{kk} are said to form the *main diagonal* of the matrix. A square matrix is symmetric if $a_{ij} = a_{ji}$ for all i and j, i.e., if the array is unchanged when the rows and columns are interchanged. Thus,

$$\begin{Vmatrix} a & 0 & x \\ 0 & b & y \\ x & y & c \end{Vmatrix}$$

is a symmetric square matrix of order three.

An algebra of matrices of the same order may be set up by defining the operations of addition, subtraction, multiplication, and division. The *sum* of two matrices is the matrix of the ordinary sums of corresponding elements. Thus,

$$\begin{Vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{Vmatrix} + \begin{Vmatrix} j & k & l \\ m & n & o \\ p & q & r \end{Vmatrix} = \begin{Vmatrix} a+j & b+k & c+l \\ d+m & e+n & f+o \\ g+p & h+q & i+r \end{Vmatrix} \tag{2}$$

Subtraction is similarly defined. The *product* of two matrices is defined as follows: The element in the ith row and jth column of the product matrix is obtained by multiplying the elements of the ith row of the left-hand matrix by the corresponding elements of the jth column of the right-hand matrix and adding the results. Thus, using a dot to indicate multiplication,

$$\begin{Vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{Vmatrix} \cdot \begin{Vmatrix} j & k & l \\ m & n & o \\ p & q & r \end{Vmatrix}$$

$$= \begin{Vmatrix} aj + bm + cp & ak + bn + cq & al + bo + cr \\ dj + em + fp & dk + en + fq & dl + eo + fr \\ gj + hm + ip & gk + hn + iq & gl + ho + ir \end{Vmatrix} \tag{3}$$

It is to be observed that the product would be different were the order of the two matrices on the left reversed; multiplication is not commutative. Division will be defined later.

We shall use the symbol $\|a_{ij}\|$ to represent a general square matrix of order k; i.e., $\|a_{ij}\|$ represents the array given in (1). In this notation, the definitions of addition, subtraction, and multiplication are

$$\|a_{ij}\| \pm \|b_{ij}\| = \|a_{ij} \pm b_{ij}\| \tag{4}$$

$$\|a_{ij}\| \cdot \|b_{ij}\| = \left\| \sum_{m=1}^{k} a_{im}b_{mj} \right\| \tag{5}$$

The *unit* matrix is defined to be the matrix which has ones for the main diagonal elements and all other elements zero. Thus,

$$\begin{Vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{Vmatrix}$$

is the unit matrix of order three. We shall use the symbol δ_{ij} to represent the elements of the unit matrix; thus δ_{ij} is defined by

$$\begin{aligned} \delta_{ij} &= 1 \qquad i = j \\ &= 0 \qquad i \neq j \end{aligned} \tag{6}$$

It is easily verified that

$$\|\delta_{ij}\| \cdot \|a_{ij}\| = \|a_{ij}\| \cdot \|\delta_{ij}\| = \|a_{ij}\| \tag{7}$$

The unit matrix plays the same role in matrix algebra that unity does in ordinary algebra.

Certain matrices have corresponding *inverse* matrices. The inverse of a matrix $\|a_{ij}\|$ is a matrix, with elements which we shall denote by a^{ij}, such that

$$\|a^{ij}\| \cdot \|a_{ij}\| = \|\delta_{ij}\| \tag{8}$$

Thus the inverse of a matrix corresponds to the quantity $1/c$ associated with a quantity c in ordinary algebra. Division of matrices is defined in terms of the inverse matrix of the denominator. Thus,

$$\frac{1}{\|b_{ij}\|} \cdot \|a_{ij}\| \text{ is defined to be } \|b^{ij}\| \cdot \|a_{ij}\| \tag{9}$$

The inverse of a matrix is often indicated by putting the exponent -1 on the matrix. Thus if a matrix $\|b_{ij}\|$ has an inverse matrix with elements b^{ij}, that fact is usually indicated by writing

$$\|b^{ij}\| = \|b_{ij}\|^{-1}$$

Since multiplication is not commutative in general, it follows that $\|b_{ij}\|^{-1} \cdot \|a_{ij}\|$ will in general be different from $\|a_{ij}\| \cdot \|b_{ij}\|^{-1}$. However, it can be shown that a matrix is commutative with its own inverse:

$$\|a_{ij}\| \cdot \|a^{ij}\| = \|a^{ij}\| \cdot \|a_{ij}\| = \|\delta_{ij}\| \tag{10}$$

Our principal problem in connection with matrices will be to find the inverse of a given matrix. This is most easily done by means of determinants.

The elements of a matrix may be used to form a determinant. We may recall the properties of determinants that are of primary interest here. A determinant is a particular function of a square array of elements, $\|a_{ij}\|$, namely, the polynomial

$$\Sigma \pm a_{1i_1}a_{2i_2} \cdots a_{ki_k} \tag{11}$$

where the sum over $i_1, i_2, \cdots, i_k$ is taken over all permutations of $1, 2, 3, \cdots, k$, and where the sign is plus or minus according as the permutation $(i_1, i_2, \cdots, i_k)$ is an even or odd permutation of $1, 2, 3, \cdots, k$ (i.e., according as the integers in $(i_1, i_2, \cdots, i_k)$ must be interchanged an even or odd number of times to bring them into the order $1, 2, 3, \cdots, k$). The function (11) is usually represented by the array in (1) except that single vertical bars instead of double bars are employed. We shall use the letter A to represent the determinant of the elements a_{ij}.

$$A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix} = \Sigma \pm a_{1i_1}a_{2i_2} \cdots a_{ki_k} \tag{12}$$

The cofactor of any element a_{ij} is the determinant of order $k - 1$ formed by omitting the ith row and jth column of A multiplied by $(-1)^{i+j}$. We shall denote the cofactor of a_{ij} by A_{ij}. Thus,

$$A_{23} = (-1)^5 \begin{vmatrix} a_{11} & a_{12} & a_{14} & a_{15} & \cdots & a_{1k} \\ a_{31} & a_{32} & a_{34} & a_{35} & \cdots & a_{3k} \\ a_{41} & a_{42} & a_{44} & a_{45} & \cdots & a_{4k} \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ a_{k1} & a_{k2} & a_{k4} & a_{k5} & \cdots & a_{kk} \end{vmatrix}$$

It is shown in the elementary theory of determinants that the value of a determinant may be obtained by adding the products of the elements of any row by their cofactors, i.e.,

$$\begin{aligned} A &= a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{ik}A_{ik} \\ &= \sum_{j=1}^{k} a_{ij}A_{ij} \end{aligned} \tag{13}$$

where any value of i may be used. By means of this result, the problem of finding the polynomial expansion (11) of a determinant is reduced to the problem of expanding determinants A_{ij} of one less order. The determinants A_{ij} may be further reduced to expressions involving determinants of order $k - 2$, and so on. Thus, always expanding on the first row, for example, the function represented by a determinant of order three may be found as follows:

$$\begin{aligned} \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} &= a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(e|i| - f|h|) - b(d|i| - f|g|) + c(d|h| - e|g|) \\ &= aei - afh - bdi + bfg + cdh - ceg \end{aligned}$$

since $|x| = x$ by (11).

One other property of determinants which we shall require is

$$\sum_{j=1}^{k} a_{ij}A_{mj} = 0 \qquad i \neq m \tag{14}$$

If the elements of any row are multiplied by the cofactors of the corresponding elements of any other row, the sum of the resulting products will vanish.

We can now determine the inverse of a given matrix in terms of its elements. Suppose the determinant of $\|a_{ij}\|$ is not zero. We shall show that the elements a^{ij} of the inverse of $\|a_{ij}\|$ are

$$a^{ij} = \frac{A_{ji}}{A} \tag{15}$$

where A is the determinant of $\|a_{ij}\|$ and A_{ji} is the cofactor of a_{ji}. To do this, we need only show that

$$\|a_{ij}\| \cdot \|a^{ij}\| = \|\delta_{ij}\|$$

By definition of a product, the element c_{ij}, say, in the product is

$$c_{ij} = \sum_{m} a_{im}a^{mj}$$

$$c_{ij} = \sum_m \frac{a_{im}A_{jm}}{A}$$
$$= \frac{1}{A}\sum_m a_{im}A_{jm}$$

From (13) it follows that the sum is equal to A when $i = j$, and from (14) the sum is zero when $i \neq j$. Hence we have at once that $c_{ij} = \delta_{ij}$.

If the determinant of a matrix vanishes, it is impossible to define its inverse, and division by such matrices is not possible. This situation is not entirely analogous to division by zero in ordinary algebra because there are many matrices with vanishing determinants whereas there is only one quantity zero in ordinary algebra.

Two properties of inverse matrices which we shall require later and which we state without proof are: (1) The determinant of the inverse of a matrix is equal to the reciprocal of the determinant of the original matrix. (2) If a matrix is symmetric, its inverse will also be symmetric.

To illustrate the computation of an inverse matrix, we shall find the inverse of

$$\|a_{ij}\| = \begin{Vmatrix} 3 & 1 & 0 \\ 2 & 4 & 1 \\ 0 & 1 & 3 \end{Vmatrix}$$

The determinant of the matrix is

$$|a_{ij}| = \begin{vmatrix} 3 & 1 & 0 \\ 2 & 4 & 1 \\ 0 & 1 & 3 \end{vmatrix}$$
$$= 3\begin{vmatrix} 4 & 1 \\ 1 & 3 \end{vmatrix} - 1\begin{vmatrix} 2 & 1 \\ 0 & 3 \end{vmatrix} + 0\begin{vmatrix} 2 & 4 \\ 0 & 1 \end{vmatrix}$$
$$= 3(12 - 1) - (6 - 0) = 27$$

The cofactors of the elements are

$$A_{11} = \begin{vmatrix} 4 & 1 \\ 1 & 3 \end{vmatrix} = 11$$
$$A_{12} = -\begin{vmatrix} 2 & 1 \\ 0 & 3 \end{vmatrix} = -6$$
$$A_{13} = \begin{vmatrix} 2 & 4 \\ 0 & 1 \end{vmatrix} = 2$$
$$A_{21} = -\begin{vmatrix} 1 & 0 \\ 1 & 3 \end{vmatrix} = -3$$

and so on; the complete matrix of cofactors is

$$\|A_{ij}\| = \begin{Vmatrix} 11 & -6 & 2 \\ -3 & 9 & -3 \\ 1 & -3 & 10 \end{Vmatrix}$$

On dividing each element of this matrix by $|a_{ij}| = 27$ and interchanging rows and columns, we have the inverse

$$\|a^{ij}\| = \begin{Vmatrix} 11/27 & -3/27 & 1/27 \\ -6/27 & 9/27 & -3/27 \\ 2/27 & -3/27 & 10/27 \end{Vmatrix}$$

as may be verified by multiplying this matrix by the original matrix to obtain the unit matrix of order three.

9.3. The Bivariate Normal Distribution in Matrix Notation. We shall denote the two variates by x_1 and x_2 instead of x and y, and their means by ξ_1 and ξ_2 in place of μ_x and μ_y. (To use μ_1 and μ_2 for the means might result in some confusion with the moments about the mean for a single variate.) The variances of x_1 and x_2 will be denoted by σ_{11} and σ_{22} instead of σ_x^2 and σ_y^2. Instead of the correlation ρ, we shall use the covariance $\rho\sigma_x\sigma_y$ as the fifth parameter and denote it by σ_{12} or σ_{21}. Both σ_{12} and σ_{21} will be used, but it is to be remembered that they are equal and represent the same parameter. The matrix

$$\|\sigma_{ij}\| = \begin{Vmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{Vmatrix} \tag{1}$$

will be referred to as the variance-covariance matrix or, more briefly, as the covariance matrix. It is a symmetric matrix. The determinant of the matrix is

$$|\sigma_{ij}| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} \tag{2}$$

which in the old notation is

$$|\sigma_{ij}| = \sigma_x^2\sigma_y^2(1 - \rho^2) \tag{3}$$

The inverse of the matrix is

$$\|\sigma^{ij}\| = \begin{Vmatrix} \dfrac{\sigma_{22}}{|\sigma_{ij}|} & -\dfrac{\sigma_{12}}{|\sigma_{ij}|} \\ -\dfrac{\sigma_{21}}{|\sigma_{ij}|} & \dfrac{\sigma_{11}}{|\sigma_{ij}|} \end{Vmatrix} \tag{4}$$

which is symmetric since $\sigma_{12} = \sigma_{21}$. In the old notation the elements of the inverse are

$$\|\sigma^{ij}\| = \begin{Vmatrix} \dfrac{1}{\sigma_x^2(1-\rho^2)} & -\dfrac{\rho}{\sigma_x\sigma_y(1-\rho^2)} \\ -\dfrac{\rho}{\sigma_x\sigma_y(1-\rho^2)} & \dfrac{1}{\sigma_y^2(1-\rho^2)} \end{Vmatrix} \tag{5}$$

The determinant of the inverse is

$$|\sigma^{ij}| = \frac{1}{|\sigma_{ij}|}$$
$$= \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \tag{6}$$

Now it is to be observed in (5) that the numbers σ^{ij} are essentially the coefficients of the terms in the exponent in equation (1.1). In fact, the exponent may be written as:

$$-\tfrac{1}{2}[\sigma^{11}(x_1-\xi_1)^2 + \sigma^{12}(x_1-\xi_1)(x_2-\xi_2) + \sigma^{21}(x_1-\xi_1)(x_2-\xi_2) + \sigma^{22}(x_2-\xi_2)^2]$$

and the constant multiplier in the distribution may be written as

$$\frac{\sqrt{|\sigma^{ij}|}}{2\pi} \quad \text{or} \quad \frac{1}{2\pi\sqrt{|\sigma_{ij}|}}$$

The bivariate density may thus be put in the form

$$f(x_1,x_2) = \frac{1}{2\pi}\sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\sigma^{ij}(x_i-\xi_i)(x_j-\xi_j)} \tag{7}$$

The double sum in the exponent is called a *quadratic form* in the variables $x_i - \xi_i$, the σ^{ij} are called the coefficients of the quadratic form, and $\|\sigma^{ij}\|$ is called the matrix of the quadratic form.

9.4. The Multivariate Normal Distribution. The multivariate normal distribution may be thought of as the distribution of a population of objects or events which may be characterized by several variables, say $x_1, x_2, \cdots, x_k$. Thus a population of human beings may be characterized by their heights (x_1), weights (x_2), head lengths (x_3), arm lengths (x_4), waist measurements (x_5), and so on. A machine tool may produce steel shapes which may be specified by several measurements of lengths and angles. Each member of the population has a set of measurements $(x_1, x_2, \cdots, x_k)$; a sample of size n drawn from such a population would consist of n such sets of measurements.

Geometric language is often used to describe a multivariate population. A given set of measurements $(x_1, x_2, \cdots, x_k)$ is referred to as the set of coordinates of a point in a k-dimensional space. The population consists of the points of the space. The distribution could

be plotted in a $(k + 1)$-dimensional space, and would plot as a so-called *hypersurface* consisting of the points $[x_1, x_2, \cdots, x_k, f(x_1, x_2, \cdots, x_k)]$. The statements are the immediate generalizations of the case of one- and two-variate populations. A distribution of a single variate x, say $f(x)$, may be plotted in a two-dimensional space and consists of the points $[x, f(x)]$ which lie on the curve $y = f(x)$. A distribution of two variates x and y may be plotted as a surface in a three-dimensional space; the points of the surface $z = f(x, y)$ have coordinates $[x, y, f(x, y)]$.

The multivariate normal density is

$$f(x_1, x_2, \cdots, x_k) = \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma^{ij}(x_i-\xi_i)(x_j-\xi_j)} \tag{1}$$

in which the matrix $\|\sigma^{ij}\|$ of the quadratic form is symmetric and has a positive determinant. This is the direct generalization of the distribution given at the end of the preceding section. We shall see later that the inverse of the matrix $\|\sigma^{ij}\|$ of the quadratic form is the matrix of variances and covariances, and that the means of the x_i are ξ_i.

In order to show that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_k) \prod_{i=1}^{k} dx_i = 1$$

we shall integrate out one of the variables, say x_1, by completing the square on that variable. First we shall change the variables to

$$y_i = x_i - \xi_i \tag{2}$$

to shorten the ensuing expressions. The quadratic form becomes $\Sigma\Sigma\sigma^{ij}y_iy_j$. Completing the square on y_1, we find

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma^{ij}y_iy_j &= \sigma^{11}y_1^2 + \sum_{i=2}^{k}\sigma^{1i}y_1y_i + \sum_{i=2}^{k}\sigma^{i1}y_iy_1 + \sum_{i=2}^{k}\sum_{j=2}^{k}\sigma^{ij}y_iy_j \\
&= \sigma^{11}\left(y_1^2 + 2y_1\frac{1}{\sigma^{11}}\sum_{i=2}^{k}\sigma^{1i}y_i\right) + \sum_{2}^{k}\sum_{2}^{k}\sigma^{ij}y_iy_j \\
&= \sigma^{11}\left(y_1 + \frac{1}{\sigma^{11}}\sum_{2}^{k}\sigma^{1i}y_i\right)^2 - \frac{1}{\sigma^{11}}\left(\sum_{2}^{k}\sigma^{1i}y_i\right)^2 + \sum_{2}^{k}\sum_{2}^{k}\sigma^{ij}y_iy_j \\
&= \sigma^{11}\left(y_1 + \frac{1}{\sigma^{11}}\sum_{2}^{k}\sigma^{1i}y_i\right)^2 - \frac{1}{\sigma^{11}}\sum_{2}^{k}\sum_{2}^{k}\sigma^{1i}\sigma^{1j}y_iy_j + \sum_{2}^{k}\sum_{2}^{k}\sigma^{ij}y_iy_j \\
&= \sigma^{11}\left(y_1 + \frac{1}{\sigma^{11}}\sum_{2}^{k}\sigma^{1i}y_i\right)^2 + \sum_{2}^{k}\sum_{2}^{k}\left(\sigma^{ij} - \frac{\sigma^{1i}\sigma^{1j}}{\sigma^{11}}\right)y_iy_j
\end{aligned} \tag{3}$$

and on substituting

$$u = y_1 + \frac{1}{\sigma^{11}} \sum_2^k \sigma^{1i} y_i \tag{4}$$

$$\bar{\sigma}^{ij} = \sigma^{ij} - \frac{\sigma^{1i}\sigma^{1j}}{\sigma^{11}} \qquad i, j = 2, 3, \cdots, k \tag{5}$$

we have

$$\sum_1^k \sum_1^k \sigma^{ij} y_i y_j = \sigma^{11} u^2 + \sum_2^k \sum_2^k \bar{\sigma}^{ij} y_i y_j$$

With this reduction we can integrate out y_1. The integral on y_1 is

$$\int_{-\infty}^{\infty} f(y_1, y_2, \cdots, y_k) dy_1 = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\sum_1^k \sum_1^k \sigma^{ij} y_i y_j} dy_1 \tag{6}$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\sigma^{11}u^2 - \frac{1}{2}\sum_2^k \sum_2^k \bar{\sigma}^{ij} y_i y_j} du$$

$$= \left(\frac{1}{2\pi}\right)^{(k-1)/2} \frac{1}{\sqrt{\sigma^{11}}} \sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\sum_2^k \sum_2^k \bar{\sigma}^{ij} y_i y_j}$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \sqrt{\sigma^{11}}\, e^{-\sigma^{11}u^2} du \tag{7}$$

in which the integral is one, as follows from the univariate normal distribution. Now let us examine the resulting function of $y_2, \cdots, y_k$, say,

$$g(y_2, y_3, \cdots, y_k) = \left(\frac{1}{2\pi}\right)^{(k-1)/2} \frac{\sqrt{|\sigma^{ij}|}}{\sqrt{\sigma^{11}}}\, e^{-\frac{1}{2}\sum_2^k \sum_2^k \bar{\sigma}^{i'j'} y_{i'} y_{j'}} \tag{8}$$

where i' and j' are indices which run from 2 to k. Suppose we denote the inverse of $\|\sigma^{ij}\|$ by $\|\sigma_{ij}\|$; then

$$\sum_{m=1}^k \sigma^{im}\sigma_{mj} = \delta_{ij} \tag{9}$$

and since $\|\sigma^{ij}\|$ is symmetric so is $\|\sigma_{ij}\|$, and we may interchange i and m in σ^{im} or j and m in σ_{mj} without invalidating the relation (9).

We shall show that the inverse of $\|\bar{\sigma}^{i'j'}\|$, $i', j' = 2, 3, \cdots, k$, is precisely $\|\sigma_{i'j'}\|$, $i', j' = 2, 3, \cdots, k$, i.e., on omitting the first row and column of the inverse of $\|\sigma^{ij}\|$, we have the inverse of $\|\bar{\sigma}^{i'j'}\|$. We need

only show that

$$\sum_{m=2}^{k} \sigma_{mj'}\bar{\sigma}^{i'm} = \delta_{i'j'} \qquad i', j' = 2, 3, \cdots, k \tag{10}$$

Referring to (5),

$$\sum_{m=2}^{k} \sigma_{mj'}\bar{\sigma}^{i'm} = \sum_{m=2}^{k} \sigma_{mj'}\left(\sigma^{i'm} - \frac{\sigma^{1i'}\sigma^{1m}}{\sigma^{11}}\right)$$

$$= \sum_{m=2}^{k} \sigma_{mj'}\sigma^{i'm} - \frac{\sigma^{1i'}}{\sigma^{11}} \sum_{m=2}^{k} \sigma_{mj'}\sigma^{1m} \tag{11}$$

and in view of (9), the first sum on the right of (11) is $\delta_{i'j'} - \sigma_{1j'}\sigma^{i'1}$, while the second sum is $\delta_{1j'} - \sigma_{1j'}\sigma^{11} = -\sigma_{1j'}\sigma^{11}$ since j' has the range $2, 3, \cdots, k$ so that $\delta_{1j'} = 0$. The expression (11) is therefore

$$\delta_{i'j'} - \sigma_{1j'}\sigma^{i'1} - \frac{\sigma^{1i'}}{\sigma^{11}}(-\sigma_{1j'}\sigma^{11}) = \delta_{i'j'}$$

so that (10) is verified.

The coefficient $\sqrt{|\sigma^{ij}|}/\sqrt{\sigma^{11}}$ is $\sqrt{|\bar{\sigma}^{ij}|}$, as may be seen as follows: σ^{11} is the cofactor of σ_{11} in $|\sigma_{ij}|$ $(i, j = 1, 2, \cdots, k)$ divided by $|\sigma_{ij}|$. The cofactor is $|\sigma_{i'j'}|$ $(i', j' = 2, 3, \cdots, k)$. Since $|\sigma^{ij}| = 1/|\sigma_{ij}|$, we have

$$\frac{\sqrt{|\sigma^{ij}|}}{\sqrt{\sigma^{11}}} = \frac{1}{\sqrt{\sigma^{11}|\sigma_{ij}|}} = \frac{1}{\sqrt{|\sigma_{i'j'}|}}$$

and since $\|\sigma_{i'j'}\|$ is the inverse of $\|\bar{\sigma}^{i'j'}\|$, their determinants are reciprocals and hence

$$\frac{\sqrt{|\sigma^{ij}|}}{\sqrt{\sigma^{11}}} = \sqrt{|\bar{\sigma}^{i'j'}|} \tag{12}$$

We find then that (8) is

$$g(y_2, y_3, \cdots, y_k) = \left(\frac{1}{2\pi}\right)^{(k-1)/2} \sqrt{|\bar{\sigma}^{i'j'}|}\, e^{-\frac{1}{2}\sum\sum\bar{\sigma}^{i'j'}y_{i'}y_{j'}} \tag{13}$$

Now suppose y_2 is integrated out of (13). The preceding argument shows that the result will be, say,

$$h(y_3, y_4, \cdots, y_k) = \left(\frac{1}{2\pi}\right)^{(k-2)/2} \sqrt{|\bar{\bar{\sigma}}^{ij}|}\, e^{-\frac{1}{2}\sum_{i=3}^{k}\sum_{j=3}^{k}\bar{\bar{\sigma}}^{ij}y_iy_j} \tag{14}$$

where $\|\bar{\bar{\sigma}}^{ij}\|$ is the inverse of the matrix obtained by striking out the first row and column of $\|\bar{\sigma}_{i'j'}\|$ or by striking out the first two rows and

columns of $\|\sigma_{ij}\|$. Proceeding in this manner suppose all variables but y_k have been integrated out; the result will be, say,

$$p(y_k) = \frac{1}{\sqrt{2\pi}} \sqrt{\sigma_0}\, e^{-\frac{1}{2}\sigma_0 y_k^2} \tag{15}$$

and we know what σ_0 is in terms of the original parameters σ^{ij}. σ_0 is the inverse of the matrix obtained by striking out the first $k-1$ rows and columns of $\|\sigma_{ij}\|$, but this leaves only one element σ_{kk} in the matrix, and its inverse is simply $1/\sigma_{kk}$. Thus $\sigma_0 = 1/\sigma_{kk}$. The integral of (15) from $-\infty$ to $+\infty$ is, of course, one, and we have shown that (1) does represent a density function.

9.5. Marginal and Conditional Distributions. The argument in the preceding section has supplied us, incidentally, with all the marginal distributions associated with the multivariate normal distribution. The marginal density for the first r variates, $x_1, x_2, \cdots, x_r$, is obtained by integrating out the remaining $k-r$ variates, and the result may be put in the form

$$\left(\frac{1}{2\pi}\right)^{r/2} \sqrt{|\tilde{\sigma}^{ab}|}\, e^{-\frac{1}{2}\sum_{a=1}^{r}\sum_{b=1}^{r}\tilde{\sigma}^{ab}(x_a-\xi_a)(x_b-\xi_b)} \tag{1}$$

where the indices a and b take on the values $1, 2, \cdots, r$. The coefficients $\tilde{\sigma}^{ab}$ of the quadratic form are obtained by striking out the last $k-r$ rows and columns of $\|\sigma_{ij}\|$ and inverting the result; i.e.,

$$\|\tilde{\sigma}^{ab}\| = \|\sigma_{ab}\|^{-1} \qquad a, b = 1, 2, \cdots, r \tag{2}$$

If one wishes to obtain the marginal distribution of any other subset of r variates, he may merely relabel those variates $z_1, z_2, \cdots, z_r$ and use the above form; or he may define indices a', b' which take on the desired values. Thus if one wanted the marginal density of x_1, x_4, x_5, he could put it in the form

$$\left(\frac{1}{2\pi}\right)^{3/2} \sqrt{|\tilde{\sigma}^{a'b'}|}\, e^{-\frac{1}{2}\sum_{a'}\sum_{b'}\tilde{\sigma}^{a'b'}(x_{a'}-\xi_{a'})(x_{b'}-\xi_{b'})} \qquad a', b' = 1, 4, 5$$

where

$$\|\tilde{\sigma}^{a'b'}\| = \begin{Vmatrix} \sigma_{11} & \sigma_{14} & \sigma_{15} \\ \sigma_{41} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{54} & \sigma_{55} \end{Vmatrix}^{-1}$$

Now let us turn to the conditional distributions. The conditional density for the first r variates, for example, is defined by

$$f(x_1, x_2, \cdots, x_r | x_{r+1}, \cdots, x_k) = \frac{f(x_1, x_2, \cdots, x_k)}{g(x_{r+1}, \cdots x_k)} \tag{3}$$

where $g(x_{r+1}, \cdots, x_k)$ is the marginal density of the last $k - r$ variates and is

$$g(x_{r+1}, \cdots, x_k) = \left(\frac{1}{2\pi}\right)^{(k-r)/2} \sqrt{|\tilde{\sigma}^{pq}|}\, e^{-\frac{1}{2}\sum_p \sum_q \tilde{\sigma}^{pq}(x_p - \xi_p)(x_q - \xi_q)} \tag{4}$$

where $p, q = r + 1, r + 2, \cdots, k$, and

$$\|\tilde{\sigma}^{pq}\| = \|\sigma_{pq}\|^{-1} \tag{5}$$

On dividing (4.1) by (4), (3) becomes:

$$\left(\frac{1}{2\pi}\right)^{r/2} \frac{\sqrt{|\sigma^{ij}|}}{\sqrt{|\tilde{\sigma}^{pq}|}}\, e^{-\frac{1}{2}\left(\sum_{ij} \sigma^{ij} y_i y_j - \sum_{pq} \tilde{\sigma}^{pq} y_p y_q\right)} \tag{6}$$

in which we have let $y_i = x_i - \xi_i$. We shall let $i, j = 1, 2, \cdots, k$; $a, b = 1, 2, \cdots, r$; and $p, q = r + 1, r + 2, \cdots, k$ throughout the remainder of this section. The conditional density (6) is a density for the y_a; the y_p are constants. We shall show that (6) is a multivariate normal density for the y_a and that the regression functions (means of the y_a) are linear functions of the y_p.

The quadratic form, $\Sigma\sigma^{ij}y_iy_j$, may be put in the form

$$\sum_{ab} \sigma^{ab} y_a y_b + 2 \sum_{ap} \sigma^{ap} y_a y_p + \sum_{pq} \sigma^{pq} y_p y_q \tag{7}$$

where the first sum involves the squares and products of the variates y_a, the second sum involves only the first powers of the variates, and the third does not involve the variates at all. First we eliminate the linear terms by substituting

$$z_a = y_a + c_a \tag{8}$$

and properly choosing values of the c_a. The substitution changes (7) to

$$\begin{aligned}\sum_{ab} \sigma^{ab}(z_a - c_a)(z_b - c_b) &+ 2 \sum_{ap} \sigma^{ap}(z_a - c_a) y_p + \sum_{pq} \sigma^{pq} y_p y_q \\ = \sum_{ab} \sigma^{ab} z_a z_b &- 2 \sum_{ab} \sigma^{ab} z_a c_b + \sum_{ab} \sigma^{ab} c_a c_b + 2 \sum_{ap} \sigma^{ap} z_a y_p - 2 \sum_{ap} \sigma^{ap} c_a y_p \\ &+ \sum_{pq} \sigma^{pq} y_p y_q \end{aligned} \tag{9}$$

The second and fourth sums on the right of (9) will cancel if we put

$$\sum_b \sigma^{ab} c_b = \sum_p \sigma^{ap} y_p \tag{10}$$

This is a set of r linear equations (for $a = 1, 2, \cdots, r$) which will determine the c's. We may solve them for the c's easily by employing the inverse of $\|\sigma^{ab}\|$, which we may denote by $\|\bar{\sigma}_{ab}\|$. On multiplying (10) by $\bar{\sigma}_{aa'}$ and summing on a, we find

$$\begin{aligned}\sum_{ap} \bar{\sigma}_{aa'} \sigma^{ap} y_p &= \sum_{ab} \bar{\sigma}_{aa'} \sigma^{ab} c_b \\ &= \sum_b \delta_{a'b} c_b \\ &= c_{a'}\end{aligned} \tag{11}$$

If we define

$$\alpha_{ap} = \sum_b \bar{\sigma}_{ab} \sigma^{bp} \tag{12}$$

then the c's are the following linear functions of the y_p:

$$c_a = \sum_p \alpha_{ap} y_p \tag{13}$$

With the substitution of (8) and (11) in (6) the part of the exponent in parentheses becomes then

$$\sum_{ab} \sigma^{ab} z_a z_b + \sum_{ab} \sigma^{ab} c_a c_b - 2 \sum_{ap} \sigma^{ap} c_a y_p + \sum_{pq} \sigma^{pq} y_p y_q - \sum_{pq} \bar{\sigma}^{pq} y_p y_q \tag{14}$$

We shall show that the last four sums cancel out. If we substitute for the c's in (14) from (13), the coefficient of $y_p y_q$ in the last four sums of (14) is, say,

$$d_{pq} = \sum_{ab} \sigma^{ab} \alpha_{ap} \alpha_{bq} - 2 \sum_a \sigma^{ap} \alpha_{aq} + \sigma^{pq} - \bar{\sigma}^{pq} \tag{15}$$

In the first sum a and b are interchanged and $\sum_{a'} \bar{\sigma}_{ba'} \sigma^{a'p}$ substituted for α_{bp} in accordance with (12). The first sum on the right of (15) becomes

$$\begin{aligned}\sum_{aba'} \sigma^{ab} \bar{\sigma}_{ba'} \sigma^{a'p} \alpha_{aq} &= \sum_{aa'} \delta_{a'}^{a} \sigma^{a'p} \alpha_{aq} \\ &= \sum_a \sigma^{ap} \alpha_{aq}\end{aligned} \tag{16}$$

and thus cancels half the second term of (15), leaving

$$d_{pq} = -\sum_a \sigma^{ap} \alpha_{aq} + \sigma^{pq} - \bar{\sigma}^{pq} \tag{17}$$

This expression is now multiplied by $\sigma_{pp'}$ and summed on p after first substituting for α_{aq} from (12); we find

$$\sum_p \sigma_{pp'} d_{pq} = -\sum_{abp} \sigma_{pp'}\sigma^{ap}\bar{\sigma}_{ab}\sigma^{bq} + \sum_p \sigma_{pp'}\sigma^{pq} - \sum_p \sigma_{pp'}\tilde{\sigma}^{pq} \tag{18}$$

$$= -\sum_{ab}\left(\delta^a_{p'} - \sum_{a'} \sigma_{a'p'}\sigma^{aa'}\right)\bar{\sigma}_{ab}\sigma^{bq} + \sum_p \sigma_{pp'}\sigma^{pq} - \delta^q_{p'} \tag{19}$$

$$= \sum_{aa'b} \sigma_{a'p'}\sigma^{aa'}\bar{\sigma}_{ab}\sigma^{bq} + \sum_p \sigma_{pp'}\sigma^{pq} - \delta^q_{p'}$$

$$= \sum_{a'b} \sigma_{a'p'}\delta^{a'}_b\sigma^{bq} + \sum_p \sigma_{pp'}\sigma^{pq} - \delta^q_{p'}$$

$$= \sum_b \sigma_{bp'}\sigma^{bq} + \sum_p \sigma_{pp'}\sigma^{pq} - \delta^q_{p'}$$

$$= \sum_i \sigma_{ip'}\sigma^{iq} - \delta^q_{p'}$$

$$= \delta^q_{p'} - \delta^q_{p'} = 0 \tag{20}$$

The $\delta^a_{p'}$ of (19) vanishes because a and p' have different ranges. Equation (20) is now multiplied by $\tilde{\sigma}^{p'q'}$ and summed on p' to show that the d_{pq} vanish.

We have shown, therefore, that the quadratic form of (6) is simply the first sum of (14):

$$\sum_{ab} \sigma^{ab} z_a z_b = \sum_{ab} \sigma^{ab}(y_a + c_a)(y_b + c_b) \tag{21}$$

and hence that the coefficients of the quadratic form in the conditional density of the y_a are the same as in the original density. Further, the regression functions, $-c_a$, are linear functions of the fixed variates y_p.

9.6. The Moment Generating Function. The joint moment generating function for $x_1, x_2, \cdots, x_k$ is

$$m(t_1, t_2, \cdots, t_k) = E(e^{\Sigma t_i x_i}) \tag{1}$$

$$= \int_{-\infty}^{\infty} \cdots \int e^{\Sigma t_i x_i} \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\Sigma\Sigma\sigma^{ij}(x_i-\xi_i)(x_j-\xi_j)} \prod^k dx_i. \tag{2}$$

Let $x_i - \xi_i = y_i$. To perform the integration, we again need to complete the squares on the y's. We shall merely exhibit the result and show that it is correct. Consider the expression

$$\sum_i \sum_j \sigma^{ij}\left(y_i - \sum_m \sigma_{mi}t_m\right)\left(y_j - \sum_n \sigma_{nj}t_n\right) = \sum_i \sum_j \sigma^{ij}y_i y_j$$
$$- \sum_i \sum_j \sum_n \sigma^{ij}y_i\sigma_{nj}t_n - \sum_i \sum_j \sum_m \sigma^{ij}y_j\sigma_{mi}t_m + \sum_i \sum_j \sum_m \sum_n \sigma^{ij}\sigma_{mi}\sigma_{nj}t_m t_n \tag{3}$$

In the second term we shall sum first on j and use the relation

$$\sum_j \sigma^{ij}\sigma_{nj} = \delta_{in}$$

to obtain

$$\sum_i \sum_n \sum_j \sigma^{ij}\sigma_{nj}y_i t_n = \sum_i \sum_n \delta_{in} y_i t_n$$
$$= \sum_i y_i t_i$$

since the sum on n of $\delta_{in}t_n = t_i$ because $\delta_{in} = 0$ except when $n = i$. Similarly, the third term in (3) reduces to $\Sigma y_j t_j = \Sigma y_i t_i$. In the fourth term of (3) we sum first on i to obtain

$$\sum_j \sum_m \sum_n \delta_{mj}\sigma_{nj}t_m t_n$$

and then sum on j to obtain

$$\sum_m \sum_n \sigma_{nm}t_m t_n$$

We have finally

$$\sum\sum \sigma^{ij}\Big(y_i - \sum_m \sigma_{mi}t_m\Big)\Big(y_j - \sum_n \sigma_{nj}t_n\Big) = \sum_i \sum_j \sigma^{ij}y_i y_j - 2\sum_i t_i y_i + \sum_i \sum_j \sigma_{ij}t_i t_j$$

and (2) may be put in the form

$$m(t_1, \cdots, t_k) = e^{\Sigma t_i \xi_i} e^{\frac{1}{2}\sum_i \sum_j \sigma_{ij}t_i t_j}$$
$$\int_{-\infty}^{\infty} \cdots \int \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{|\sigma^{ij}|}\, e^{-\frac{1}{2}\sum\sum \sigma^{ij}(y_i - \sum_m \sigma_{mi}t_m)(y_j - \sum_n \sigma_{nj}t_n)} \prod dy_i$$

The integral here is clearly one, since it is the integral of a multivariate normal density with parameters $\xi_i' = \Sigma\sigma_{mi}t_m = \Sigma\sigma_{ni}t_n$. Hence the moment generating function is

$$m(t_1, \cdots, t_k) = e^{\Sigma t_i \xi_i + \frac{1}{2}\Sigma\Sigma\sigma_{ij}t_i t_j} \tag{4}$$

On differentiating m with respect to t_r and then putting all $t_i = 0$, we find

$$E(x_r) = \xi_r$$

and the second derivatives show that

$$E(x_r^2) = \sigma_{rr} + \xi_r^2$$
$$E(x_r x_s) = \sigma_{rs} + \xi_r \xi_s$$

remembering that $\sigma_{rs} = \sigma_{sr}$. The variances and covariances of the x_i are therefore σ_{ii} and σ_{ij}; hence the inverse of the matrix $\|\sigma^{ij}\|$ of the quadratic form of the multivariate normal distribution is in fact the matrix of variances and covariances of the distribution.

As in the case of the bivariate distribution we may define *correlations* ρ_{ij} between x_i and x_j by the relations

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \qquad i \neq j$$

and these correlations may be used as parameters instead of the covariances. It can be shown that if $|\sigma^{ij}|$ is positive, as is required by the definition of the distribution, all the correlations must lie between -1 and $+1$. If all the correlations (or covariances) are zero, then the multivariate distribution reduces to the product of k univariate normal distributions with variances $1/\sigma^{ii}$.

9.7. Estimators. If random samples of size n, $(x_{1\alpha}, x_{2\alpha}, \cdots, x_{k\alpha})$, $\alpha = 1, 2, \cdots, n$, are drawn from a k-variate normal population, the joint density of the observations is

$$\left(\frac{1}{2\pi}\right)^{nk/2} |\sigma^{ij}|^{n/2} e^{-\frac{1}{2}\sum_i \sum_j \sum_\alpha \sigma^{ij}(x_{i\alpha} - \xi_i)(x_{j\alpha} - \xi_j)} \tag{1}$$

and the logarithm of the likelihood is

$$L = -\frac{nk}{2} \log 2\pi + \frac{n}{2} \log |\sigma^{ij}| - \frac{1}{2} \sum_i \sum_j \sum_\alpha \sigma^{ij}(x_{i\alpha} - \xi_i)(x_{j\alpha} - \xi_j) \tag{2}$$

To estimate the parameters ξ_i and σ^{ij}, we solve the equations obtained by putting the derivatives of L with respect to these parameters equal to zero. Considering first the means,

$$\begin{aligned}\frac{\partial L}{\partial \xi_1} &= \sigma^{11} \sum_\alpha (x_{1\alpha} - \xi_1) + \frac{1}{2} \sum_{j=2}^{k} \sum_\alpha \sigma^{1j}(x_{j\alpha} - \xi_j) + \frac{1}{2} \sum_{i=2}^{k} \sum_\alpha \sigma^{i1}(x_{i\alpha} - \xi_i) \\ &= \sum_{i=1}^{k} \sum_\alpha \sigma^{i1}(x_{i\alpha} - \xi_i)\end{aligned} \tag{3}$$

since $\sigma^{1i} = \sigma^{i1}$.

And in general for ξ_r we have

$$\frac{\partial L}{\partial \xi_r} = \sum_i \sum_\alpha \sigma^{ir}(x_{i\alpha} - \xi_i) \qquad r = 1, 2, \cdots, k \tag{4}$$

If we substitute $\bar{x}_i = (1/n) \sum_\alpha x_{i\alpha}$ in the last expression and equate it to zero, we have a set of k equations:

$$n \sum_{i=1}^{k} \sigma^{ir}(\bar{x}_i - \xi_i) = 0 \qquad r = 1, 2, \cdots, k \tag{5}$$

to be solved for the ξ_i. On dividing by n, then multiplying by σ_{rs}, and summing on r, we have

$$\sum_r \sum_i \sigma_{rs}\sigma^{ir}(\bar{x}_i - \xi_i) = 0$$

or

$$\sum_i \delta_{is}(\bar{x}_i - \xi_i) = 0$$

or

$$\bar{x}_s - \xi_s = 0 \qquad s = 1, 2, \cdots, k$$

The estimators $\hat{\xi}_i$ of the population means ξ_i are therefore the sample means.

$$\hat{\xi}_i = \bar{x}_i = \frac{1}{n} \sum_\alpha x_{i\alpha} \tag{6}$$

To estimate the σ^{ij}, we must differentiate L with respect to each of these parameters. We have $\sigma^{ij} = \sigma^{ji}$; however it will be simpler to regard σ^{ij} as different from σ^{ji}. We seek the maximum of L subject to the restrictions on the variables, $\sigma^{ij} = \sigma^{ji}$, but we shall find first the maximum of L without observing these restrictions. Certainly the unrestricted maximum will be at least as large as the restricted maximum. We have

$$\begin{aligned}\frac{\partial L}{\partial \sigma^{rs}} &= \frac{n}{2}\frac{1}{|\sigma^{ij}|} \text{ cofactor of } \sigma^{rs} - \frac{1}{2}\sum_\alpha (x_{r\alpha} - \xi_r)(x_{s\alpha} - \xi_s) \\ &= \frac{n}{2}\sigma_{rs} - \frac{1}{2}\sum_\alpha (x_{r\alpha} - \xi_r)(x_{s\alpha} - \xi_s) \qquad r, s = 1, 2, \cdots, k \end{aligned} \tag{7}$$

On putting this expression equal to zero for all pairs (r, s), we have a set of k^2 equations to solve for the σ_{ij}. The solutions will obviously involve $\hat{\xi}_i$, and we have already solved for those in equation (6). Let us now define

$$a_{ij} = \frac{1}{n}\sum_\alpha (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \tag{8}$$

Then (7), after substituting $\bar{x}_i$ for $\hat{\xi}_i$, becomes

$$\frac{n}{2}\sigma_{rs} - \frac{n}{2}a_{rs}$$

On equating this to zero we have

$$\hat{\sigma}_{rs} = a_{rs} \qquad r, s, = 1, 2, \cdots, k \tag{9}$$

and if we let $\|a^{ij}\|$ be the inverse of $\|a_{ij}\|$, we have

$$\hat{\sigma}^{ij} = a^{ij} \tag{10}$$

We have located the unrestricted maximum, but it turns out to be equivalent to the restricted maximum because it is obvious from (8) that $a_{ij} = a_{ji}$; hence $\hat{\sigma}^{ij} = \hat{\sigma}^{ji}$. Thus the same maximum would have been located had we used the restrictions $\sigma^{ij} = \sigma^{ji}$ originally; the only point of omitting the restrictions is that it simplifies the differentiation of the determinant in (2).

The maximum likelihood estimators of the means, variances, and covariances are therefore

$$\hat{\xi}_i = \frac{1}{n}\sum_{\alpha} x_{i\alpha}$$

$$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{\alpha}(x_{i\alpha} - \hat{\xi}_i)(x_{j\alpha} - \hat{\xi}_i) \tag{11}$$

and the estimators of the parameters σ^{ij} are given by the inverse of $\|\hat{\sigma}_{ij}\|$,

$$\|\hat{\sigma}^{ij}\| = \|\hat{\sigma}_{ij}\|^{-1} \tag{12}$$

9.8. Problems

1. Show that the contour lines for the bivariate normal density [i.e., curves for which $f(x, y) = \text{constant}$] are ellipses.

2. Show that any plane perpendicular to the x, y plane intersects the normal surface in a curve of the normal form.

3. If the exponent of the exponential in a bivariate normal density is $-\frac{1}{6}[4(x+1)^2 - 2(x+1)(y-2) + (y-2)^2]$, what are the means, variances, and covariance of the variates?

4. What is the moment generating function for the distribution specified in Prob. 3?

5. What is the moment generating function for moments about the means for the bivariate normal distribution?

6. Find the inverse of the matrix $\begin{Vmatrix} 3 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{Vmatrix}$.

7. Find the variances and covariances of normal variates which have the quadratic form $2x_1^2 + x_2^2 + 4x_3^2 - x_1x_2 - 2x_1x_3$ in their distribution.

8. What is the marginal density of x_1 and x_3 in Prob. 7?

9. What is the conditional density of x_1 and x_3 in Prob. 7?

10. If the matrix of Prob. 6 is the matrix $\|\sigma^{ij}\|$ of a normal distribution of x_1, x_2, x_3, x_4, show that the conditional distribution of x_1 and x_2 is the same as the marginal distribution of x_1 and x_2, hence that the pair (x_1, x_2) is distributed independently of the pair (x_3, x_4).

11. Show that the determinant with k rows and columns,

$$\begin{vmatrix} a & b & b & \cdots & b \\ b & a & b & \cdots & b \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ b & b & b & \cdots & a \end{vmatrix}$$

which has a's in the main diagonal and b's everywhere else, has the value

$$(a - b)^{k-1}[a + (k - 1)b]$$

Before expanding the determinant, subtract the second row from the first, the third from the second, and so on; then add the first column to the second, the second to the third, and so on.

12. Given the sample (2.5, 7.0), (4.0, 9.0), (0.4, 1.7), (1.2, 2.0), (0.3, 0.0), (1.5, 3.7) from a normal bivariate population, find the maximum-likelihood estimate of the regression function for the conditional distribution of x_2. Plot the sample observations and the regression function.

13. Consider any multivariate density $f(x_1, x_2, \cdots, x_k)$. One can define

$$\begin{aligned} &\text{The means:} && \xi_i = E(x_i) \\ &\text{The variances:} && \sigma_{ii} = E[(x_i - \xi_i)^2] \\ &\text{The covariances:} && \sigma_{ij} = E[(x_i - \xi_i)(x_j - \xi_j)] \\ &\text{The correlations:} && \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \end{aligned}$$

What is the mean and variance of any linear function $y = \Sigma a_i x_i$ of the x's?

14. Referring to Prob. 13, what is the correlation between two linear functions $y = \Sigma a_i x_i$ and $z = \Sigma b_i x_i$ $(y \neq kz)$?

15. What is the covariance matrix for the multinomial distribution [equation (3.5.2)]?

16. Referring to Prob. 13, the conditional density of the first r x's is

$$f(x_1, x_2, \cdots, x_r | x_{r+1}, \cdots, x_k) = \frac{f(x_1, x_2, \cdots, x_k)}{g(x_{r+1}, \cdots, x_k)}$$

where g represents the marginal density of the remaining variates. The conditional distribution has means variances and covariances which may be functions of the $x_{r+1}, \cdots, x_k$ and may be denoted by $\xi_i(x_{r+1}, \cdots, x_k)$ (the regression functions) and $\sigma_{ij}(x_{r+1}, \cdots, x_k)$ where now $i, j = 1, 2, \cdots, r$. Show that the expected value of the regression function $\xi_i(x_{r+1}, \cdots, x_k)$ is the mean of x_i under the unconditional distribution.

17. Show that the $\sigma_{ij}(x_{r+1}, \cdots, x_k)$ of Prob. 16 are constants for the multivariate normal distribution.

18. Verify the details of the sequence of equations (5.18 to 5.20).

19. The expected values of the $\sigma_{ij}(x_{r+1}, \cdots, x_k)$ defined in Prob. 16 are called *variances and covariances about the regression functions* and are usually denoted by

$$\sigma_{ij\cdot(r+1)\cdots k} = E[\sigma_{ij}(x_{r+1}, \cdots, x_k)]$$

The *partial correlation coefficients* of the conditional distribution are defined by

$$\rho_{ij\cdot(r+1)\cdots k} = \frac{\sigma_{ij\cdot(r+1)\cdots k}}{\sqrt{\sigma_{ii\cdot(r+1)\cdots k}\sigma_{jj\cdot(r+1)\cdots k}}}$$

Find $\rho_{12\cdot 3}$ in terms of p_1, p_2, p_3, and p_4 for the multinomial distribution, taking the number of classes to be four.

20. What is $\sigma_{11\cdot 2}$ for the bivariate normal distribution?

21. Find the conditional density of x_1 and x_2, given x_3, for the trivariate normal distribution, and show that the regression functions are linear. (Simplify the algebra by using variates $y_i = x_i - \xi_i$. The means of y_1 and y_2 are $(\sigma_{13}/\sigma_{33})y_3$ and $(\sigma_{23}/\sigma_{33})y_3$.

22. Find the variances and covariances about the regression functions for the conditional distribution of Prob. 21.

23. Show, for the trivariate normal distribution, that

$$\rho_{12\cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1 - \rho_{13}^2)(1 - \rho_{23}^2)}}$$

24. Let $x_1, x_2, \cdots, x_{2k}$ denote scores on $2k$ questions in an aptitude test. Let the scores be normally distributed, each with the same mean

and variance (μ and σ^2), and such that the correlation between any pair of questions is $\rho > 0$. If $y_1 = \sum_1^k x_{2i-1}$ and $y_2 = \sum_1^k x_{2i}$ are total scores on the odd and even questions, find the correlation between y_1 and y_2 and show that it can be made as near unity as one pleases by making the test sufficiently long.

25. Let $x_{1\cdot 23\cdots r}$ represent the deviation of x_1 from its regression function in the conditional distribution of x_1, given $x_2, x_3, \cdots, x_r$. Show for a trivariate normal distribution that x_1, $x_{2\cdot 1}$, $x_{3\cdot 21}$ are independently normally distributed.

26. Generalize the result of Prob. 25 to k variates.

27. Let $x_1, x_2, \cdots, x_k$ have the multivariate normal distribution and consider the conditional distribution of x_1, given the other $k - 1$ variates. Let the regression function be denoted by z; the correlation between x_1 and z is called the *multiple correlation coefficient* of x_1 on z and is denoted by $R_{1\cdot 23\cdots k}$. Show for a trivariate normal distribution that

$$\sigma_{11\cdot 23} = \sigma_{11}(1 - R^2_{1\cdot 23})$$

28. Referring to Prob. 27, show that

$$\sigma_{11\cdot 23\cdots k} = \sigma_{11}(1 - R^2_{1\cdot 23\cdots k})$$

29. Show that

$$1 - R^2_{1\cdot 23} = (1 - \rho^2_{12})(1 - \rho^2_{13\cdot 2})$$

30. Show that

$$1 - R^2_{1\cdot 23\cdots k} = (1 - \rho^2_{12})(1 - \rho^2_{13\cdot 2})(1 - \rho^2_{14\cdot 23}) \cdots (1 - \rho^2_{1k\cdot 23\cdots (k-1)})$$

CHAPTER 10

SAMPLING DISTRIBUTIONS

10.1. Distributions of Functions of Random Variables. In order to study further the problem of estimation, it is necessary to have the distributions of the estimators. In this section we shall consider methods of obtaining such distributions, and then in the remaining sections of the chapter the methods will be employed to obtain certain distributions of particular interest.

A variate x may be transformed by some function of x, say $u(x)$, to define a new variate u. We may think of the population over which x varies to be changed to a new population over which u varies. A sample value x_0, for example, drawn from the x population may be interpreted as determining an observation $u_0 = u(x_0)$ from the u population. The density of u, say $g(u)$, will be determined by the transformation $u(x)$ together with the density $f(x)$ of x.

If x is a discrete variate, the distribution of a function $u(x)$ is determined directly by the laws of probability. If x takes on the values $0, 1, 2, \cdots, r$, for example, with probabilities $f(0), f(1), \cdots, f(r)$, then the possible values of u, say $u_0, u_1, \cdots, u_s$, are determined by substituting the successive values of x in $u(x)$, which we shall assume to be a single-valued function of x. It may be that several values of x give rise to the same value of u. The probability that u takes on a given value, say u_i, is

$$g(u_i) = \Sigma' f(x) \tag{1}$$

where the sum, Σ', is taken over all values of x such that $u(x) = u_i$. Thus suppose x takes on the values 0, 1, 2, 3, 4, 5 with probabilities $p_0, p_1, p_2, p_3, p_4, p_5$; the density of $u = (x - 2)^2$ is

$$\begin{aligned} g(0) &= p_2 \\ g(1) &= p_1 + p_3 \\ g(4) &= p_0 + p_4 \\ g(9) &= p_5 \end{aligned}$$

and 0, 1, 4, 9 are all the possible values of u. Similarly if u is a function of several discrete variates $x_1, x_2, \cdots, x_k$ with a joint density

$f(x_1, x_2, \cdots, x_k)$, the probability that $u(x_1, x_2, \cdots, x_k)$ takes on a particular one of its values u_i is

$$g(u_i) = \Sigma' f(x_1, x_2, \cdots, x_k) \tag{2}$$

where Σ' is taken over all sets of values of the x's such that $u(x_1, x_2, \cdots, x_k) = u_i$.

The basic and often the simplest method for finding distributions of functions of continuous random variables was given in Prob. 28 of Chap. 4. If x has density $f(x)$ and $u(x)$ is a function of x, then the

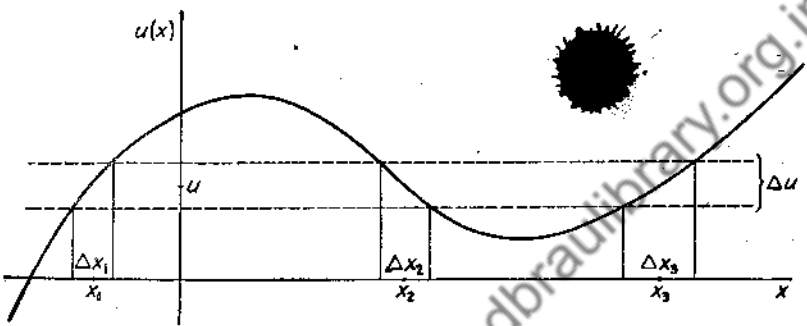


FIG. 43.

cumulative distribution of u is readily found. Let $G(u)$ denote the cumulative distribution; then

$$G(u) = P[u(x) < u] \tag{3}$$

$$= \int_{u(x)<u} f(x)dx \tag{4}$$

in which the integral is taken over that part of the x axis where the function $u(x)$ is less than u. If, for example,

$$u(x) = x^3 - 2 \tag{5}$$

then

$$G(u) = \int_{-\infty}^{\sqrt[3]{u+2}} f(x)dx = F(\sqrt[3]{u+2}) \tag{6}$$

Of course the density function may be obtained by differentiating the cumulative distribution.

It will be instructive to consider another approach to this problem of finding the distributions of functions of continuous variates.

We shall first investigate functions of a single random variate x. To see how $f(x)$ and $u(x)$ determine $g(u)$, we may consider the situation illustrated in Fig. 43, where a particular function $u(x)$ is plotted. We

wish to determine $g(u)$ at the point marked u on the u axis between the horizontal dotted lines. If we solve the equation $u = u(x)$ for x, we may obtain one or more values of x; thus in the figure there are three values, x_1, x_2, x_3, which correspond to the given value of u. A small interval Δu about u determines corresponding intervals Δx_1, Δx_2, and Δx_3 about the points x_i which correspond to u. The function $g(u)$ must be such a function that

$$P(u \text{ lies in } \Delta u) = \int_{\Delta u} g(u)du \tag{7}$$

where the symbolism on the right means that the integral is to be taken over the interval Δu. We have already seen (Sec. 4.2) that a value u' may be found in the interval such that

$$\int_{\Delta u} g(u)du = g(u')\Delta u \tag{8}$$

Now u will lie in the interval Δu provided x lies in any one of the intervals Δx_1, Δx_2, Δx_3; hence we may state

$$P(u \text{ in } \Delta u) = P(x \text{ in } \Delta x_1) + P(x \text{ in } \Delta x_2) + P(x \text{ in } \Delta x_3) \tag{9}$$

and since

$$P(x \text{ in } \Delta x_i) = \int_{\Delta x_i} f(x)dx = f(x_i')\Delta x_i \tag{10}$$

for a properly chosen value x_i' in Δx_i, we have

$$g(u')\Delta u = f(x_1')\Delta x_1 + f(x_2')\Delta x_2 + f(x_3')\Delta x_3 \tag{11}$$

From this relation it is clear that $g(u)$ may be determined by dividing through by Δu and taking the limit as $\Delta u \to 0$.

The curve $u = u(x)$ may also be represented over Δx_1 by the equation $x = x_1(u)$ obtained by solving $u = u(x)$ for x. Similarly over Δx_2 the curve may be represented by $x_2(u)$, and over Δx_3 by $x_3(u)$. From (11) we have

$$\lim_{\Delta u \to 0} g(u') = \lim_{\Delta u \to 0} \left[f(x_1') \frac{\Delta x_1}{\Delta u} + f(x_2') \frac{\Delta x_2}{\Delta u} + f(x_3') \frac{\Delta x_3}{\Delta u} \right] \tag{12}$$

and when $\Delta u \to 0$ in such a way as to collapse on u, all the Δx_i also approach zero so that they collapse on the corresponding x_i. The values u' and x_i' necessarily approach u and x_i since the primed values must lie within the corresponding intervals. The ratios $\Delta x_i/\Delta u$, of course, approach the derivatives of the x_i when Δu approaches zero. It follows then that

$$g(u) = f(x_1) \frac{dx_1}{du} + f(x_2) \frac{dx_2}{du} + f(x_3) \frac{dx_3}{du}$$

except that one revision is required. Some of the derivatives may be negative; thus at x_2 in the figure u decreases with increasing x hence dx_2/du is negative. We are, however, interested in the positive areas in (9), and for this reason we must change the signs of any negative derivatives. We shall use a subscript + to indicate that a quantity is to have its sign changed if it is negative. We shall write, therefore,

$$g(u) = f(x_1)\frac{dx_1}{du_+} + f(x_2)\frac{dx_2}{du_+} + f(x_3)\frac{dx_3}{du_+} \tag{13}$$

and since we shall want $g(u)$ to be a function of u instead of the x_i, we shall substitute the functions $x_i(u)$ for the x_i in this relation.

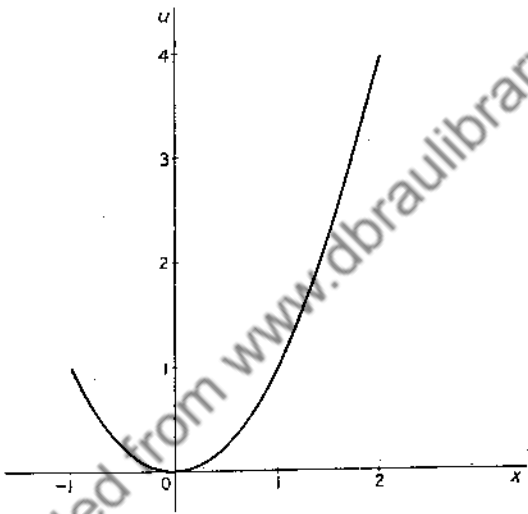


FIG. 44.

To illustrate the above ideas, we may consider the variate x with density

$$f(x) = \tfrac{2}{9}(x + 1) \qquad -1 < x < 2 \tag{14}$$

and transform x to u by the relation $u = x^2$. The function u is plotted in Fig. 44. The range of u is clearly $0 < u < 4$. If $u < 1$, there are two values of x which correspond to each value of u; we may designate them by

$$\begin{aligned} x_1(u) &= -\sqrt{u} \qquad x < 0 \\ x_2(u) &= \sqrt{u} \qquad x > 0 \end{aligned} \tag{15}$$

For $u > 1$, there is only one corresponding value of x, namely,

$$x(u) = \sqrt{u}$$

We must therefore define the distribution of u in two parts. If $0 < u < 1$, we have by (13):

$$\begin{aligned} g(u) &= \frac{2}{9}[x_1(u) + 1]\frac{dx_1}{du_+} + \frac{2}{9}[x_2(u) + 1]\frac{dx_2}{du_+} \\ &= \frac{2}{9}(-\sqrt{u} + 1)\frac{1}{2\sqrt{u}} + \frac{2}{9}(\sqrt{u} + 1)\frac{1}{2\sqrt{u}} \\ &= \frac{2}{9\sqrt{u}} \end{aligned} \tag{16}$$

while if $1 < u < 4$,

$$\begin{aligned} g(u) &= \frac{2}{9}[x(u) + 1]\frac{dx}{du_+} \\ &= \frac{2}{9}(\sqrt{u} + 1)\frac{1}{2\sqrt{u}} \end{aligned} \tag{17}$$

The general procedure is now clear. To find the distribution of any function $u(x)$ of a random variable x, we find, for every u, all the points x_i such that $u(x_i) = u$, and express the x_i as functions of u, say $x_i(u)$. The density of u is

$$g(u) = \sum_i f(x_i(u))\frac{dx_i}{du_+} \tag{18}$$

where $f(x)$ is the density of x. Often we shall deal with monotone functions $u(x)$, functions which are single-valued and such that $x(u)$ is also single-valued. In this case the sum in (18) would consist of only one term, and we have

$$g(u) = f(x(u))\frac{dx}{du_+} \tag{19}$$

for monotone functions $u(x)$.

When u is a function of several random variables, the distribution of u may be obtained as a marginal distribution. Suppose $x_1, x_2, \cdots, x_k$ have a density $f(x_1, x_2, \cdots, x_k)$ and the density of $u(x_1, x_2, \cdots, x_k)$ is required. We may eliminate one of the x's, say x_1, in terms of u by solving the equation

$$u(x_1, x_2, \cdots, x_k) = u$$

for x_1 to obtain a function $x_1(u, x_2, x_3, \cdots, x_k)$, or several such functions $x_{1i}(u, x_2, \cdots, x_k)$ if u is not a monotone function of x_1. Using a similar argument to that used to obtain (18), we may obtain a density

$$g(u, x_2, \cdots, x_k) = \sum_i f(x_{1i}(u, x_2, \cdots, x_k), x_2, \cdots, x_k) \frac{\partial x_{1i}}{\partial u_+} \quad (20)$$

and the density of u may then be found by integrating out $x_2, x_3, \cdots, x_k$ in g. We shall illustrate this method in the next section where we shall find the distribution of

$$u(x_1, x_2, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

for samples from a normal population.

The procedure described above may be generalized to determine the joint distribution of several functions $u_1(x_1, \cdots, x_k), u_2(x_1, \cdots, x_k), \cdots, u_r(x_1, \cdots, x_k)(r \leq k)$ of k random variables. We may put

$$\begin{aligned} u_1(x_1, \cdots, x_k) &= u_1 \\ u_2(x_1, \cdots, x_k) &= u_2 \\ \cdots\cdots\cdots\cdots & \\ u_r(x_1, \cdots, x_k) &= u_r \end{aligned} \quad (21)$$

and solve the resulting set of equations for $x_1, x_2, \cdots, x_r$ to obtain a set of r functions $x_j(u_1, u_2, \cdots, u_r, x_{r+1}, \cdots, x_k)$, or if the solution is not unique, we may have several such sets of r functions. The joint density of the u's and the remaining x's can be shown to be

$$g(u_1, u_2, \cdots, u_r, x_{r+1}, \cdots, x_k) = \sum f(x_1, x_2, \cdots, x_k) \left|\frac{\partial x_i}{\partial u_j}\right|_+ \quad (22)$$

where the sum is taken over all sets of solutions of (21) and where $|\partial x_i/\partial u_j|_+$ is the positive value of the determinant of the partial derivatives of $x_i(u_1, \cdots, u_r, x_{r+1}, \cdots, x_k)$ with respect to the $u_j(i, j = 1, 2, \cdots, r)$. We omit the proof of (22); it is essentially the same as the derivation of the formulas for transforming variables in multiple integrals, which may be found in any textbook on advanced calculus.

Use of Moment Generating Functions. There is a second method of determining distributions of functions of random variables which we shall find to be particularly useful. If $u(x_1, x_2, \cdots, x_k)$ is a function of random variables x_i which are distributed by $f(x_1, x_2, \cdots, x_k)$, we may find the moment generating function of u

$$\begin{aligned} m(t) &= E(e^{tu}) \\ &= \int \cdots \int e^{tu(x_1, x_2, \cdots, x_k)} f(x_1, x_2, \cdots, x_k) \Pi dx_i \end{aligned} \quad (23)$$

If the resulting function of t can be recognized as the moment generating function of some known distribution, then it will follow that u has that distribution by virtue of the theorem given at the end of Chap. 5.

This method is quite powerful in connection with certain techniques of advanced mathematics (the theory of Laplace transforms and Fourier transforms) which enable one to determine the distribution associated with any given moment generating function. The method can also be generalized to determine the joint distribution of several functions of random variables.

10.2. Distribution of the Sample Mean for Normal Populations. If samples $(x_1, x_2, \cdots, x_n)$ of size n are drawn from a normal population, the joint density function for the observations is

$$f(x_1, x_2, \cdots, x_n) = \prod_i \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}[(x_i-\mu)/\sigma]^2}$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} e^{-\frac{1}{2}\Sigma[(x_i-\mu)/\sigma]^2} \tag{1}$$

and if the variates are transformed to

$$y_i = \frac{x_i - \mu}{\sigma}$$

the density becomes

$$h(y_1, \cdots, y_n) = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\Sigma y_i^2} \tag{2}$$

in accordance with equation (1.22) with $r = k = n$, since $|\partial x_i/\partial y_j|$ is a diagonal determinant with elements σ in the main diagonal positions and zeros elsewhere. The value of the determinant is readily seen to be σ^n.

To find the distribution of $\bar{y}$, we eliminate y_1 from (2) by the substitution

$$y_1 = n\bar{y} - \sum_2^n y_i = y_1(\bar{y}, y_2, \cdots, y_k) \tag{3}$$

and obtain the density

$$g(\bar{y}, y_2, y_3, \cdots, y_n) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} n e^{-\frac{1}{2}[(n\bar{y} - \sum_2^n y_i)^2 + \sum_2^n y_i^2]} \tag{4}$$

in accordance with (1.14) since $\partial y_1/\partial \bar{y} = n$. We now wish to find the marginal distribution of $\bar{y}$. The density in (4) may be regarded as a multivariate normal distribution of $\bar{y}, y_2, \cdots, y_n$, and examination of the exponent shows that

$$\|\sigma^{ij}\| = \begin{Vmatrix} n^2 & -n & -n & -n & \cdots & -n \\ -n & 2 & 1 & 1 & \cdots & 1 \\ -n & 1 & 2 & 1 & \cdots & 1 \\ -n & 1 & 1 & 2 & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ -n & 1 & 1 & 1 & \cdots & 2 \end{Vmatrix} \tag{5}$$

The determinant $|\sigma^{ij}|$ must necessarily be n^2, since in (4) it is seen that $\sqrt{|\sigma^{ij}|} = n$. We have seen in Sec. 9.5 that the marginal distribution of one of a set of normally distributed variates is a normal distribution with the same variance that the variate has in the joint distribution. We need therefore to find σ_{11}, which is obtained by dividing the cofactor of σ^{11} by $|\sigma^{ij}|$. The elements of the cofactor are obtained by striking out the first row and column of (5), and the determinant of the resulting array is easily found to be n. Hence

$$\sigma_{11} = \frac{n}{n^2} = \frac{1}{n}$$

The density of $\bar{y}$ is therefore

$$h(\bar{y}) = \frac{1}{\sqrt{2\pi}} \sqrt{n}\, e^{-\frac{1}{2}n\bar{y}^2} \tag{6}$$

Since

$$\bar{y} = \frac{\bar{x} - \mu}{\sigma} \tag{7}$$

we may transform (6) by (7) to obtain the density of $\bar{x}$,

$$n(\bar{x}) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sigma} e^{-(n/2)[(\bar{x}-\mu)/\sigma]^2} \tag{8}$$

by equation (1.13) since $d\bar{y}/d\bar{x} = 1/\sigma$.

The distribution (8) is the distribution approached by the distribution of $\bar{x}$ for any population with finite variance as n becomes large, as we have seen in Sec. 7.6. We have shown here that the distribution is exactly the distribution of the sample mean for normal populations whether or not the sample size is large.

10.3. The Chi-square Distribution. We shall obtain the distribution of

$$u = \sum_{i=1}^{k} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \tag{1}$$

where the x_i are normally and independently distributed with means μ_i and variance σ_i^2. In the joint distribution of the x_i we again transform the variates to

$$y_i = \frac{x_i - \mu_i}{\sigma_i}$$

in order to simplify the equations; u is then simply Σy_i^2. The method of moment generating functions will be employed to obtain its distribution.

The moment generating function of u is

$$m(t) = \left(\frac{1}{2\pi}\right)^{k/2} \int\int \cdots \int e^{t\Sigma y_i^2} e^{-\frac{1}{2}\Sigma y_i^2} \prod_i dy_i \tag{2}$$

and the multiple integral may be written as the product of k integrals of the form

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)y_i^2}\, dy_i \tag{3}$$

The integral (3) has the value $1/\sqrt{1-2t}$ since multiplication of the integral by $\sqrt{1-2t}$ makes it represent the area under a normal curve with variance $1/(1-2t)$. It follows that

$$m(t) = \left(\frac{1}{1-2t}\right)^{k/2} \qquad t < \frac{1}{2} \tag{4}$$

The moment generating function is of the form of the moment generating function for a gamma distribution (Sec. 6.3) with $\alpha = (k/2) - 1$ and $\beta = 2$. We may conclude therefore that the density of u is

$$f(u) = \frac{1}{[(k/2) - 1]!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-\frac{1}{2}u} \qquad u > 0 \tag{5}$$

This particular form of the gamma distribution is usually referred to as chi-square distribution with k *degrees of freedom*. The variate u is commonly designated by the square of the greek letter chi,

$$\chi^2 = \sum_1^k \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \tag{6}$$

hence the name for this distribution. The phrase *degrees of freedom* refers to the number of independent squares in the sum in (6); we may think of it, however, as merely a name for the parameter k in the density (5).

We may notice here that (5) gives essentially the distribution of the

maximum-likelihood estimator for σ^2 in normal populations when μ is known. If one considers samples of size n from a normal population with known mean μ, the maximum-likelihood estimator for σ^2 is found to be

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \mu)^2 = \frac{\sigma^2}{n} u$$

where $u = \sum^n [(x_i - \mu)/\sigma]^2$ has the chi-square distribution with n degrees of freedom. The density for the estimator is therefore

$$f(\hat{\sigma}^2) = \frac{1}{[(n/2) - 1]!}\left(\frac{n}{2\sigma^2}\right)^{n/2} (\hat{\sigma}^2)^{(n/2)-1} e^{-n\hat{\sigma}^2/2\sigma^2} \tag{7}$$

since

$$\frac{du}{d\hat{\sigma}^2} = \frac{n}{\sigma^2}$$

This is a gamma density with $\alpha = (n/2) - 1$ and $\beta = 2\sigma^2/n$.

The chi-square distribution is partially tabulated in Table III; the most complete tabulation is Karl Pearson's "Tables of the Incomplete Gamma Function" (Cambridge University Press, London, 1922).

10.4. Independence of the Sample Mean and Variance for Normal Populations. Ordinarily the mean of a population is unknown, and we are rather more interested in the estimator $(1/n)\Sigma(x_i - \bar{x})^2$ for σ^2 than in the estimator $(1/n)\Sigma(x_i - \mu)^2$ considered in the preceding section. We shall now derive the distribution of this estimator and show incidentally that it is distributed independently of the sample mean.

We shall let

$$y_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

$$u = n\bar{y}^2 = \frac{1}{n}\left(\sum y_i\right)^2 \tag{2}$$

$$v = \sum_1^n (y_i - \bar{y})^2 \tag{3}$$

and find the joint moment generating function for u and v, say,

$$m(t_1, t_2) = E(e^{t_1u + t_2v}) \tag{4}$$

$$= \int\int \cdots \int \left(\frac{1}{2\pi}\right)^{n/2} e^{(t_1/n)(\Sigma y_i)^2 + t_2\Sigma(y_i - \bar{y})^2 - \frac{1}{2}\Sigma y_i^2} \prod_1^n dy_i$$

$$= \int\int \cdots \int \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}[\Sigma y_i^2 - (2t_1/n)(\Sigma y_i)^2 - 2t_2\Sigma(y_i - \bar{y})^2]} \prod_1^n dy_i \tag{5}$$

The quadratic form may be written

$$\sum y_i^2 - \frac{2t_1}{n}\left(\sum y_i\right)^2 - 2t_2 \sum (y_i - \bar{y})^2$$

$$= \sum y_i^2 - \frac{2t_1}{n}\left(\sum y_i\right)^2 - 2t_2 \sum y_i^2 + 2nt_2\bar{y}^2 \tag{6}$$

$$= (1 - 2t_2) \sum y_i^2 - \frac{2(t_1 - t_2)}{n}\left(\sum y_i\right)^2$$

$$= \Sigma\Sigma\sigma^{ij} y_i y_j \tag{7}$$

where

$$\sigma^{ii} = 1 - 2t_2 - \frac{2(t_1 - t_2)}{n} = a$$

$$\sigma^{ij} = -\frac{2(t_1 - t_2)}{n} = b \qquad i \neq j$$

A determinant of order n with a's in the main diagonal and b's elsewhere has the value

$$(a - b)^{n-1}[a + (n - 1)b]$$

Hence

$$|\sigma^{ij}| = \left[1 - 2t_2 - \frac{2(t_1 - t_2)}{n} + \frac{2(t_1 - t_2)}{n}\right]^{n-1}$$

$$\left[1 - 2t_2 - \frac{2(t_1 - t_2)}{n} - \frac{n-1}{n}\,2(t_1 - t_2)\right]$$

$$= (1 - 2t_2)^{n-1}(1 - 2t_1) \tag{8}$$

From the multivariate normal distribution it follows that

$$\int_{-\infty}^{\infty} \cdots \int \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\Sigma\Sigma\sigma^{ij}y_iy_j} \prod dy_i = \frac{1}{\sqrt{|\sigma^{ij}|}} \tag{9}$$

hence the integral in (5) has the value

$$m(t_1, t_2) = \left(\frac{1}{1 - 2t_1}\right)^{\frac{1}{2}} \left(\frac{1}{1 - 2t_2}\right)^{(n-1)/2} \tag{10}$$

The fact that the joint-moment generating function factors into a function of t_1 alone and a function of t_2 alone implies that u and v are independently distributed. We shall not prove this rigorously but merely indicate the argument. Similar reasoning to that employed in Sec. 5.4 will show that if two distributions of several variates have the same joint-moment generating function, then the two distributions are the same. We have a density, say $f(u, v)$, with joint-moment

generating function (10). Given the marginal distributions $f_1(u)$ and $f_2(v)$, we may form the bivariate function

$$g(u, v) = f_1(u)f_2(v) , \tag{11}$$

which is clearly a density function. Furthermore its moment generating function must be

$$m(t_1, 0)m(0, t_2) \tag{12}$$

where

$$m(t_1, t_2) = \int\int e^{t_1u+t_2v}f(u, v)du\, dv \tag{13}$$

Since (12) and (13) are identical by (10), it follows that $g(u, v)$ and $f(u, v)$ are the same density and hence that $f(u, v)$ is equal to the product of its marginal densities.

The two factors of equation (10) are each of the form of the moment generating function for a chi-square distribution; hence it follows that u and v are each independently distributed by chi-square distributions, the first having one degree of freedom, and the second $n - 1$ degrees of freedom. The fact that $u = n\bar{y}^2$ is distributed as chi square with one degree of freedom is in accord with the results of Secs. 2 and 3. For we have seen that $\bar{y}$ is normally distributed with zero mean and variance $1/n$, and from the result of Sec. 3 with $k = 1$ it follows that

$$u = \frac{(\bar{y} - 0)^2}{1/n} = n\bar{y}^2 = n\left(\frac{\bar{x} - \mu}{\sigma}\right)^2 \tag{14}$$

must have the chi-square distribution with one degree of freedom.

The function

$$v = \sum_1^n (y_i - \bar{y})^2 = \sum_1^n\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 \tag{15}$$

has the distribution given by equation (3.5) with k replaced by $n - 1$ instead of n, as would be the case if the deviations were measured from the population mean. It is sometimes said that one degree of freedom is lost by taking the sum of squares of deviations from the sample mean rather than the population mean, or that one degree of freedom is used up in estimating the mean. While v in equation (15) is the sum of n squares, the squares are not all functionally independent. The relation $\Sigma y_i = n\bar{y}$ enables one to compute any one of the deviations $y_i - \bar{y}$, given the other $n - 1$ of them.

In terms of v of (15), the estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 \tag{16}$$

has the value

$$\hat{\sigma}^2 = \frac{\sigma^2 v}{n}$$

The density for this estimator is therefore:

$$f(\hat{\sigma}^2) = \frac{1}{[(n-3)/2]!}\left(\frac{n}{2\sigma^2}\right)^{(n-1)/2}(\hat{\sigma}^2)^{(n-3)/2}e^{-(n\hat{\sigma}^2/2\sigma^2)} \tag{17}$$

All the results of this section apply only to normal populations. It can be proved that for no other distributions are (1) the sample mean and sample variance independently distributed, or (2) the sample mean exactly normally distributed, or (3) the sum of squares of deviations, from either the population or sample mean, exactly distributed by the chi-square law.

10.5. The F Distribution. A distribution which we shall later find to be of considerable practical interest is that of the ratio of two quantities independently distributed by chi-square laws. Suppose u and v are independently distributed by chi-square distributions with m and n degrees of freedom, respectively. Their joint density is, by (3.5),

$$f(u, v) = \frac{1}{[(m-2)/2]![(n-2)/2]!2^{(m+n)/2}}u^{(m-2)/2}v^{(n-2)/2}e^{-\frac{1}{2}(u+v)} \tag{1}$$

We shall find the distribution of the quantity

$$F = \frac{u/m}{v/n} = \frac{nu}{mv} \tag{2}$$

which is sometimes referred to as the *variance ratio*. We shall find the density of F by eliminating u in terms of F in (1) and then integrating out v from the resulting density. Since

$$\frac{\partial u}{\partial F} = \frac{mv}{n} \tag{3}$$

and since F is a monotonic function of u, the joint density of F and v is, say,

$$g(F, v) = \frac{1}{[(m-2)/2]![(n-2)/2]!2^{(m+n)/2}} v^{(n-2)/2}\left(\frac{mvF}{n}\right)^{(m-2)/2} e^{-\frac{1}{2}[v+(mvF/n)]}\frac{mv}{n} \tag{4}$$

To integrate out v, we must evaluate the integral

$$\int_0^\infty v^{(m+n-2)/2}e^{-\frac{1}{2}[1+(mF/n)]v}\,dv \tag{5}$$

of the factors in (4) which involve v. We observe that the integrand is, apart from certain constants, the integral of a gamma density over its whole range. In fact, if the integral were multiplied by

$$\frac{\{\frac{1}{2}[1 + (mF/n)]\}^{(m+n)/2}}{[(m + n - 2)/2]!} \tag{6}$$

it would be exactly the area under the gamma density with

$$\alpha = \frac{(m + n - 2)}{2}$$

and $\beta = \dfrac{1}{\frac{1}{2}[1 + (mF/n)]}$, and would have the value one. Hence the value of (5) is the reciprocal of the expression (6). The density of F is therefore

$$\begin{aligned} h(F) &= \int_0^\infty g(F, v)dv \\ &= \frac{1}{\left(\frac{m-2}{2}\right)!\left(\frac{n-2}{2}\right)!\,2^{\frac{m+n}{2}}}\left(\frac{m}{n}\right)^{\frac{m}{2}} F^{\frac{m-2}{2}} \int_0^\infty v^{\frac{m+n-2}{2}} e^{-\frac{1}{2}\left(1+\frac{mF}{n}\right)v}\, dv \\ &= \frac{\left(\frac{m+n-2}{2}\right)!}{\left(\frac{m-2}{2}\right)!\left(\frac{n-2}{2}\right)!}\left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{F^{\frac{m-2}{2}}}{\left(1 + \frac{mF}{n}\right)^{\frac{m+n}{2}}} \qquad F > 0 \end{aligned} \tag{7}$$

a function with two parameters m and n. These parameters are also called *degrees of freedom;* thus (7) is called the F density with m *and* n *degrees of freedom;* the number of degrees of freedom of the variate u in the numerator of F is always quoted first.

Five points on the upper tail of the cumulative distribution of F are given in Table V. More complete tables may be found in the reference cited in the footnote to Table V and in Fisher and Yates, "Statistical Tables" (Oliver & Boyd, Ltd., Edinburgh and London, 1938). The reciprocals of the numbers in Table V provide five points on the lower tail of the cumulative distribution. To evaluate in general an integral of the form

$$P(a < F < b) = \int_a^b h(F)dF$$

one may transform the distribution to the beta distribution and use Karl Pearson's "Tables of the Incomplete Beta Function" (Cambridge

University Press, London, 1932). The required transformation is

$$w = \frac{mF/n}{1 + (mF/n)} \tag{8}$$

which changes (7) to a beta density with parameters $\alpha = (m - 2)/$ and $\beta = (n - 2)/2$.

10.6. "Student's" t Distribution. Another distribution of considerable practical importance is that of the ratio of a normally distributed variate to the square root of a variate independently distributed by the chi-square distribution. More precisely, if x is normally distributed with mean μ and variance σ^2, if u has the chi-square distribution with k degrees of freedom, and if x and u are independently distributed, we seek the distribution of

$$t = \frac{(x - \mu)/\sigma}{\sqrt{u/k}} \tag{1}$$

and letting

$$y = \frac{x - \mu}{\sigma}$$

t becomes $\dfrac{y}{\sqrt{u/k}}$. The joint density of y and u is

$$f(y, u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \frac{1}{[(k - 2)/2]!\,2^{k/2}} u^{(k-2)/2} e^{-\frac{1}{2}u} \tag{2}$$

and we find the distribution of t by the same procedure as was used in the preceding section. We substitute for y in terms of t ($y = t\sqrt{u/k}$) in (2) and then integrate out u from the resulting function. The final result is

$$h(t) = \frac{[(k - 1)/2]!}{\sqrt{k\pi}\,[(k - 2)/2]!} \frac{1}{[1 + (t^2/k)]^{(k+1)/2}} \qquad -\infty < t < \infty \tag{3}$$

a distribution with one parameter k, which is also referred to as the number of degrees of freedom of the distribution. Since $[(x - \mu)/\sigma]^2$ has the chi-square distribution with one degree of freedom, it is evident from (1) that t^2 has the F distribution with one and k degrees of freedom. The cumulative form of the distribution is partially tabulated in Table IV.

10.7. Distribution of Sample Means for Binomial and Poisson Populations. In the preceding sections we have illustrated the two methods of finding distributions of functions of continuous random variables described in the first section. Here we shall illustrate the technique for discrete variates in two cases of particular interest.

If $x_1, x_2, \cdots, x_n$ is a sample of size n from the binomial population which has density

$$f(x) = p^x q^{1-x} \qquad x = 0, 1 \tag{1}$$

the joint density of the x's is simply

$$f(x_1, x_2, \cdots, x_n) = p^{\Sigma x_i} q^{n-\Sigma x_i} \qquad x_i = 0, 1 \tag{2}$$

The sample mean is

$$\bar{x} = \frac{1}{n} \sum x_i$$

a function of the random variates, and it is evident that the only possible values of $\bar{x}$ are $0, 1/n, 2/n, \cdots, 1$. The probability, $g(j/n)$, that $\bar{x}$ takes on the value j/n is obtained by summing (2) over all sets $(x_1, x_2, \cdots, x_n)$ such that $(1/n)\Sigma x_i = j/n$, or such that $\Sigma x_i = j$. For all such sets, $f(x_1, x_2, \cdots, x_n)$ has the same value $p^j q^{n-j}$; hence the sum may be evaluated by multiplying this value by the number of sets $(x_1, x_2, \cdots, x_n)$ with the required specification. The number of such sets is the number of arrangements of j ones and $n - j$ zeros, which is $\binom{n}{j}$; hence

$$g\left(\frac{j}{n}\right) = \binom{n}{j} p^j q^{n-j} \qquad \frac{j}{n} = 0, \frac{1}{n}, \frac{2}{n}, \cdots, 1 \tag{3}$$

as we have found already in Sec. 7.7.

In a similar manner we may find the distribution of the mean of a sample, $x_1, x_2, \cdots, x_n$, from a Poisson population. The joint density of the observations is

$$f(x_1, x_2, \cdots, x_n) = \frac{e^{-n\mu}\mu^{\Sigma x_i}}{\prod_i x_i!} \qquad x_i = 0, 1, 2, \cdots \tag{4}$$

using μ for the parameter of the distribution. The sample mean $\bar{x}$ can obviously have any of the values j/n where $j = 0, 1, 2, \cdots$. For a particular value j/n, the x's must be such that $\Sigma x_i = j$; hence

$$\begin{aligned} g\left(\frac{j}{n}\right) &= \sum_{\Sigma x_i = j} \frac{e^{-n\mu}\mu^j}{\Pi x_i!} \\ &= e^{-n\mu}\mu^j \sum_{\Sigma x_i = j} \frac{1}{\prod_i x_i!} \end{aligned} \tag{5}$$

The sum can be performed with the aid of the multinomial theorem which, on putting all $x_i = 1$ in equation (2.5.2), states that

$$\sum \frac{j!}{\Pi x_i!} = n^j$$

The sum is therefore $n^j/j!$, and the required density is

$$g\left(\frac{j}{n}\right) = \frac{e^{-n\mu}(n\mu)^j}{j!} \qquad \bar{x} = \frac{j}{n} = 0, \frac{1}{n}, \frac{2}{n}, \cdots \tag{6}$$

The function may be written explicitly as a function of $\bar{x}$:

$$g(\bar{x}) = \frac{e^{-n\mu}(n\mu)^{n\bar{x}}}{(n\bar{x})!} \tag{7}$$

We may notice that since there is a unique correspondence between $\bar{x} = j/n$ and $j = \Sigma x_i$, the density of j is

$$h(j) = \frac{e^{-n\mu}(n\mu)^j}{j!} \qquad j = 0, 1, 2, \cdots$$

and hence that the sum of n observations from a Poisson population has a Poisson distribution with the parameter equal to n times the parameter of the original distribution.

10.8. Large-sample Distribution of Maximum-likelihood Estimators. We have investigated several special problems in sampling theory not only to illustrate the methods of finding sampling distributions, but because the particular distributions we have obtained are important in applied statistics. They are sometimes referred to as "small-sample distributions," though of course they hold for large or small samples and the term is merely meant to indicate that they are valid for small samples. In this section we shall consider a distribution much more general, in the sense that it is more or less independent of the form of the population distribution, but valid only for large samples.

We shall first consider the case of one parameter θ in a density $f(x_i; \theta)$, and we shall show that the maximum-likelihood estimator $\hat{\theta}(x_1, x_2, \cdots, x_n)$ for θ from samples of size n is approximately normally distributed under rather general conditions where n is large. Before doing so, it is necessary to consider the variate

$$u(x) = \frac{\partial}{\partial\theta} \log f(x; \theta) \tag{1}$$

The expected value of u is

$$E(u) = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right] f(x;\theta)dx \tag{2}$$

$$= \int \frac{\partial}{\partial\theta} f(x;\theta)dx \tag{3}$$

If $f(x;\theta)$ is such that the operations of differentiation and integration may be interchanged, then

$$E(u) = \frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f(x;\theta)dx$$

$$= \frac{\partial}{\partial\theta}(1) = 0 \tag{4}$$

Hence, if this condition is satisfied, the variance of u is

$$\sigma_u^2 = \int \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right]^2 f(x;\theta)dx \tag{5}$$

$$= \int \frac{1}{f(x;\theta)} \left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 dx \tag{6}$$

and this may be put in another form which is more useful for our purpose. On differentiating (2) with respect to θ, we have

$$0 = \frac{\partial}{\partial\theta} \int \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right) f(x;\theta)dx$$

$$= \int \left(\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)\right) f(x;\theta)dx + \int \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right) \frac{\partial f(x;\theta)}{\partial\theta} dx$$

$$= \int \left(\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)\right) f(x;\theta)dx + \int \frac{1}{f(x;\theta)} \left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 dx \tag{7}$$

The integral in (6) is therefore minus the first integral in (7), and we may write

$$\sigma_u^2 = -E\left(\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)\right) \tag{8}$$

Now suppose a sample of size n is drawn from the population. This will give rise to a sample of u values:

$$u_i = u(x_i) = \frac{\partial}{\partial\theta} \log f(x_i;\theta) \qquad i = 1, 2, \cdots, n \tag{9}$$

Applying the central-limit theorem (Sec. 7.6) to the sample of u's, we may state that

$$\bar{u} = \frac{1}{n} \sum u_i$$

is approximately normally distributed for large n with zero mean and variance σ_u^2/n. Remembering that the likelihood of the sample of x's is $\Pi f(x_i; \theta)$ and that its logarithm (Sec. 8.4) is

$$L = \Sigma \log f(x_i; \theta) \tag{10}$$

we have

$$\bar{u} = \frac{1}{n} \frac{\partial L}{\partial \theta} \tag{11}$$

Hence it follows that $\partial L/\partial \theta$ is approximately normally distributed for large n with mean zero and variance $n\sigma_u^2$.

This last result enables us to find the distribution of the estimator $\hat{\theta}$. We shall suppose that $\hat{\theta}$ is a root of

$$\frac{\partial L}{\partial \theta} = 0 \tag{12}$$

i.e., that L actually has zero slope at its maximum value. And we shall suppose that $\partial L(\hat{\theta})/\partial \theta$ as a function of $\hat{\theta}$ may be expanded in a Taylor series about θ:

$$\frac{\partial L(\hat{\theta})}{\partial \theta} = \frac{\partial L(\theta)}{\partial \theta} + \frac{\partial^2 L(\theta)}{\partial \theta^2} (\hat{\theta} - \theta) + \frac{\partial^3 L(\tilde{\theta})}{2!\, \partial \theta^3} (\hat{\theta} - \theta)^2 \tag{13}$$

where $\tilde{\theta}$ is some point between θ and $\hat{\theta}$. Since $\hat{\theta}$ is a root of $\partial L(\theta)/\partial \theta$, (13) vanishes, and we have

$$\frac{\partial L(\theta)}{\partial \theta} = - \frac{\partial^2 L(\theta)}{\partial \theta^2} (\hat{\theta} - \theta) - \frac{\partial^3 L(\tilde{\theta})}{2!\, \partial \theta^3} (\hat{\theta} - \theta)^2 \tag{14}$$

Now we have seen that

$$\frac{1}{\sqrt{n}\, \sigma_u} \frac{\partial L(\theta)}{\partial \theta}$$

is approximately normally distributed for large n with zero mean and unit variance. Using (14), this expression is

$$\frac{1}{\sqrt{n}\, \sigma_u} \frac{\partial L(\theta)}{\partial \theta} = - \frac{1}{\sqrt{n}\, \sigma_u} \frac{\partial^2 L(\theta)}{\partial \theta^2} (\hat{\theta} - \theta) - \frac{1}{\sqrt{n}\, \sigma_u} \frac{\partial^3 L(\tilde{\theta})}{2!\, \partial \theta^3} (\hat{\theta} - \theta)^2 \tag{15}$$

and on the right we shall substitute $w = \sqrt{n}\, \sigma_u(\hat{\theta} - \theta)$ to get

$$\frac{1}{\sqrt{n}\, \sigma_u} \frac{\partial L(\theta)}{\partial \theta} = - w \frac{1}{\sigma_u^2} \left[\frac{1}{n} \frac{\partial^2 L(\theta)}{\partial \theta^2} \right] - \frac{w^2}{\sqrt{n}\, \sigma_u^3} \left[\frac{1}{n} \frac{\partial^3 L(\theta)}{2!\, \partial \theta^3} \right] \tag{16}$$

The first bracket on the right of (16) is simply an average for samples of size n of $\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)$ and by virtue of (8) has a mean value σ_u^2. Furthermore, if this quantity has a finite variance, $(1/n)(\partial^2 L/\partial\theta^2)$ will approach σ_u^2 with probability approaching one as n becomes infinite. The first term of (16) is therefore nearly w for large n. The second term of (16) approaches zero because of the factor $1/\sqrt{n}$ if we assume that the average of the third derivative of $\log f(x;\tilde{\theta})$ cannot become infinite for any possible value of $\tilde{\theta}$. The right of (16) is therefore approximately w, and since the left of (16) is approximately normal with zero mean and unit variance, it follows that w has approximately the same distribution. We have finally that $\hat{\theta}$ is, for large samples, approximately normally distributed with mean θ (the true parameter value) and variance $1/n\sigma_u^2$, where σ_u^2 is defined by (8). The mean θ will be the exact mean of $\hat{\theta}$ for any sample size only if $\hat{\theta}$ is an unbiased estimator. In general, we have seen that maximum-likelihood estimators are not unbiased so that θ is the *large-sample mean*, i.e., the value approached by the mean as n becomes large. Similarly $1/n\sigma_u^2$ may be the exact variance, or it may be only the limiting form of the exact variance as n becomes large, the *large-sample variance*. One could, of course, compute the variance of $\hat{\theta}$ directly by

$$E[\hat{\theta} - E(\hat{\theta})]^2 = \int \cdots \int [\hat{\theta} - E(\theta)]^2 \Pi f(x_i;\theta) dx_i$$

rather than by means of equation (8), but this is usually the more difficult computation.

The above argument is not, of course, a proof of the asymptotic normality of $\hat{\theta}$; we have merely outlined the nature of the proof. A rigorous demonstration requires careful evaluation of the errors in the various approximations. While the maximum-likelihood estimator is approximately normally distributed for large samples under rather general conditions, it is to be remarked that several conditions on the original distribution must be fulfilled:

(1) It must be permissible to interchange the operations of integration with respect to x and differentiation with respect to θ.

(2) The expected value of $\frac{\partial}{\partial\theta} \log f(x;\theta)$ must be zero.

(3) $\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)$ must have finite mean and variance.

(4) $\frac{1}{n}\frac{\partial^3}{\partial\theta^3} L(\theta)$ must remain bounded for all possible values of θ.

(5) The derivative of $L(\theta)$ must vanish at its maximum.

These conditions will not be fulfilled, for example, if the parameter is the range or a function of the range, for then (1) is not satisfied. We have seen in particular that if θ is the range of a rectangular distribution, condition (5) is not fulfilled.

For a wide class of distributions, however, the maximum-likelihood estimator is approximately normally distributed about the true parameter value as a mean for large samples. This is a powerful tool for solving many important problems of applied statistics as we shall see in the following chapters. The theorem is applicable to discrete as well as to continuous distributions. The only change in the reasoning for discrete distributions would be replacement of the integral signs by summation signs.

A straightforward extension of the argument will provide an analogous result for the large-sample distribution of several parameters. We shall merely state the result:

The maximum-likelihood estimators $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k$ for the parameters of a density $f(x; \theta_1, \theta_2, \cdots, \theta_k)$ from samples of size n are, for large samples, approximately distributed by the multivariate normal distribution with means $\theta_1, \theta_2, \cdots, \theta_k$ and with coefficients $\|n\sigma^{ij}\|$ in the quadratic form, where

$$\sigma^{ij} = -E\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(x; \theta_1, \theta_2, \cdots, \theta_k)\right] \tag{17}$$

The variances and covariances of the estimators are $\|(1/n)\sigma_{ij}\|$, where

$$\|\sigma_{ij}\| = \|\sigma^{ij}\|^{-1} \tag{18}$$

The conditions under which this theorem is true are essentially the same as those given in the case of one parameter.

The theorems obviously depend in no way on the fact that we have used univariate distributions. The variate x in all the statements of this section may be replaced by a set of variates $(x, y, z, \cdots)$.

10.9 Applications of the Large-sample Theory. To illustrate the use of the theorem just given, we may find the large-sample distribution for the estimators of the two parameters of the normal distribution. We shall write it in the form

$$f(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}\, e^{-(1/2\theta_2)(x-\theta_1)^2} \tag{1}$$

For samples of size n we have seen that the maximum-likelihood estimators are

$$\hat{\theta}_1 = \frac{1}{n}\sum x_i \tag{2}$$

$$\hat{\theta}_2 = \frac{1}{n}\sum (x_i - \hat{\theta}_1)^2 \tag{3}$$

In accordance with the theorem, these estimators will be approximately normally distributed for large samples with means θ_1 and θ_2 and coefficients $n\sigma^{ij}$ in the quadratic form, where

$$\sigma^{ij} = -E\left(\frac{\partial^2 \log f}{\partial\theta_i\,\partial\theta_j}\right) \tag{4}$$

Since

$$\log f = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\theta_2 - \frac{1}{2\theta_2}(x - \theta_1)^2 \tag{5}$$

the required derivatives are

$$\frac{\partial^2 \log f}{\partial\theta_1^2} = -\frac{1}{\theta_2}$$

$$\frac{\partial^2 \log f}{\partial\theta_1\,\partial\theta_2} = -\frac{x - \theta_1}{\theta_2^2}$$

$$\frac{\partial^2 \log f}{\partial\theta_2^2} = \frac{1}{2\theta_2^2} - \frac{(x - \theta_1)^2}{\theta_2^3}$$

and because

$$E(x) = \theta_1 \qquad E(x - \theta_1)^2 = \theta_2$$

the σ^{ij} are readily seen to be

$$\|\sigma^{ij}\| = \begin{Vmatrix} \dfrac{1}{\theta_2} & 0 \\ 0 & \dfrac{1}{2\theta_2^2} \end{Vmatrix} \tag{6}$$

The large-sample distribution of the estimators is, therefore, say,

$$g(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{2\pi}\frac{n}{\sqrt{2\theta_2^3}}\, e^{-\frac{n}{2}\left[\frac{(\hat{\theta}_1 - \theta_1)^2}{\theta_2} + \frac{(\hat{\theta}_2 - \theta_2)^2}{2\theta_2^2}\right]} \tag{7}$$

with large-sample variances and covariances given by

$$\left\|\frac{1}{n}\sigma_{ij}\right\| = \begin{Vmatrix} \dfrac{\theta_2}{n} & 0 \\ 0 & \dfrac{2\theta_2^2}{n} \end{Vmatrix} \tag{8}$$

Since $\sigma_{12} = 0$, the estimators are shown to be independently distributed for large samples; we have already seen, of course, in Sec. 4 that they are actually independent for any sample size. The large-sample distribution of $\hat{\theta}_1$ is exactly the normal distribution as given in (7). But the exact distribution of $\hat{\theta}_2$ is given by the gamma distribution for any sample size and this appears to conflict with the normal distribution indicated in (7). However, it can be shown that the exact distribution of $\hat{\theta}_2$ does approach the normal form

$$\frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{2}} \frac{1}{\theta_2} e^{-\frac{n}{2} \frac{(\hat{\theta}_2 - \theta_2)^2}{2\theta_2^2}}$$

as n becomes large (see Prob. 38, Chap. 6).

As a second illustration, we shall obtain the large-sample distribution of the parameters of a multinomial distribution.

Suppose the elements of a population may be classified into $k + 1$ categories, say $A_1, A_2, \cdots, A_{k+1}$. We shall describe an element by the set of variables $(x_1, x_2, \cdots, x_{k+1})$ where, if the element belongs to A_i, $x_i = 1$ and all the other x's are zero. If the probability is p_i that an element drawn at random belongs to A_i, then the joint density of the x's is

$$f(x_1, x_2, \cdots, x_{k+1}) = p_1^{x_1} p_2^{x_2} \cdots p_{k+1}^{x_{k+1}} \qquad x_i = 0, 1: \Sigma x_i = 1 \quad (9)$$

where $\Sigma p_i = 1$. Summing $f(x_1, \cdots, x_{k+1})$ over all possible sets of x's, namely, $(1, 0, 0, \cdots, 0)$, $(0, 1, 0, 0, \cdots, 0)$, $(0, 0, 1, 0, \cdots, 0)$, and so on, we have

$$\sum f(x_1, \cdots, x_{k+1}) = \sum_{i=1}^{k+1} p_i = 1$$

The distribution (9) is a multivariate distribution with k functionally independent parameters; we shall take them to be $p_1, p_2, \cdots, p_k$ and think of p_{k+1} as a symbol for $1 - p_1 - p_2 - \cdots - p_k$.

Let a sample of size n be drawn, and let n_i be the number of sample elements in A_i; then $\Sigma n_i = n$ and the likelihood of the sample is

$$\prod_{i=1}^{k+1} p_i^{n_i}$$

the logarithm of which is

$$L(p_1, p_2, \cdots, p_k) = \sum_{i=1}^{k+1} n_i \log p_i \qquad (10)$$

The estimators are found by putting the first derivatives of L equal to zero and solving for the parameters. The equations are

$$\begin{aligned} \frac{\partial L}{\partial p_1} &= \frac{n_1}{p_1} - \frac{n_{k+1}}{p_{k+1}} = 0 \\ \frac{\partial L}{\partial p_2} &= \frac{n_2}{p_2} - \frac{n_{k+1}}{p_{k+1}} = 0 \end{aligned} \tag{11}$$

and so on, remembering that p_{k+1} represents $1 - p_1 - p_2 - \cdots - p_k$. On multiplying the first equation by $p_1 p_{k+1}$, the second by $p_2 p_{k+1}$, and so on, and adding the results, one finds $\hat{p}_{k+1} = n_{k+1}/n$, and then that

$$\hat{p}_i = \frac{n_i}{n} \qquad i = 1, 2, \cdots, k \tag{12}$$

We wish to find the approximate distribution of the estimators in (12) for large samples. Applying the theorem of the preceding section, we know that the distribution is normal and that the means are p_i. We need only to find the coefficients $n\sigma^{ij}$ of the quadratic form. By equation (8.17)

$$\sigma^{ij} = -E\left(\frac{\partial^2}{\partial p_i\, \partial p_j} \log f\right) \tag{13}$$

Differentiating $\log f$, we have

$$\begin{aligned} \frac{\partial^2}{\partial p_i\, \partial p_j} \log f &= -\frac{x_{k+1}}{p_{k+1}^2} \qquad \text{if } i \neq j \\ &= -\frac{x_i}{p_i^2} - \frac{x_{k+1}}{p_{k+1}^2} \qquad \text{if } i = j \end{aligned} \tag{14}$$

and taking expected values,

$$\begin{aligned} E(x_i) &= \sum x_i \prod_1^{k+1} p_i^{x_i} = p_i \\ E(x_{k+1}) &= \sum x_{k+1} \prod_1^{k+1} p_i^{x_i} = p_{k+1} \end{aligned} \tag{15}$$

Thus

$$\begin{aligned} \sigma^{ij} &= \frac{1}{p_{k+1}} \qquad \text{if } i \neq j \\ &= \frac{1}{p_i} + \frac{1}{p_{k+1}} \qquad \text{if } i = j \end{aligned} \tag{16}$$

and we may write these two relations as one using the symbol δ_{ij},

$$\sigma^{ij} = \frac{\delta_{ij}}{p_i} + \frac{1}{p_{k+1}} \qquad i, j = 1, 2, \cdots, k \tag{17}$$

The value of the determinant $|\sigma^{ij}|$ can be shown to be $1 \Big/ \prod_1^{k+1} p_i$; hence the approximate large-sample distribution of the estimators is, say,

$$g(\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_k) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \sqrt{\frac{1}{\prod_1^{k+1} p_i}}\, n^{\frac{k}{2}} e^{-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} n\left(\frac{\delta_{ij}}{p_i}+\frac{1}{p_{k+1}}\right)(\hat{p}_i-p_i)(\hat{p}_j-p_j)} \tag{18}$$

The inverse of $\|\sigma^{ij}\|$ has elements

$$\begin{aligned} \sigma_{ii} &= p_i(1 - p_i) \\ \sigma_{ij} &= -p_i p_j \qquad i \neq j \qquad i, j = 1, 2, \cdots, k \end{aligned} \tag{19}$$

as may be verified by computing the product $\|\sigma^{ij}\| \cdot \|\sigma_{ij}\|$. The large-sample variances and covariances of the estimators are therefore given by multiplying (19) by $1/n$. These happen to be, in fact, the exact variances and covariances for any sample size

10.10. Problems

1. Apply the method of equation (1.4) to the example treated in equations (1.14) to (1.17).

2. If x is distributed by $f(x) = 2x$, $0 < x < 1$, find the distribution of $u = (3x - 1)^2$.

3. If x is distributed by $f(x) = 1$, $0 < x < 1$, find the distribution of $\bar{x}$ for samples x_1, x_2 of size two. Observe that the range of x_2 for fixed $\bar{x}$ is $0 < x_2 < 2\bar{x}$ when $\bar{x} < \frac{1}{2}$, and $2\bar{x} - 1 < x_2 < 1$ when $\bar{x} > \frac{1}{2}$.

4. If x is normally distributed with mean μ and variance σ^2, show by transforming the variate that $u = [(x - \mu)/\sigma]^2$ has the chi-square distribution with one degree of freedom.

5. Obtain the distribution of the mean of a sample of size n from a normal population by using the moment generating function.

6. If $\chi_1^2, \chi_2^2, \chi_3^2, \cdots \chi_k^2$ are independently distributed by chi-square laws with $n_1, n_2, \cdots, n_k$ degrees of freedom, respectively, show by means of the moment generating function that $u = \Sigma\chi_i^2$ has the chi-square distribution with $n = \Sigma n_i$ degrees of freedom.

7. Using an argument similar to that given for the derivation of the chi-square distribution and the fact that $|(1 - 2t)\sigma^{ij}| = (1 - 2t)^k|\sigma^{ij}|$, show that the quadratic form of a k-variate normal distribution has the chi-square distribution with k degrees of freedom.

8. Find the mean and variance of a chi-square variate with k degrees of freedom.

9. Use the integral of the F distribution over the whole range to obtain an identity in the parameters m and n, and then use the identity to obtain the mean and variance of F.

10. Find the .95 probability level of F for two and four degrees of freedom by direct integration of the distribution function.

✓**11.** Show that the transformation

$$w = \frac{mF/n}{1 + (mF/n)}$$

changes the F distribution to the beta distribution.

12. Show, by transforming the variate in the t distribution, that $u = t^2$ has the F distribution.

✓**13.** If $x_1, x_2, \cdots, x_n$ is a random sample from a normal population, show that

$$u = \frac{\bar{x} - \mu}{\sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n(n-1)}}}$$

has the t distribution with $n - 1$ degrees of freedom.

✓**14.** If x_1 and x_2 are a random sample of two from a population with $f(x) = e^{-x}$, $x > 0$, show that $u = x_1 + x_2$ and $v = x_1/x_2$ are independently distributed.

✓**15.** If x, y, z have the joint density

$$f(x, y, z) = \frac{6}{(1 + x + y + z)^4} \qquad x, y, z > 0$$

find the distribution of $u = x + y + z$.

✓**16.** If x_1 and x_2 are a random sample of two from a population with the uniform distribution over the unit interval, find the distribution of $u = x_1x_2$.

✓**17.** If x and y have the bivariate normal distribution, show that

$$u = \frac{x - \mu_x}{\sigma_x} + \frac{y - \mu_y}{\sigma_y}$$

and

$$v = \frac{x - \mu_x}{\sigma_x} - \frac{y - \mu_y}{\sigma_y}$$

are independently normally distributed with zero means and variances $2(1 + \rho)$ and $2(1 - \rho)$.

18. If x and y are independently normally distributed with zero means and unit variances, show that $u = x^2 + y^2$ and $v = x/y$ are independently distributed. What are the names of the individual distributions of u and v?

19. Show that "Student's" distribution approaches the normal form when the number of degrees of freedom becomes infinite.

20. If $x_1, x_2, \cdots, x_n$ are a random sample from a normal population, find the joint distribution of

$$u = \sum_1^k x_i \quad \text{and} \quad v = \sum_r^n x_i \qquad 0 < r < k < n$$

21. If x and y are independently distributed by chi-square laws with m and n degrees of freedom, respectively, show that $u = x + y$ and $v = x/y$ are independently distributed.

22. Consider samples of size n from a bivariate normal distribution. Using the notation of Sec. 9.7, show that

$$\frac{\sqrt{n-1}\,(\hat{\xi}_1 - \hat{\xi}_2 - \xi_1 + \xi_2)}{\sqrt{\hat{\sigma}_{11} + \hat{\sigma}_{22} - 2\hat{\sigma}_{12}}}$$

has "Student's" distribution with $n - 1$ degrees of freedom.

23. If x and y are horizontal and vertical components of the deviations of a shot from the center of a target, and if x and y have a bivariate normal distribution with zero means, $\rho = 0.1$, and standard deviations of 10 inches, find the equation of an ellipse which will contain a shot with probability .95. (Use the result of Prob. 7.)

24. Find the mean and variance of $(1/n)\Sigma(x_i - \bar{x})^2$ for samples of size n from a normal population, and show that they approach the large-sample mean and variance, σ^2 and $2\sigma^4/n$, as n increases.

25. If $x_1, x_2, \cdots, x_k$ are independently and normally distributed with means μ_i and variance σ_i^2, show that

$$u = \sum^k a_i x_i$$

where the a_i are constants, is normally distributed with mean $\Sigma a_i\mu_i$ and variance $\Sigma a_i^2\sigma_i^2$. Then deduce the distribution of the sample mean from a normal population by putting $a_i = 1/n$.

26. Obtain a result similar to that of Prob. 25 when the x_i have the multivariate normal distribution.

27. Find the large-sample distribution for the estimator of the parameter β in the gamma distribution.

28. Find the large-sample distribution for the estimator of the parameter of the Poisson distribution.

29. If $(x_{1a}, x_{2a}, \cdots, x_{ka})$, $a = 1, 2, \cdots, n$ is a sample of size n from the multinomial population with density

$$\prod_1^k p_i^{x_i} \qquad x_i = 0, 1; \sum x_i = 1; \sum p_i = 1$$

find the distribution of the variates $n_i = \sum_a x_{ia}$, and find their variances and covariances.

30. Verify that $\|\sigma_{ij}\|$ defined in equation (9.19) is the inverse of $\|\sigma^{ij}\|$ given by equation (9.17).

31. Evaluate the determinant of $\|\sigma_{ij}\|$ in Prob. 30.

32. If $x_1, x_2, \cdots, x_n$ are independently normally distributed with the same mean but different variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$, show that $u = \dfrac{\Sigma y_i/\sigma_i^2}{\Sigma 1/\sigma_i^2}$ and $v = \Sigma(y_i - u)^2/\sigma_i^2$ are independently distributed. Show also that u is normal, while v has the chi-square distribution with $n - 1$ degrees of freedom.

33. Let s^2 denote $\Sigma(x_i - \bar{x})^2/(n - 1)$, the mean square for samples of size n. For three samples from normal populations (with variances σ_1^2, σ_2^2, and σ_3^2), the sample sizes being n_1, n_2, and n_3, find the joint density of

$$u = \frac{s_1^2}{s_3^2} \quad \text{and} \quad v = \frac{s_2^2}{s_3^2}$$

where the s_1^2 s_2^2, and s_3^2 are the sample mean squares.

34. Let a sample of size n_1 from a normal population (with variance σ_1^2) have mean square s_1^2, and let a second sample of size n_2 from a second normal population (with mean μ_2 and variance σ_2^2) have mean $\bar{x}$ and mean square s_2^2. Find the joint density of

$$u = \frac{\sqrt{n_2}\,\bar{x}}{s_2} \quad \text{and} \quad v = \frac{s_1^2}{s_2^2}$$

CHAPTER 11

INTERVAL ESTIMATION

11.1. Confidence Intervals. A point estimate of a parameter is not very meaningful without some measure of the possible error in the estimate. An estimate $\hat{\theta}$ of a parameter θ should be accompanied by some interval about $\hat{\theta}$, possibly of the form $\hat{\theta} - d$ to $\hat{\theta} + d$, together with some measure of assurance that the true parameter θ does lie within the interval. Estimates are often given in such form. Thus the electronic charge may be estimated to be $(4.770 \pm .005)10^{-10}$ electrostatic unit with the idea that the first factor is very unlikely to be outside the range 4.765 to 4.775. A cost accountant for a publishing company in trying to allow for all factors which enter into the cost of producing a certain book (actual production costs, proportion of plant overhead, proportion of executive salaries, etc.) may estimate the cost to be 83 ± 4.5 cents per volume with the implication that the correct cost very probably lies between 78.5 and 87.5 cents per volume. The Bureau of Labor Statistics may estimate the number of unemployed to be $2.4 \pm .3$ millions at a given time, feeling rather sure that the actual number is between 2.1 and 2.7 millions.

In order to give precision to these ideas, we shall consider a particular example. Suppose a sample (1.2, 3.4, 0.6, 5.6) of four observations is drawn from a normal population with unknown mean μ and known standard deviation 3. The maximum-likelihood estimate of μ is the mean of the sample observations:

$$\bar{x} = 2.7 \tag{1}$$

We wish to determine upper and lower limits which are rather certain to contain the true parameter value between them.

In general, for samples of size four from the given distribution, the quantity

$$y = \frac{\bar{x} - \mu}{3/2} \tag{2}$$

will be normally distributed with zero mean and unit variance. $\bar{x}$ is the sample mean, and $3/2$ is $\sigma/\sqrt{n}$. Thus the quantity y has a density

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{3}$$

which is independent of the true value of the unknown parameter, and we can compute the probability that y will be between any two arbitrarily chosen numbers. Thus, for example,

$$P(-1.96 < y < 1.96) = \int_{-1.96}^{1.96} f(y)dy = .95 \tag{4}$$

In this relation the inequality $-1.96 < y$, or

$$-1.96 < \frac{\bar{x} - \mu}{3/2}$$

is equivalent to the inequality

$$\mu < \bar{x} + 3/2(1.96) = \bar{x} + 2.94$$

and the inequality

$$y < 1.96$$

is equivalent to

$$\mu > \bar{x} - 2.94$$

We may therefore rewrite (4) in the form

$$P(\bar{x} - 2.94 < \mu < \bar{x} + 2.94) = .95 \tag{5}$$

and substituting 2.7 for $\bar{x}$,

$$P(-.24 < \mu < 5.64) = .95 \tag{6}$$

Thus two limits have been obtained (−.24, 5.64), which we may say are 95 per cent certain to contain the true parameter value between them.

The meaning of (6) needs to be examined carefully. It appears that μ is the variable and that the statement implies that the probability that the variable μ lies between −.24 and 5.64 is .95. This is, of course, nonsense. μ is a fixed number, the mean of the population sampled. Furthermore the true mean μ either does or does not lie between −.24 and 5.64. The only correct probability statements possible in this situation are

$$P(-.24 < \mu < 5.64) = 1$$

if μ actually is between the numbers, or

$$P(-.24 < \mu < 5.64) = 0$$

if μ is not between the numbers. It is possible, however, to give (6) a meaningful interpretation.

The statement in equation (5) does have meaning. The probability that the *random interval*, $\bar{x} - 2.94$ to $\bar{x} + 2.94$, covers the true mean μ is .95. That is, if samples of four were repeatedly drawn from the population, and if the random interval $\bar{x} - 2.94$ to $\bar{x} + 2.94$ were computed for each sample, then 95 per cent of those intervals would be expected to contain the true mean μ. We do therefore have considerable confidence that the interval $-.24$ to 5.64 does cover the true mean. The measure of our confidence is .95 because *before* the sample was drawn, .95 was the probability that the interval we were going to construct would cover the true mean. In (5) the number .95 is a true probability; in (6) it is not a true probability although it is a measure of our confidence in the truth of the statement on the left of (6). We shall call it the *confidence coefficient*, or the *fiducial probability*, to distinguish it from our ordinary concept of probability. And we shall rewrite (6) as

$$P_F(-.24 < \mu < 5.64) = .95 \tag{7}$$

and read it "The fiducial probability that the interval $-.24$ to 5.64 covers the true mean is .95." The word *fiducial* indicates nothing more than that the probability associated with the given interval was .95 before the sample was drawn.

The interval $-.24$ to 5.64 is called a *confidence interval;* more specifically it is called a 95 per cent confidence interval, the confidence coefficient, or fiducial probability, being expressed as a percentage. We can obtain intervals with any desired degree of confidence. Thus, since

$$P(-2.58 < y < 2.58) = .99 \tag{8}$$

a 99 per cent confidence interval for the true mean is obtained by converting the inequalities as before and substituting $\bar{x} = 2.7$ to get

$$P_F(-1.17 < \mu < 6.57) = .99 \tag{9}$$

It is to be observed that there are, in fact, many possible intervals with the same fiducial probability. Thus, for example, since

$$P(-1.68 < y < 2.70) = .95 \tag{10}$$

another 95 per cent confidence interval for μ is given by

$$P_F(-1.35 < \mu < 5.22) = .95 \tag{11}$$

This interval is inferior to the one obtained before because its length 6.57 is greater than the length 5.88 of the interval in (7); it gives less precise information about the location of μ. Any numbers a and b such

that ordinates at those points include 95 per cent of the area under $f(y)$ will determine a 95 per cent confidence interval. Ordinarily one would want the confidence interval to be as short as possible, and it is made so by making a and b as close together as possible, because the relation $P(a < y < b) = .95$ gives rise to a confidence interval of length $(\sigma/\sqrt{n})(b - a)$. The distance $b - a$ will be minimized for fixed area when $f(a) = f(b)$, as is evident on referring to Fig. 45. If the point b is moved a short distance to the left, the point a will need to be moved a lesser distance to the left in order to keep the area the same; this operation decreases the length of the interval and will continue to do so as long as $f(b) < f(a)$. Since $f(y)$ is symmetric about $y = 0$ in the present example, the minimum value of $b - a$ for fixed area occurs

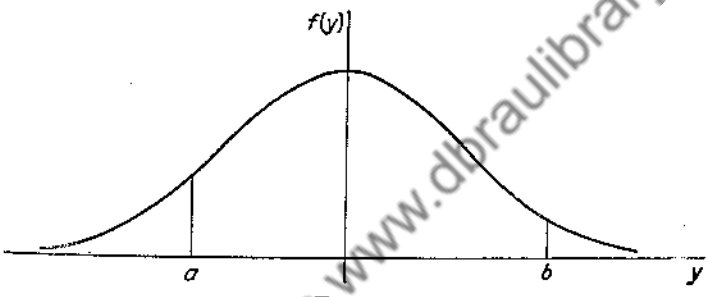


FIG. 45.

when $b = -a$. Thus (7) gives the shortest 95 per cent confidence interval, and (9) gives the shortest 99 per cent confidence interval for μ.

The general method illustrated here is as follows: One finds, if possible, a function of the sample observations and the parameter to be estimated (the function y above) which has a distribution independent of the parameter and any other parameters. Then any probability statement of the form $P(a < y < b) = \gamma$, where y is the function, will give rise to a fiducial statement about the parameter. This technique is applicable in many important problems, but in many others it is not, because it is impossible to find functions of the desired form which are distributed independently of any parameters. These latter problems can be dealt with by a more general technique to be described in Sec. 5.

The idea of interval estimation can be extended to include simultaneous estimation of several parameters. Thus the two parameters of the normal distribution may be estimated by some plane region R

in the so-called parameter space, the space of all possible combinations of values of μ and σ^2. A 95 per cent confidence region is a region constructible from the sample such that if samples were repeatedly drawn and a region constructed for each sample, 95 per cent of those regions on the average would include the true parameter point (μ_0, σ_0^2).

Confidence intervals and regions provide good illustrations of uncertain inferences. In (7) the inference is made that the interval -0.24 to 5.64 covers the true parameter value, but that statement is not made categorically. A measure, .05, of the uncertainty of the inference is an essential part of the statement.

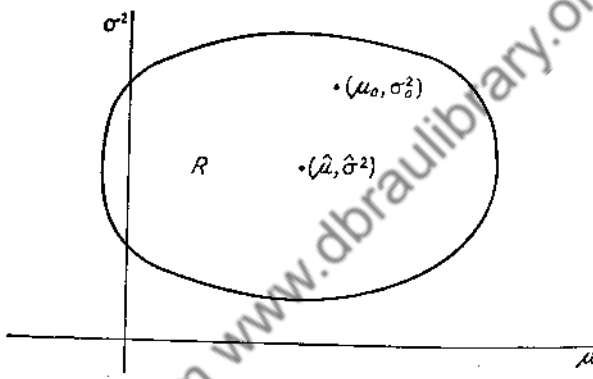


FIG. 46.

11.2. Confidence Intervals for the Mean of a Normal Distribution. The method used in the preceding section cannot ordinarily be used to estimate the mean of a normal population, because the variance σ^2 is not ordinarily known. The function y takes the form (for samples of size n)

$$y = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{1}$$

and on converting the inequalities in, say,

$$P(-1.96 < y < 1.96) = .95 \tag{2}$$

one finds

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95 \tag{3}$$

For a given sample, $\bar{x}$ and n are known, but σ is not, so that limits for μ cannot be computed. Of course, an estimate $\hat{\sigma}$ could be substituted for σ, but then the probability statement would no longer be exact and might be very far wrong for small samples.

The way around this difficulty was shown by W. S. Gossett (who wrote under the pseudonym of "Student") in a classic paper which introduced the t distribution. He is regarded as the founder of the modern theory of exact statistical inference. The quantity

$$t = \frac{\bar{x} - \mu}{\sqrt{\Sigma(x_i - \bar{x})^2/n(n-1)}} \quad (4)$$

$$= \frac{(\bar{x}-\mu)/\sigma/n^{1/2}}{\sqrt{\Sigma(x_i-\bar{x})^2/\sigma^2/(n-1)}}$$

involves only the parameter μ and has the t distribution with $n - 1$ degrees of freedom which does not involve any unknown parameters. It is therefore possible to find a number, say $t_{.05}$, such that

$$P(-t_{.05} < t < t_{.05}) = \int_{-t_{.05}}^{t_{.05}} f(t; n-1)dt = .95 \quad (5)$$

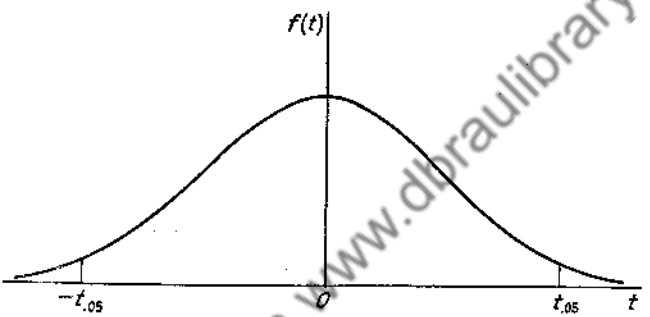


FIG. 47.

and then to convert the inequalities to obtain

$$P\left[\bar{x} - t_{.05}\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}} < \mu < \bar{x} + t_{.05}\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}}\right] = .95 \quad (6)$$

in which the limits can be computed for a given sample to obtain a 95 per cent confidence interval.

The number $t_{.05}$ is called the 5 per cent level of t and locates points which cut off 2.5 per cent of the area under $f(t)$ on each tail. Since $f(t)$ is symmetric about $t = 0$, (6) gives the minimum 95 per cent confidence interval. Other confidence intervals can be obtained by using other levels of t. Thus a 99 per cent confidence may be found by using the number $t_{.01}$, which cuts off area .005 on each tail of the t distribution.

Figure 48 shows the result of computing 50 per cent confidence intervals for 15 samples of size four actually drawn from a normal population with zero mean and unit variance. The intervals are

shown as horizontal lines above the μ axis, and, as expected, about half of them cover the true mean zero. Similarly if 95 per cent confidence intervals were used, about 95 per cent of them would be expected to cover the true mean. If one consistently uses 95 per cent confidence intervals to estimate parameters and states each time that the interval contains the true parameter value, he can expect to be wrong in 5 per cent of those statements.

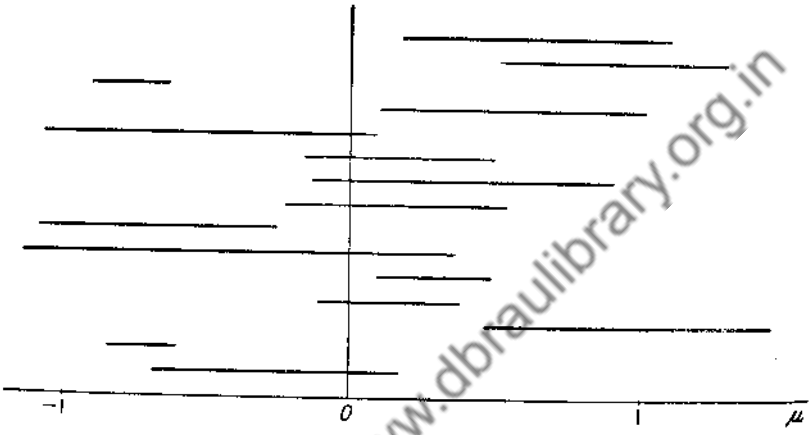


FIG. 48.

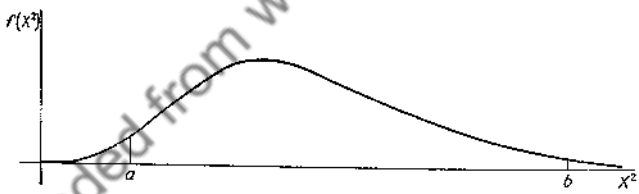


FIG. 49.

11.3. Confidence Intervals for the Variance of a Normal Distribution. For samples of size n from a normal population, the quantity

$$\chi^2 = \frac{\Sigma(x_i - \bar{x})^2}{\sigma^2} \tag{1}$$

where $\bar{x}$ is the sample mean, has the chi-square distribution with $n - 1$ degrees of freedom. Hence a confidence interval with confidence coefficient γ may be set up by finding two numbers, say a and b, such that

$$P(a < \chi^2 < b) = \int_a^b f(\chi^2)d\chi^2 = \gamma \tag{2}$$

On converting the inequalities, we obtain

$$P\left[\frac{\Sigma(x_i - \bar{x})^2}{b} < \sigma^2 < \frac{\Sigma(x_i - \bar{x})^2}{a}\right] = \gamma \tag{3}$$

which will determine a confidence interval for σ^2.

Since the length of the confidence interval is

$$\left(\frac{1}{a} - \frac{1}{b}\right)\sum (x_i - \bar{x})^2 \tag{4}$$

the shortest confidence interval for a given sample would be obtained by choosing a so as to minimize $[(1/a) - (1/b)]$ for the chosen value of γ. The required computation is so tedious that it is rarely done in practice, and tables giving the required levels have not been published. The ordinary chi-square tables give numbers χ^2 such that

$$P(\chi^2 > \chi_\epsilon^2) = \int_{\chi_\epsilon^2}^{\infty} f(\chi^2)d\chi^2 = \epsilon \tag{5}$$

for selected values of ϵ. In setting up, say, a 95 per cent confidence interval, one merely chooses $a = \chi^2_{.975}$ and $b = \chi^2_{.025}$, i.e., selects a and b so that area .025 is cut off from each tail of the distribution. This very nearly minimizes the length of the confidence interval unless the number of degrees of freedom is quite small.

11.4. Confidence Region for Mean and Variance of a Normal Distribution. In constructing a region for the joint estimation of the mean μ_0 and variance σ_0^2 of a normal distribution, one might at first sight be inclined to use the individual estimates given by the t and the χ^2 distributions. That is, for example, one might construct a .9025 ($= .95^2$) region as in Fig. 50 by using the two relations:

$$P\left[\bar{x} - t_{.05}\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}} < \mu_0 < \bar{x} + t_{.05}\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}}\right] = .95 \tag{1}$$

$$P\left[\frac{\Sigma(x_i - \bar{x})^2}{\chi^2_{.025}} < \sigma_0^2 < \frac{\Sigma(x_i - \bar{x})^2}{\chi^2_{.975}}\right] = .95 \tag{2}$$

assuming that the probability of both occurrences is the product of the separate probabilities. This is incorrect because t and χ^2 are not independently distributed. The joint probability that the two intervals cover the true parameter values is not equal to the product of the separate probabilities. Hence the probability that the rectangular region of Fig. 50 covers the true parameter point (μ_0, σ_0^2) is not .9025.

A confidence region may be set up, however, by using the distributions of $\bar{x}$ and $\Sigma(x_i - \bar{x})^2$, which are independently distributed. If, for example, a 95 per cent confidence region is desired, we may find numbers a, a', and b' such that

$$P\left(-a < \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < a\right) = \sqrt{.95} \cong .975 \tag{3}$$

$$P\left[a' < \frac{\Sigma(x_i - \bar{x})^2}{\sigma_0^2} < b'\right] = \sqrt{.95} \tag{4}$$

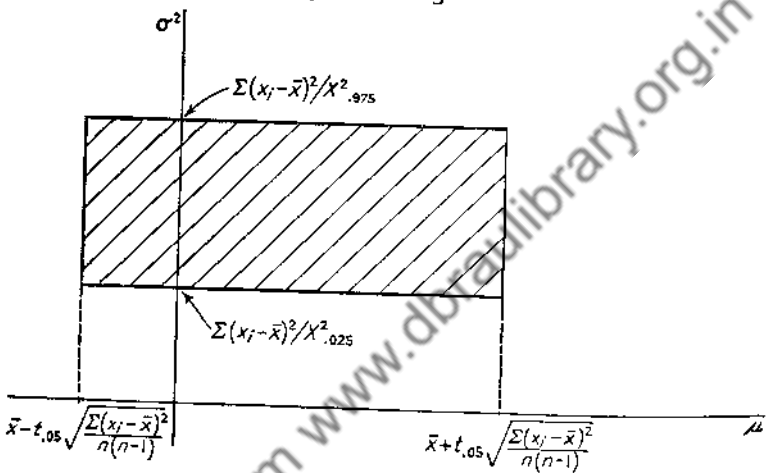


FIG. 50.

The joint probability

$$P\left[-a < \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < a,\ a' < \frac{\Sigma(x_i - \bar{x})^2}{\sigma_0^2} < b'\right] = .95 \tag{5}$$

because of the independence of the distributions. The four inequalities in (5) determine a region in the parameter space which is easily found by plotting its boundaries. One merely replaces the inequality signs by equality signs and plots each of the four resulting relations as functions of μ and σ^2 in the parameter space. A region such as the shaded area in Fig. 51 will result. A confidence region for (μ_0, σ_0) would be obtained in exactly the same way; the relations would be plotted as functions of σ instead of σ^2, and the parabola in Fig. 51 would become a pair of straight lines

$$\mu = \bar{x} \pm \frac{a\sigma}{\sqrt{n}}$$

intersecting at $\bar{x}$ on the μ axis.

The region we have constructed does not have minimum area, but it is easily constructible from existing tables and will differ but little from the region of minimum area unless the sample size is small. The minimum region is roughly elliptical in shape and difficult to construct.

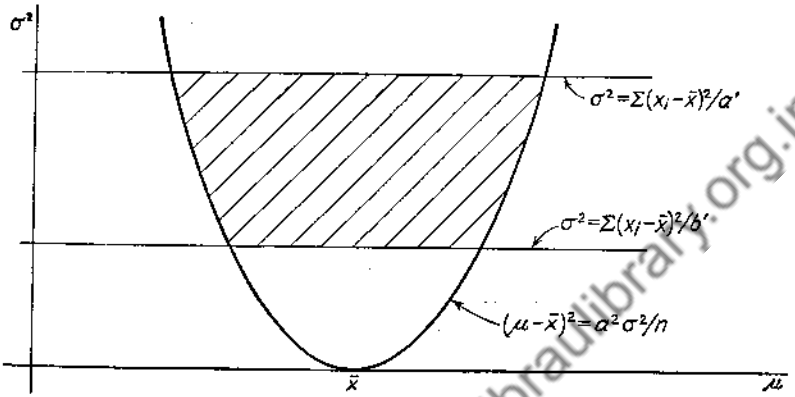


FIG. 51.

11.5. A General Method for Obtaining Confidence Intervals. The method used in the preceding sections for determining confidence intervals and regions required that functions of the sample and parameters be found which were distributed independently of the parameters. It is possible to set up confidence intervals, however, whether or not such functions exist.

Given a population with density $f(x; \theta)$ and an estimator $\hat{\theta}(x_1, x_2, \cdots, x_n)$ for samples of size n (one would ordinarily use the maximum-likelihood estimator), we may determine the density, say $g(\hat{\theta}; \theta)$, of the estimator. We shall suppose, for definiteness, that a 95 per cent confidence interval is desired. If any arbitrary number, say θ', is substituted for θ in $g(\hat{\theta}, \theta)$, the distribution of $\hat{\theta}$ will be completely specified and it will be possible to make probability statements about $\hat{\theta}$. In particular, we may find two numbers h_1 and h_2 such that

$$P(\hat{\theta} < h_1) = \int_{-\infty}^{h_1} g(\hat{\theta}; \theta')d\hat{\theta} = .025 \tag{1}$$

$$P(\hat{\theta} > h_2) = \int_{h_2}^{\infty} g(\hat{\theta}; \theta')d\hat{\theta} = .025 \tag{2}$$

The numbers h_1 and h_2 will depend, of course, on the number substituted for θ in $g(\hat{\theta}; \theta)$. In fact, we may write h_1 and h_2 as functions of θ: $h_1(\theta)$ and $h_2(\theta)$. The values of these functions for any value of θ

are determined by equations (1) and (2). Obviously

$$P[h_1(\theta) < \hat{\theta} < h_2(\theta)] = \int_{h_1(\theta)}^{h_2(\theta)} g(\hat{\theta}; \theta)d\hat{\theta} = .95 \tag{3}$$

The functions $h_1(\theta)$ and $h_2(\theta)$ may be plotted against θ as in Fig. 52. A vertical line through any chosen value θ' of θ will intersect the two curves in points which, projected on the $\hat{\theta}$ axis, will give limits between which θ will fall with probability .95.

Having constructed the two curves $\hat{\theta} = h_1(\theta)$ and $\hat{\theta} = h_2(\theta)$, we may construct a confidence interval for θ as follows: Draw a sample of size n and compute the value of the estimator, say $\hat{\theta}'$. A horizontal line

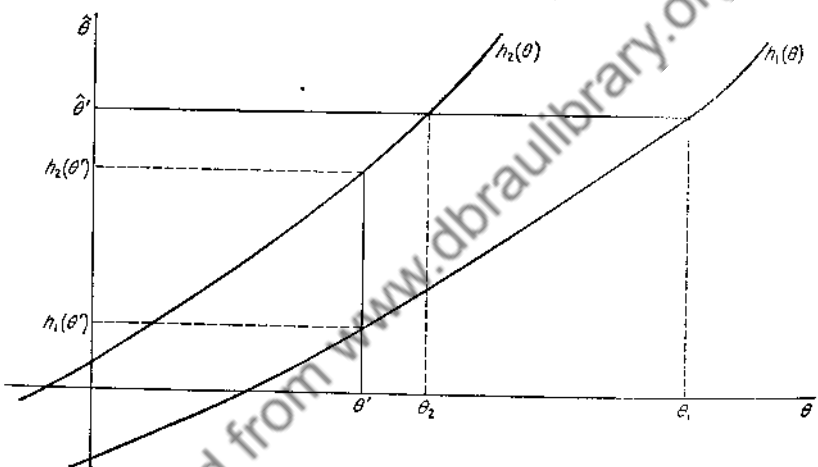


FIG. 52.

through the point $\hat{\theta}'$ on the $\hat{\theta}$ axis (Fig. 52) will intersect the two curves at points which may be projected on the θ axis and labeled θ_1 and θ_2 as in the figure. These two numbers define the confidence interval, for it is easily shown that

$$P_F(\theta_2 < \theta < \theta_1) = .95 \tag{4}$$

Suppose that we were in fact sampling from a population that had θ' as the value of θ. The probability that the estimate $\hat{\theta}$ will fall between $h_1(\theta')$ and $h_2(\theta')$ is .95. If the estimate does fall between these limits, then the horizontal line will cut the vertical line through θ' at some point between the curves and the corresponding interval (θ_2, θ_1) will cover θ'. If the estimate does not fall between $h_1(\theta')$ and $h_2(\theta')$, the horizontal line does not cut the vertical line between the curves and the corresponding interval (θ_2, θ_1) does not cover θ. It

follows, therefore, that the probability is exactly .95 that an interval (θ_2, θ_1) constructed by this method will cover θ'. And this statement is true for any population value of θ.

It is sometimes possible to determine the limits θ_2 and θ_1 for a given estimate without actually finding the functions $h_1(\theta)$ and $h_2(\theta)$. Referring to Fig. 52, the limits for θ are at points θ_2 and θ_1 such that $h_1(\theta_1) = \hat{\theta}'$ and $h_2(\theta_2) = \hat{\theta}'$. In terms of the definition of h_1 and h_2, we may say that θ_1 is the value of θ for which

$$\int_{-\infty}^{\hat{\theta}'} g(\hat{\theta};\, \theta)d\hat{\theta} = .025 \tag{5}$$

and θ_2 is the value of θ for which

$$\int_{\hat{\theta}'}^{\infty} g(\hat{\theta};\, \theta)d\hat{\theta} = .025 \tag{6}$$

If the left-hand sides of these two equations can be given explicit expressions in terms of θ, and if the equations can be solved for θ uniquely, then those roots are the 95 per cent confidence limits for θ.

If $h_1(\theta)$ and $h_2(\theta)$ are not monotonic functions of θ, the confidence interval may in fact be a set of intervals. Thus suppose the curves of Fig. 52 bent down farther to the right so that the horizontal line at $\hat{\theta}'$ cut them again, for example, at points θ_3 and θ_4. Then the confidence interval would actually consist of two intervals (θ_2, θ_1) and (θ_3, θ_4). The fiducial statement about θ would then be of the form

$$P_F(\theta_2 < \theta < \theta_1, \text{ or } \theta_3 < \theta < \theta_4) = .95 \tag{7}$$

However, in most situations encountered in practice there will be a single interval, or it will be possible to select a single interval on the basis of other evidence concerning the experiment which produced the sample observations.

The method described here for obtaining confidence intervals may be extended to the case of several parameters, but a geometrical representation becomes impossible even for two parameters. Suppose a distribution depends on two parameters θ_1 and θ_2; we may find a plane region R in the $\hat{\theta}_1$, $\hat{\theta}_2$ plane such that

$$P(\hat{\theta}_1, \hat{\theta}_2 \text{ in } R) = \iint_R g(\hat{\theta}_1, \theta_2;\, \theta_1, \theta_2)d\hat{\theta}_1\, d\hat{\theta}_2 = .95 \tag{8}$$

By considering all possible pairs of values of θ_1 and θ_2, we can generate a four-dimensional region in the θ_1, θ_2, $\hat{\theta}_1$, $\hat{\theta}_2$ space which is analogous to the two-dimensional region between the curves in Fig. 52. Now

suppose a sample is drawn and the estimates $\hat{\theta}_1'$ and $\hat{\theta}_2'$ calculated. The intersection of the two hyperplanes $\hat{\theta}_1 = \hat{\theta}_1'$ and $\hat{\theta}_2 = \hat{\theta}_2'$ with the four-dimensional region will determine a two-dimensional region, which, when projected on the θ_1, θ_2 plane, will be a 95 per cent confidence region for θ_1, θ_2.

The argument may be extended to cover the case of k parameters. The method will determine a confidence region for all the parameters of a distribution. If one wishes to estimate some but not all of a set of parameters, the method can not be used in general, though it may be modified to handle the problem in special circumstances. There is as yet no general solution to the problem of setting up confidence regions for a part of a set of k parameters in a distribution function except in the case of large samples.

Illustrative example: As a simple illustration, we may consider the estimation of α in

$$f(x; \alpha) = \frac{2}{\alpha^2}(\alpha - x) \qquad 0 < x < \alpha \tag{9}$$

for samples of size one. If x is the observation, the maximum-likelihood estimator is found to be $\hat{\alpha} = 2x$ by solving

$$\frac{\partial}{\partial \alpha}\left[\frac{2}{\alpha^2}(\alpha - x)\right] = 0$$

$\frac{2}{\alpha^2} + \frac{-2}{\alpha^3}2(\alpha - x) = 0 \quad 2\alpha - 4\alpha + 4x = 0$

for α. The distribution of the estimator is $\alpha \quad \hat{\alpha} = 2x$

$$g(\hat{\alpha}; \alpha) = \frac{1}{2\alpha^2}(2\alpha - \hat{\alpha}) \qquad 0 < \hat{\alpha} < 2\alpha \tag{10}$$

so that 95 per cent confidence intervals are obtained by determining $h_1(\alpha)$ and $h_2(\alpha)$ so that

$$\int_0^{h_1(\alpha)} g(\hat{\alpha}; \alpha)d\hat{\alpha} = .025 \tag{11}$$

$$\int_{h_2(\alpha)}^{2\alpha} g(\hat{\alpha}; \alpha)d\hat{\alpha} = .025 \tag{12}$$

The integrations are easily performed in this case and give, on solving for h_1 and h_2,

$$h_1(\alpha) = 2(1 - \sqrt{.975})\alpha \tag{13}$$

$$h_2(\alpha) = 2(1 - \sqrt{.025})\alpha \tag{14}$$

These plot as straight lines, as in Fig. 53. For a given observation, say $x = 2$, the estimate is $\hat{\alpha}' = 4$ and the 95 per cent confidence inter-

val is given by

$$P_F\left(\frac{2}{1-\sqrt{.025}} < \alpha < \frac{2}{1-\sqrt{.975}}\right) = .95 \tag{15}$$

Actually, since

$$u = \frac{2\alpha - \hat{\alpha}}{\alpha}$$

is distributed independently of α, it was not necessary to use the general method in this problem. We could have found a confidence interval for α by getting .95 limits for u and then converting the inequalities to get a statement about α.

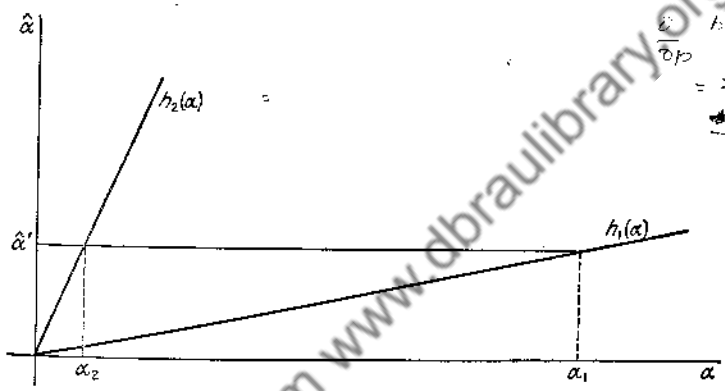


FIG. 53.

11.6. Confidence Intervals for the Parameter of a Binomial Distribution. We shall apply the general method described in the preceding section to a problem which requires its use. If a sample, $x_1, x_2, \cdots, x_n$, is drawn from a binomial population with

$$f(x; p) = p^x(1-p)^{1-x} \qquad x = 0, 1 \tag{1}$$

the maximum-likelihood estimator of p is

$$\hat{p} = \frac{y}{n} \tag{2}$$

where $y = \Sigma x_i$ can have the values $0, 1, 2, \cdots, n$. The density of $\hat{p}$ is

$$g(\hat{p}; p) = \binom{n}{n\hat{p}} p^{n\hat{p}}(1-p)^{n(1-\hat{p})} \qquad \hat{p} = 0, \frac{1}{n}, \frac{2}{n}, \cdots, 1 \tag{3}$$

and it is not possible to find a function of $\hat{p}$ and p which is distributed independently of p.

Again we shall suppose for definiteness that a 95 per cent confidence interval is to be constructed. The first step is to determine the functions $h_1(p)$ and $h_2(p)$. For $p = .4$, for example, we would, in accordance with the preceding section, seek a number $h_1(.4)$ such that

$$P[\hat{p} < h_1(.4)] = \sum_{y=0}^{nh_1} \binom{n}{y} (.4)^y(.6)^{n-y} = .025 \tag{4}$$

However, in view of the discreteness of the distribution, nh_1 in the sum must be an integer, and it will be impossible to make the sum exactly .025 for every value of p. This need not worry us though. We do not need a curve $h_1(p)$ defined at every p. The only points of interest are those which correspond to the possible values of $\hat{p}$. It is, in fact, possible to use the technique indicated by equations (5.5) and (5.6) of the preceding section, because an explicit expression for the probabilities on the left of these equations is immediately at hand. Assuming we have an estimate

$$\hat{p}' = \frac{k}{n} \tag{5}$$

the 95 per cent confidence upper limit p_1 may be determined by finding the value of p for which

$$\sum_{y=0}^{k} \binom{n}{y} p^y(1 - p)^{n-y} = .025 \tag{6}$$

and the lower limit p_2 is the value of p for which

$$\sum_{y=k}^{n} \binom{n}{y} p^y(1 - p)^{n-y} = .025 \tag{7}$$

If k is zero, the lower limit is taken to be zero, and if $k = n$, the upper limit is taken to be one.

For small values of n, equations (6) and (7) may be solved by trial and error for the roots p_1 and p_2, but this computation rapidly becomes tedious with increasing n. A simple method of solution is provided by Pearson's tables of the incomplete beta function. The cumulative form of the beta distribution is

$$F(x; \alpha, \beta) = \frac{(\alpha + \beta + 1)!}{\alpha!\beta!} \int_0^x t^\alpha(1 - t)^\beta \, dt \tag{8}$$

and repeated integration by parts gives

$$F(x;\alpha,\beta) = -\sum_{i=0}^{\alpha}\binom{\alpha+\beta+1}{i}x^i(1-x)^{\alpha+\beta+1-i}+1 \tag{9}$$

It follows that partial binomial sums are given by the table of $F(x;\alpha,\beta)$. We may write equation (6) as

$$\sum_{y=0}^{k}\binom{n}{y}p^y(1-p)^{n-y} = 1 - F(p;k,n-k-1) = .025 \tag{10}$$

and find at once in the table the value of p which corresponds to $F = .975$ for the given values of k and $n-k-1$. Similarly, since

$$\sum_{k}^{n}\binom{n}{y}p^y(1-p)^{n-y} = 1 - \sum_{0}^{k-1}\binom{n}{y}p^y(1-p)^{n-y}$$

we may find the lower confidence limit by putting (7) in the form

$$\sum_{k}^{n}\binom{n}{y}p^y(1-p)^{n-y} = F(p;k-1,n-k) = .025 \tag{11}$$

For values of n beyond the range of the table, the normal approximation to the binomial distribution may be used to obtain confidence intervals for p, as is shown in the following section.

11.7. Confidence Intervals for Large Samples. We have seen in Chap. 10 that for large samples, the maximum-likelihood estimator $\hat{\theta}$ for a parameter θ in a density $f(x;\theta)$ is approximately normally distributed about θ under rather general conditions. When these conditions are satisfied, it is possible to obtain approximate confidence intervals quite easily. The large-sample variance of the estimator is, say,

$$\sigma^2(\theta) = \frac{-1}{nE\left[\partial^2 \log f(x;\theta)/\partial\theta^2\right]} \tag{1}$$

and we have indicated that it is a function of θ since it ordinarily will depend on θ. For large samples, therefore, a confidence interval with fiducial probability γ may be determined by converting the inequalities in

$$P\left[-d_\gamma < \frac{\hat{\theta}-\theta}{\sigma(\theta)} < d_\gamma\right] \cong \gamma \tag{2}$$

where d_γ is chosen so that

$$\int_{-d_\gamma}^{d_\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt = \gamma$$

As an example, we may consider the binomial distribution with parameter p; the variance of $\hat{p}$ is

$$\sigma^2(p) = \frac{p(1 - p)}{n} \tag{3}$$

An approximate γ confidence interval, for example, is obtained by converting the inequalities in

$$P\left[-d_\gamma < \frac{\hat{p} - p}{\sqrt{p(1 - p)/n}} < d_\gamma\right] \cong \gamma \tag{4}$$

to get

$$P\left[\frac{2n\hat{p} + d_\gamma^2 - d_\gamma\sqrt{4n\hat{p} + d_\gamma^2 - 4n\hat{p}^2}}{2(n + d_\gamma^2)} < p \right.$$
$$\left. < \frac{2n\hat{p} + d_\gamma^2 + d_\gamma\sqrt{4n\hat{p} + d_\gamma^2 - 4n\hat{p}^2}}{2(n + d_\gamma^2)}\right] \cong \gamma \tag{5}$$

These expressions for the limits may be simplified if we recall that in deriving the large-sample distribution, we neglect certain terms containing the factor $1/\sqrt{n}$; i.e., the asymptotic normal distribution is correct only to within error terms of size $k/\sqrt{n}$. We may therefore neglect terms of this order in the limits in (5) without affecting the accuracy of the approximation. This means simply that we may omit all the d_γ^2 in (5), because they always occur added to a term with factor n and will be negligible, relative to n when n is large, to within the degree of approximation we are assuming. Thus (5) may be rewritten as

$$P\left[\hat{p} - d_\gamma\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + d_\gamma\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right] \cong \gamma \tag{6}$$

In particular,

$$P\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right. \cong .95$$

gives an approximate 95 per cent confidence interval for p for large samples.

We may observe that (6) is just the expression that would have been obtained had $\hat{p}$ been substituted for p in $\sigma^2(p)$. This substitution

would imply that

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

is approximately normally distributed with zero mean and unit variance. It is, in fact, true in general that in the asymptotic normal distribution of a maximum-likelihood estimator $\hat{\theta}$, the variance $\sigma^2(\theta)$ may be replaced by its estimator $\sigma^2(\hat{\theta})$ without appreciably affecting the accuracy of the approximation. We shall not prove this fact but shall use it because it greatly simplifies the conversion of inequalities in a probability statement to get confidence intervals.

For large samples, therefore, an approximate confidence interval with confidence coefficient γ is given by

$$P(\hat{\theta} - d_\gamma\sigma(\hat{\theta}) < \theta < \hat{\theta} + d_\gamma\sigma(\hat{\theta})) \cong \gamma \tag{7}$$

when $\hat{\theta}$ is asymptotically normally distributed, and $\sigma(\hat{\theta})$ in this expression is the maximum-likelihood estimate of the standard deviation of $\hat{\theta}$.

11.8. Confidence Regions for Large Samples. When a distribution involves several parameters $(\theta_1, \theta_2, \cdots, \theta_k)$, we have seen in Chap. 10 that under rather general conditions the large-sample maximum-likelihood estimates, $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$, are approximately normally distributed with means $(\theta_1, \theta_2, \cdots, \theta_k)$ and coefficients of the quadratic form given by

$$\|\sigma^{ij}(\theta_1, \cdots, \theta_k)\| = \left\| -nE\left[\frac{\partial^2 \log f(x; \theta_1, \theta_2, \cdots, \theta_k)}{\partial\theta_i\,\partial\theta_j}\right]\right\| \tag{1}$$

The coefficients will, in general, be functions of the θ_i as we have indicated.

Now we have seen that the quadratic form of a k-variate normal distribution has the chi-square distribution with k degrees of freedom. We may conclude, therefore, that the quantity

$$u = \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma^{ij}(\theta_1, \cdots, \theta_k)(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) \tag{2}$$

is approximately distributed by the chi-square distribution with k degrees of freedom for large samples. Here again, the accuracy of the approximation is not impaired by substituting the estimates of the θ_i for the θ_i in $\sigma_{ij}(\theta_1, \cdots, \theta_n)$; the quantity

$$v = \Sigma\Sigma\sigma^{ij}(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) \tag{3}$$

is also approximately distributed by the chi-square law with k degrees of freedom.

The variate v enables us to set up a very simple confidence region for the θ_i. If $\chi^2_{1-\gamma}$ is the $1 - \gamma$ level of the chi-square distribution, then

$$P(v < \chi^2_{1-\gamma}) = \gamma \tag{4}$$

determines a confidence region in the parameter space. The boundary of the region is given by the equation

$$\Sigma\Sigma\sigma^{ij}(\hat{\theta}_1, \cdots, \hat{\theta}_k)(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) = \chi^2_{1-\gamma} \tag{5}$$

which is the equation of an ellipsoid in the $(\theta_1, \theta_2, \cdots, \theta_k)$ space with its center at $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$.

If one is interested in estimating only a part of a set of k parameters, for example, the set $(\theta_1, \theta_2, \cdots, \theta_r)$ where $r < k$, we first find the marginal distribution of the maximum-likelihood estimators for this set. If we let (a, b) be indices which have the range $1, 2, \cdots, r$, then the coefficients $\bar{\sigma}^{ab}$ of the quadratic form of the large-sample normal distribution of $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r$ are given by

$$\|\bar{\sigma}^{ab}\| = \|\sigma_{ab}\|^{-1}$$

where the matrix $\|\sigma_{ab}\|$ is obtained by striking out the last $k - r$ rows and columns in $\|\sigma_{ij}\|$. The $\bar{\sigma}^{ab}$ will, in general, be functions of all k of the original parameters $\theta_1, \theta_2, \cdots, \theta_k$. If we substitute the $\hat{\theta}_i$ for the θ_i in $\bar{\sigma}^{ab}$, we shall obtain the maximum-likelihood estimators $\hat{\bar{\sigma}}^{ab}$ of the $\bar{\sigma}^{ab}$. The quadratic form

$$w = \sum_a \sum_b \hat{\bar{\sigma}}^{ab}(\hat{\theta}_a - \theta_a)(\hat{\theta}_b - \theta_b)$$

is approximately distributed like chi square with r degrees of freedom and will serve to determine an ellipsoidal confidence region in the $\theta_1, \theta_2, \cdots, \theta_r$ space for those parameters.

As an example of the estimation of more than one parameter, we may consider the large-sample estimation of the mean and variance of a normal population. We have seen in Sec. 10.9 that $\bar{x}$ and $\hat{\sigma}^2$ are approximately distributed with means μ and σ^2 and with coefficients of the quadratic form

$$\|\sigma^{ij}(\mu, \sigma^2)\| = \begin{Vmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{Vmatrix} \tag{6}$$

If we substitute $\hat{\sigma}^2$ for σ^2 in (6), then the quadratic form becomes

$$v = \frac{n}{\hat{\sigma}^2}(\bar{x} - \mu)^2 + \frac{n}{2\hat{\sigma}^4}(\hat{\sigma}^2 - \sigma^2)^2 \tag{7}$$

which is approximately distributed like chi square with two degrees of freedom for large samples. In particular, let us suppose that we have an actual sample of 100 observations (3.4, 5.1, $\cdots$, 2.2) with

$$\bar{x} = \tfrac{1}{100}\Sigma x_i = 4$$
$$\hat{\sigma}^2 = \tfrac{1}{100}\Sigma(x_i - \bar{x})^2 = 5$$

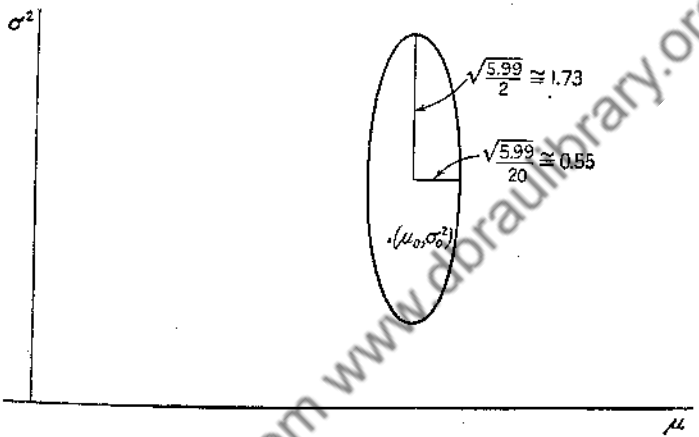


FIG. 54.

since the .05 level of chi square with two degrees of freedom is 5.99, a 95 per cent confidence region for μ and σ^2 is determined by

$$P_{\bar{x}}[20(4 - \mu)^2 + 2(5 - \sigma^2)^2 < 5.99)] = .95 \tag{8}$$

The values of μ and σ^2 which satisfy the inequality in (8) are the points within the ellipse

$$20(4 - \mu)^2 + 2(5 - \sigma^2)^2 = 5.99$$

which is plotted in Fig. 54. This is the 95 per cent confidence region for the true parameter point, say (μ_0, σ_0^2). Before the sample was drawn, the probability was about .95 that the region we were going to construct would cover the true parameter point.

The large-sample confidence intervals and regions presented in this and the preceding section have an optimum property which we shall point out but not prove. In the earlier sections of the chapter, we were concerned with finding the shortest interval for a given fiducial

probability. Thus the shortest 95 per cent interval for the mean of a normal population when σ is known is given by

$$P\left(\bar{x} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right) = .95$$

and the length of the interval is $2 \times 1.96\sigma/\sqrt{n}$, where n is the sample size. Now let us suppose that, instead of using $\bar{x} = (1/n)\Sigma x_i$ to construct the confidence interval, we used only one of the observations, say the first. The estimator is simply

$$\tilde{\mu} = x_1$$

and the confidence interval is given by

$$P(\tilde{\mu} - 1.96\sigma < \mu < \tilde{\mu} + 1.96\sigma) = .95$$

which has length $2 \times 1.96\sigma$. This interval is $\sqrt{n}$ times as long as the one obtained by using the sample mean as the estimator.

It is now evident that the length of a confidence interval for a parameter depends strongly on what function of the sample observations is chosen as an estimator. The optimum property of the large-sample intervals and regions based on maximum-likelihood estimators is this:

Large-sample confidence intervals and regions based on maximum-likelihood estimators will be smaller on the average than intervals and regions determined by any other estimators of the parameters.

This property of maximum-likelihood estimators is closely related to the fact that they are efficient, i.e., that they have smaller variance in large samples than other estimators. By "other estimators" we mean functionally different estimators; one would obtain essentially the same confidence regions by using estimators which were functions of the maximum-likelihood estimators. The phrase "on the average" refers to the fact that confidence regions usually vary in size from sample to sample (see Fig. 48), and for a given sample a region determined by some other estimators may be smaller than the region determined by the maximum-likelihood estimators. But for repeated sampling, the average size of the regions determined by maximum-likelihood estimators will be smaller than the average size of regions determined by other estimators.

11.9. Problems

1. Find a 90 per cent confidence interval for the mean of a normal distribution with $\sigma = 3$, given the sample (2.3, −.2, −.4, −.9). What would be the confidence interval if σ were unknown?

2. The breaking strengths in pounds of five specimens of manilla rope of diameter $\frac{3}{16}$ inch were found to be 560, 480, 540, 570, 540. Estimate the mean breaking strength by a 95 per cent confidence interval, assuming normality. Estimate the point at which only 5 per cent of such specimens would be expected to break.

3. Referring to Prob. 2, estimate σ^2 by a 90 per cent confidence interval; also σ.

4. Referring to Prob. 2, plot an 81 per cent confidence region for the joint estimation of μ and σ^2; for μ and σ.

5. Five samples were drawn from populations assumed to be normal and assumed to have the same variance. The values of $s^2 = \Sigma(x_i - \bar{x})^2$ and n, the sample size, were

s^2:	40	22	17	42	45
n:	6	4	3	7	8

Find 98 per cent confidence limits for the common variance.

6. The largest observation x' of a sample of n from a rectangular density $f(x) = 1/\theta$ $(0 < x < \theta)$ has the density

$$f(x') = \frac{n(x')^{n-1}}{\theta^n} \qquad 0 < x' < \theta$$

Show that $u = x'/\theta$ is distributed independently of θ. Using u, find the shortest confidence interval for θ for fiducial probability γ.

7. Compute a 95 per cent confidence interval for the range of a rectangular distribution given the sample (2.6, 1.2, 4.3, 1.6), and given that the lower limit of the range is zero.

8. To test two promising new lines of hybrid corn under normal farming conditions, a seed company selected eight farms at random in Iowa and planted both lines in experimental plots on each farm. The yields (converted to bushels per acre) for the eight locations were:

Line A:	86	87	56	93	84	93	75	79
Line B:	80	79	58	91	77	82	74	66

Assuming the two yields are jointly normally distributed, estimate the difference between the mean yields by a 95 per cent confidence interval. (Refer to Prob. 22 of Chap. 10.)

9. Using the density

$$f(x) = \frac{4x^3}{\theta^4} \qquad 0 < x < \theta$$

for the largest of four observations from a rectangular population, set up a general system of 95 per cent confidence intervals for θ by finding

the functions $h_1(\theta)$ and $h_2(\theta)$ and plotting these in the $(\hat{\theta}, \theta)$ plane. Find the interval for the sample given in Prob. 7. Why does it differ from the interval found in that problem?

10. Referring to Prob. 9, plot the functions $h_1(\theta)$ and $h_2(\theta)$ for samples of size eight. Then show in general that the lengths of the intervals decrease as the sample size n increases.

11. The sample (2.3, 1.2, 0.9, 3.2) was drawn from a population distributed by $f(x) = \alpha e^{-\alpha x}$, $x > 0$. Find a 90 per cent confidence interval for α.

12. Referring to Prob. 11, find 90 per cent confidence intervals for the mean and for the variance of the distribution. What is the fiducial probability that both these intervals cover the true mean and true variance, respectively?

13. One head and two tails resulted when a coin was tossed three times. Find a 90 per cent confidence interval for the probability of a head.

14. 160 heads and 240 tails resulted from 400 tosses of a coin. Find a 90 per cent confidence interval for the probability of a head. Find a 99 per cent confidence interval. Does this appear to be a true coin?

15. A sample of 2000 voters were asked their attitude toward a certain political proposal. 1200 favored the proposal; 600 opposed it; and 200 were undecided. Assuming this was a random sample from a trinomial population, construct a 95 per cent confidence region for p_1 and p_2, the proportions of individuals for and against the proposal. (Use the results of Sec. 10.9.)

16. Plot a 95 per cent confidence region like that of Fig. 51 for the example used in Sec. 8 and compare it with the region of Fig. 54.

17. Integrate by parts [integrating $(1 - t)^s$ and differentiating t^r] to show

$$\int_0^x t^r(1 - t)^s\, dt = -\frac{1}{s + 1} x^r(1 - x)^{s+1} + \frac{r}{s + 1} \int_0^x t^{r-1}(1 - t)^{s+1}\, dt$$

18. Apply the above result repeatedly to obtain a cumulative form for the beta distribution, $F(x; \alpha, \beta)$.

19. Show that

$$F(x; \alpha, \beta) = \sum_{i=\alpha+1}^{\alpha+\beta+1} \binom{\alpha + \beta + 1}{i} x^i(1 - x)^{\alpha-\beta+1+i}$$

by using the result of Prob. 18. This is the form that would have arisen had the integration by parts been done the other way—differentiating $(1 - t)^s$ and integrating t^r.

20. Given a sample of size 100 from a normal population with $\hat{\mu} = 3$, $\hat{\sigma}^2 = .25$, what is the maximum-likelihood estimate of the number α for which

$$\int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2\sigma^2)(x-\mu)^2}\,dx = .05$$

21. Find the large-sample distribution of $\hat{\mu}$ and $\hat{\sigma}$ for samples from a normal population. Since it is known that $\hat{\mu}$ and $\hat{\sigma}$ will be normally and independently distributed with means μ and σ, it is only necessary to find their variance.

22. Referring to the above problem, find the large-sample distribution of $\hat{\mu} + k\hat{\sigma}$ where k is a given constant. Use this to obtain a 95 per cent confidence interval for α in Prob. 20.

23. Develop a method for estimating the ratio of the variances of two normal populations by a confidence interval.

24. Develop a method for estimating the parameter of the Poisson distribution by a confidence interval. (Refer to Prob. 33 of Chap. 6.)

25. Work through the details of the derivation of equation (2.6).

26. What is the probability that the length of a t confidence interval will be less than σ for samples of size 20?

27. Compare the average length of a 95 per cent confidence interval for the mean of a normal population based on the t distribution with the length that the interval would have were the variance known.

28. Show that the length and the variance of the length of the t confidence interval approach zero with increasing sample size.

29. How large a sample must be drawn from a normal population to make the probability .95 that a 90 per cent confidence interval (based on t) for the mean will have length less than $\sigma/5$?

30. Show that the length of the confidence interval for σ (of a normal population) approaches zero with increasing sample size.

31. Consider a *truncated* normal population with density

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma\alpha} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \qquad x < a$$

where

$$\alpha = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}\,dx$$

Show that $\frac{\partial}{\partial\mu} \log f(x)$ and $\frac{\partial}{\partial\sigma} \log f(x)$ have zero expectations.

32. Referring to Prob. 31, let $\hat{\mu}$ and $\hat{\sigma}$ be maximum-likelihood estimators of μ and σ. Show that the matrix of coefficients of the quad-

ratic form for the large-sample distribution of $\hat{\mu}$ and $\hat{\sigma}$ is

$$\|\sigma_{ij}\| = \begin{Vmatrix} \dfrac{n(1 - tb - b^2)}{\sigma^2} & \dfrac{-nb(1 + tb + t^2)}{\sigma^2} \\ \dfrac{-nb(1 + tb + t^2)}{\sigma^2} & \dfrac{n(2 - tb - t^2b^2 - t^3b)}{\sigma^2} \end{Vmatrix}$$

where $b = \sigma f(a)$, and where $t = (a - \mu)/\sigma$.

CHAPTER 12

TESTS OF HYPOTHESES

12.1. Introduction. There are two major areas of statistical inference—the estimation of parameters and the testing of hypotheses. We shall study the second of these two areas in this chapter. Our aim will be to develop general methods for testing hypotheses and to apply those methods to some common problems. The methods will be of further use in later chapters.

In experimental research, the object is sometimes merely to estimate parameter. Thus one may wish to estimate the yield of a new hybrid line of corn. But more often the ultimate purpose will involve some use of the estimate. One may wish, for example, to compare the yield of the new line with that of a standard line and perhaps recommend that the new line replace the standard line if it appears superior. This is a common situation in research. One may wish to determine whether a new method of sealing light bulbs will increase the life of the bulbs, whether a new germicide is more effective in treating a certain infection than a standard germicide, whether one method of preserving foods is better than another in so far as retention of vitamins is concerned, and so on.

Using the light-bulb example as an illustration, let us suppose that the average life of bulbs made under a standard manufacturing procedure is 1400 hours. It is desired to test a new procedure for manufacturing the bulbs. The statistical model here is this: We are dealing with two populations of light bulbs—those made by the standard process and those made by the proposed process. We know (from numerous past investigations) that the mean of the first population is about 1400. The question is whether the mean of the second population is greater than or less than 1400. To answer this question, we set up a *null hypothesis*, namely, the hypothesis that the two means are the same. On the basis of a sample from the second population we shall either accept or reject the *null hypothesis*. (Naturally we hope that the new process is better and that the *null hypothesis* will be rejected.) The reason for this roundabout way of doing things will become apparent later.

To test the null hypothesis, a number of bulbs are made by the new process and their lives measured. Suppose the mean of this sample of observations is 1550 hours. The indication is that the new process is better, but suppose the estimate of the standard deviation of the mean $\hat{\sigma}/\sqrt{n}$ is 125 (n being the sample size). Then a 95 per cent confidence interval for the mean of the second population (assuming normality) is roughly 1300 to 1800 hours. The sample mean 1550 could very easily have come from a population with mean 1400. We have no strong grounds for rejecting the null hypothesis. If, on the other hand, $\hat{\sigma}/\sqrt{n}$ were 25, then we could very confidently reject the null hypothesis and pronounce the proposed manufacturing process to be superior.

The testing of hypotheses is seen to be closely related to the problem of estimation. It will be instructive, however, to develop the theory of testing independently of the theory of estimation, at least in the beginning.

12.2. Test of a Hypothesis against a Single Alternative. In the example considered above, there were many alternatives to the null hypothesis; the mean of the second population could have been any positive number within a fairly wide range. To introduce the basic notions of testing hypotheses, we shall consider the very simple case of one alternative. Suppose it is known that a population has either the density $f_0(x)$ or the density $f_1(x)$, and suppose it is desired to test on the basis of one observation whether the true density is $f_0(x)$ or $f_1(x)$. Let us designate by

$$H_0\text{: the hypothesis that } f(x) = f_0(x)$$

and by

$$H_1\text{: the alternative hypothesis that } f(x) = f_1(x)$$

We shall call H_0 the null hypothesis; rejection of H_0 will be equivalent to acceptance of H_1.

To test H_0, we shall choose a number A (see Fig. 55) and make an observation x_1. If $x_1 < A$, we shall accept H_0; if $x_1 > A$, we shall reject H_0.

There are two kinds of error possible in this test. We may reject H_0 when it is in fact true; i.e., the population may have $f_0(x)$ as its distribution even though the observed x did exceed A. This is called the *Type* I *error* of the test, and for the example of Fig. 55 its probability is obviously

$$\int_A^\infty f_0(x)dx$$

This probability is often called the *significance level* of the test. A second possible error is the acceptance of H_0 when it is false; i.e., the observation may be less than A even though the true population distribution is $f_1(x)$. This is called the *Type II error* of the test, and in the example we are considering its probability is

$$\int_{-\infty}^{A} f_1(x)dx$$

The interval $x < A$ is called the *acceptance region* for the null hypothesis, and the interval $x > A$ is called the *rejection region*, or more often the *critical region*. The construction of a test is nothing more than a matter of dividing the x axis into two regions, and this

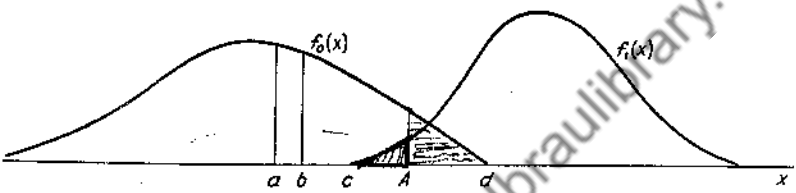


FIG. 55.

may be done quite arbitrarily. We might set up a test as follows (see Fig. 55):

Accept H_0 if $x < a$ or $x > b$
Reject H_0 if $a < x < b$

This is clearly a poorer test than the one described first. We may make the two tests comparable on one score by making the probabilities of their Type I errors the same, say .05; i.e., A may be chosen so that

$$\int_{A}^{\infty} f_0(x)dx = .05$$

and a and b chosen so that

$$\int_{a}^{b} f_0(x)dx = .05$$

The superiority of the first test is then apparent in the Type II errors,

$$\int_{-\infty}^{A} f_1(x)dx < \int_{-\infty}^{a} f_1(x)dx + \int_{b}^{\infty} f_1(x)dx$$

The second test is much more likely to accept H_0 when it is false.

A good test is clearly one which makes the probabilities of both errors as small as possible. However, it is impossible to reduce both errors simultaneously with a single observation. The common procedure is

to fix the Type I error arbitrarily (make it have probability .05, for example) and then choose the critical region so as to minimize the probability of a Type II error. The quantity

$$1 - \text{probability of a Type II error}$$

is called the *power* of the test. The power of the first test (based on the intervals $x < A$ and $x > A$) we have described is

$$1 - \int_{-\infty}^{A} f_1(x)dx = \int_{A}^{\infty} f_1(x)dx$$

In terms of this concept, the principle for setting up a test is to fix the probability of a Type I error and then choose a critical region so as to maximize the power of the test.

Returning to the example of Fig. 55, we can now set up the best test of the null hypothesis for given size of the Type I error. Suppose we wish the Type I error to have probability .05. Our problem is to divide the x axis into two regions (two intervals or two collections of intervals), one of which will be the acceptance region and the other the critical region. We may concentrate on the critical region, and having selected it, the remainder of the axis will be the acceptance region. The critical region is to be such that the area under $f_0(x)$ over the critical region is .05, and such that the power will be maximized, i.e., such that the area under $f_1(x)$ will be as large as possible over the critical region.

Certainly the critical region will include every x to the right of $x = d$, the upper limit of the range of $f_0(x)$. We can include still more of the area under $f_1(x)$ so long as we do not make the area under $f_0(x)$ exceed .05. The best values of x to choose are obviously those for which $f_1(x)$ is as large as possible relative to $f_0(x)$. We want $f_1(x)$ to be large so that the area under $f_1(x)$ will be large, and we want $f_0(x)$ to be small so that as much of the area under $f_1(x)$ can be included as possible without taking in more than .05 of the area under $f_0(x)$. The best critical region is clearly the interval $x > A$ where A is chosen so that

$$\int_{A}^{\infty} f_0(x)dx = .05$$

Other best tests would be determined by changing the specification of the probability of the Type I error. In the present illustration, for example, the Type I error could be made zero, and the best critical region would be $x > d$. This is the test one would make if he were particularly anxious to avoid rejecting H_0 when it was true, but was

not greatly concerned about rejecting H_1 when it was true. To refer back to the light-bulb illustration, H_0 may refer to the standard manufacturing process. One would not want to go to the expense of changing the process unless he was rather certain the new process was superior. Of course, such a decision as this would not ordinarily be based on one observation in practice.

The general method of setting up critical regions in the case of one alternative is quite simple. Suppose we are testing H_0 against H_1 as before. The inequality

$$\frac{f_1(x)}{f_0(x)} > k \tag{1}$$

where k is an arbitrarily chosen number, will be satisfied by certain values of x. These values of x form a critical region for a best test, the test for which the Type I error is given by integrating $f_0(x)$ over the region. Thus in Fig. 55 if we choose

$$k = \frac{f_1(A)}{f_0(A)}$$

the set of values of x for which (1) is satisfied is just the set $x > A$. By reducing k, we would get another set of x values, $x > A'$, where A' would be some number to the left of A. The test would be more powerful (would have greater probability of accepting H_1 when it was true) but would have larger probability of a Type I error. By changing k, the probability of a Type I error may be made to have any desired value. A general criterion for constructing tests may be stated thus:

To set up a best test for a given probability α of a Type I error, one chooses as a critical region the set R of points x such that:

$$f_1(x) > kf_0(x)$$

where k is selected so that:

$$\int_R f_0(x)dx = \alpha$$

This criterion refers to a test for a single alternative H_1 and a single observation. It is almost obvious that the given method of choosing R will maximize the power of the test. A formal proof would go somewhat as follows: Consider the possibility of replacing a small interval $\Delta x'$ about a point x' in R by an interval $\Delta x''$ about a point x'' not in R. (We may think of R as the interval $x > A$ in Fig. 55.) Let the length of $\Delta x''$ be so chosen that the probability of the Type I error is

unchanged by the replacement, i.e., so that approximately

$$f_0(x'')\Delta x'' = f_0(x')\Delta x'$$

The substitution will decrease the power by about $f_1(x')\Delta x'$ and increase it by about $f_1(x'')\Delta x''$. Since x' is in R,

$$f_1(x')\Delta x' > kf_0(x')\Delta x'$$

and since x'' is not in R,

$$f_1(x'')\Delta x'' \leq kf_0(x'')\Delta x''$$

The right-hand sides of these last two expressions are equal, however; hence

$$f_1(x'')\Delta x'' < f_1(x')\Delta x'$$

and any such replacement would necessarily reduce the power of the test.

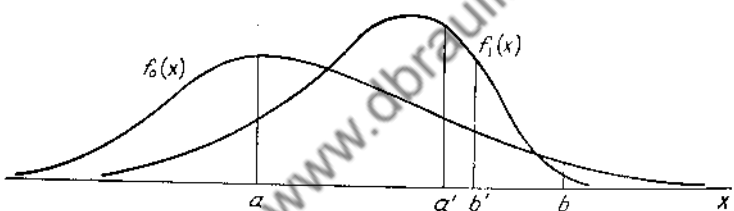


FIG. 56.

To illustrate the method further, we may consider the situation in Fig. 56. A critical region for $k = \frac{1}{2}$ is given by the interval

$$a < x < b$$

The corresponding acceptance region is, of course, the pair of intervals $x < a$ and $x > b$. The test has fairly high power in that H_0 will often be rejected when H_1 is true, but its Type I error is large. If we choose a test with small probability, say .05, of a Type I error, then the critical region would become $a' < x < b'$, and the null hypothesis would be accepted 95 per cent of the time when it was true. But now the power of the test is small; H_0 will not often be rejected when it is false, i.e., when H_1 is true. The power is, however, as large as it can be made for the given probability of a Type I error. This situation can be improved by taking more observations; we have been considering only tests based upon a single observation.

When a test is to be based on a sample of several observations, the construction is essentially the same as that we have already examined.

Suppose a sample of two observations is to be used to test H_0 against H_1. The sample density is

$$f(x_1)f(x_2)$$

defined over the x_1, x_2 plane. A test is defined by selecting a critical region R in the plane, accepting H_0 if the sample point (x_1, x_2) falls outside R, and rejecting H_0 if the sample point falls inside R. Here again the best test is given by selecting R to be the set of points (x_1, x_2) such that

$$\frac{f_1(x_1)f_1(x_2)}{f_0(x_1)f_0(x_2)} > k$$

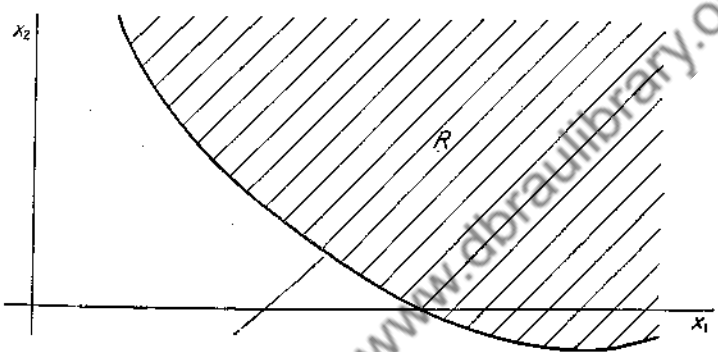


FIG. 57.

The probability of a Type I error is

$$\iint_R f_0(x_1)f_0(x_2)dx_1\,dx_2$$

and for that probability the power of the test

$$\iint_R f_1(x_1)f_1(x_2)dx_1\,dx_2$$

is maximized.

The generalization to samples of size n is immediate. The sample observations $(x_1, x_2, \cdots, x_n)$ may be plotted as a point in an n-dimensional space. The space is divided into two regions—the critical region R and the acceptance region. If the sample point falls in R, H_0 is rejected; otherwise H_0 is accepted. The best critical region R will consist of those points $(x_1, x_2, \cdots, x_n)$ in the n-dimensional space for which the *likelihood ratio*

$$\frac{f_1(x_1)f_1(x_2)\cdots f_1(x_n)}{f_0(x_1)f_0(x_2)\cdots f_0(x_n)}$$

exceeds some number k, and k is so chosen that the test has the desired probability of a Type I error. This probability is, of course,

$$\iint_R \cdots \int f_0(x_1)f_0(x_2) \cdots f_0(x_n)dx_1\, dx_2 \cdots dx_n$$

We shall not actually have to deal with n-dimensional spaces because we shall be concerned with tests of parameter values and such tests can often be based on the distribution of an estimator of the parameter.

12.3. Tests for Several Alternative Hypotheses. A common problem in testing hypotheses is that of testing a particular parameter value, say θ_0, against a set of other values of θ for a family of distributions $f(x;\ \theta)$. The basic ideas may be illustrated by a particular example. Suppose a population is known to have a normal distribution with $\sigma^2 = 1$, and suppose it is further known that the mean μ is

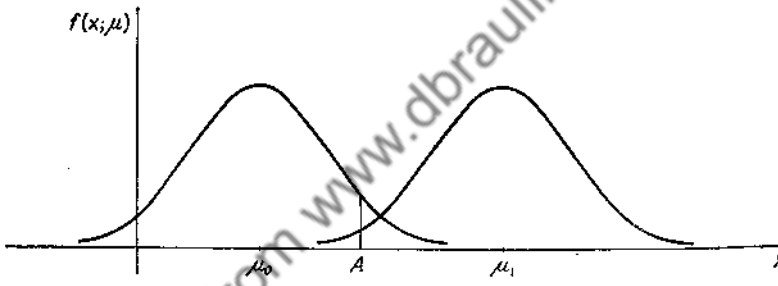


FIG. 58.

greater than or equal to some given number μ_0. On the basis of an observation x, we shall test the null hypothesis,

$$H_0\colon \mu = \mu_0 \tag{1}$$

The alternatives to this hypothesis are all the values $\mu > \mu_0$. On the basis of an observation x, we shall accept H_0 (state that $\mu = \mu_0$) or reject H_0 (state that $\mu > \mu_0$). We shall require a test for which the probability of a Type I error is, say, .05.

If a particular value μ' of μ is considered, the best test of μ_0 against that value is given by choosing as a critical region the set of points for which

$$f(x;\ \mu') > kf(x;\ \mu_0) \tag{2}$$

or, using the specific form of the distribution,

$$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu')^2} > k\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu_0)^2} \tag{3}$$

After canceling $1/\sqrt{2\pi}$ and taking logarithms, this inequality may be put in the form

$$x > \frac{2 \log k + \mu'^2 - \mu_0^2}{2(\mu' - \mu_0)} \tag{4}$$

The best critical region is, therefore, an interval $x > A$, and A is to be chosen so that

$$\int_A^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_0)^2}\, dx = .05 \tag{5}$$

The value of A is determined from the normal tables to be

$$A = \mu_0 + 1.65$$

in the present example. It was, of course, to be expected that the critical region would be of the form $x > A$.

An important thing to observe here is that the critical region is independent of the selected value μ'. Any value of μ greater than μ_0 would have given rise to exactly the same critical region. For we should have found that the best critical region was of the form $x > A$ regardless of the value given μ', and the determination of the value of A depends only on μ_0 and the selected probability of a Type I error.

We shall see later that this is not a general situation. It is not in general true that the inequality

$$f(x; \theta) > kf(x; \theta_0)$$

will give rise to the same critical region for all possible values of θ alternative to a value θ_0 specified by a null hypothesis. When it is true that all alternatives give rise to the same critical region, the test is called a *uniformly most powerful test*. We shall see that uniformly most powerful tests do exist for many important problems in statistics, while there are other equally important problems which do not have uniformly most powerful tests.

Going back to the problem of testing μ_0 against all $\mu > \mu_0$, let us consider the power of the test for a particular value of μ. The power is the probability of rejecting H_0 when it is false (when the true mean is $\mu > \mu_0$) and is given by

$$\int_A^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}\, dx$$

This quantity will clearly be a function of μ; it will be denoted by $P(\mu)$ and will be called the *power function* of the test. When the true mean

μ is far to the right of μ_0, the power will be nearly one, while when μ is near μ_0, the power will be small; at $\mu = \mu_0$ the power becomes equal to the Type I error, the probability that x falls in the critical region when $\mu = \mu_0$. The function is plotted in Fig. 59.

In view of the fact that the test we are considering is a uniformly most powerful test, we can make the following statement about its power function: the power function of any other test with the same probability of a Type I error will lie entirely below the curve of Fig. 59 (except, of course, that it will have the same value at μ_0. The general

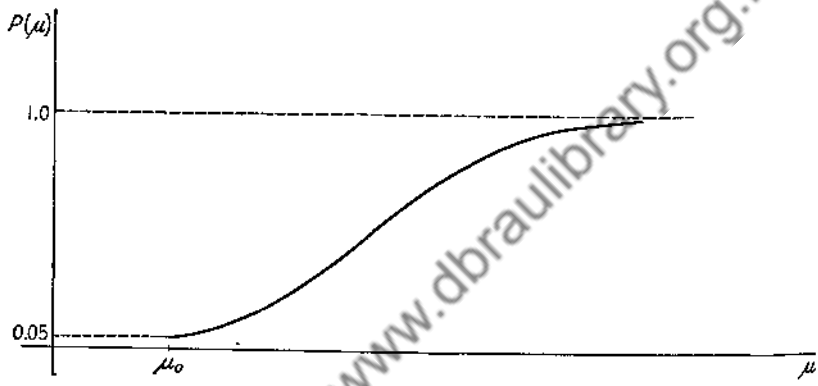


FIG. 59.

problem of studying tests can be set up in terms of the power function. For one parameter we may consider the test of the null hypothesis:

$$H_0: \theta = \theta_0$$

for the parameter of a density $f(x; \theta)$, where the possible values of θ lie in some interval which may be finite or infinite. In Fig. 60 are plotted several power functions for fixed Type I error. If a test exists which has a power function such as $P_1(\theta)$, then we have a very fine test indeed, and it can be shown that such tests can be obtained in general for large samples. For small samples, power functions are more likely to look like $P_2(\theta)$ and $P_3(\theta)$. And generally speaking, there will be no absolute criterion for choosing between tests. The test represented by $P_2(\theta)$ is better than that represented by $P_3(\theta)$ for $\theta > \theta_0$ and for $a < \theta < b$. But the test represented by $P_3(\theta)$ is better for $\theta < a$ and $b < \theta < \theta_0$.

The situation just described is typical. It will be possible to set up tests which are best for certain alternatives to H_0 but which are poor for other alternatives, and other tests will be better for these other

alternatives. The choice of a test must depend on the particular problem at hand and on the end one is most anxious to gain by the test. Thus, for example, if one had to make a choice between the two tests represented by $P_2(\theta)$ and $P_3(\theta)$ in Fig. 60, he would choose $P_3(\theta)$ if he wanted to be fairly certain to reject H_0 when θ was quite far from θ_0 in either direction. But $P_2(\theta)$ would be chosen if he were particularly concerned with the alternatives $\theta > \theta_0$.

We may mention here that an *unbiased* test is one such that its power function has a minimum at $\theta = \theta_0$. The test represented by $P_2(\theta)$ is biased. There are values of θ (just to the left of θ_0) for which the probability $1 - P(\theta)$ of accepting the null hypothesis is larger than for the null hypothesis itself.

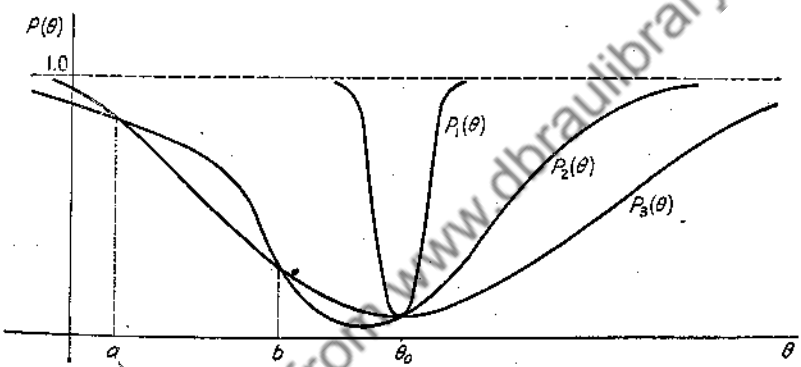


FIG. 60.

12.4. Simple and Composite Hypotheses. We turn now to hypotheses involving distributions with several parameters, and we may consider the general density $f(x; \theta_1, \theta_2, \cdots, \theta_k)$. The distribution may have several variates $x, y, z, \cdots$ without in any way changing the ensuing development. The *parameter space* with coordinates $\theta_1, \theta_2, \cdots, \theta_k$ will be denoted by the Greek letter Ω. A particular distribution in the family of distributions will be represented by a point in Ω. Thus if numerical values $\theta_{10}, \theta_{20}, \cdots, \theta_{k0}$ are substituted for $\theta_1, \theta_2, \cdots, \theta_k$ in $f(x; \theta_1, \theta_2, \cdots, \theta_k)$, a specific distribution function is determined. The numerical values $(\theta_{10}, \theta_{20}, \cdots, \theta_{k0})$ may be thought of as the coordinates of a point in a k-dimensional space. Thus the family of normal distributions with

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2}$$

may be represented by the upper half plane of Fig. 61. The coordinates of any point in the plane determine a particular member of the family. This upper half plane is Ω for the given family.

A *simple* null hypothesis is one which states that a distribution is one specific member of a given family. A *composite* null hypothesis is one which states that a distribution belongs to some subspace of the parameter space. We shall be primarily interested in subspaces of

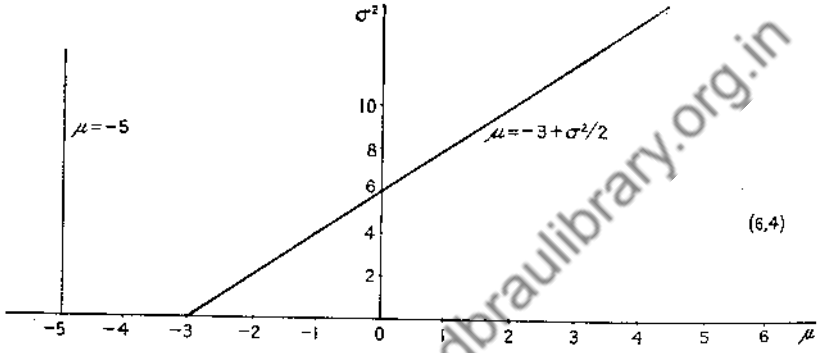


FIG. 61.

lower dimensionality than that of Ω. Referring to the two-parameter family of normal distributions, the null hypothesis:

$$H_0: \mu = 6, \qquad \sigma = 2$$

is a simple hypothesis because it completely specifies a single distribution in the family. The null hypothesis:

$$H_0: \mu = -5$$

is satisfied by all distributions with mean -5 regardless of the value of σ^2; this null hypothesis selects a subspace (the line $\mu = -5$) of the parameter space and is a composite hypothesis. Similarly

$$H_0: \mu = -3 + \frac{\sigma^2}{2}$$

is a composite hypothesis satisfied by all distributions with parameter values which satisfy the given relation.

Of course a simple hypothesis which selects a single point of the parameter space may be regarded as a special case of a composite hypothesis, because a point is a subspace; we shall use the word *composite* only when the subspace has more than one point. The symbol ω will be used to designate the subspace determined by the null hypothesis whether it is simple or composite.

For a general family of distributions $f(x; \theta_1, \theta_2, \cdots, \theta_k)$, a null hypothesis will state that the actual distribution belongs to some subspace ω of the complete parameter space Ω. If ω is a point, the hypothesis is simple; otherwise the hypothesis is composite.

12.5. The Likelihood-ratio Test and Its Large-sample Distribution. There are many ways to set up tests of hypotheses, and the best test in any given situation depends on the form of the distribution function and what alternatives are considered to be of primary importance. We shall not be able to study all the various methods of constructing tests but shall confine our attention to one method which usually leads to a very good test.

The *likelihood-ratio test* is closely related to maximum-likelihood estimation and to the ratio test described in Sec. 2 for a single alternative. Let $x_1, x_2, \cdots, x_n$ be a sample of size n from a population with density $f(x; \theta_1, \theta_2, \cdots, \theta_k)$. On the basis of this sample it is desired to test the null hypothesis:

$$H_0: f(x; \theta_1, \theta_2, \cdots, \theta_k) \text{ belongs to the subspace } \omega \text{ of } \Omega$$

The likelihood of the sample is

$$L = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \cdots, \theta_k) \tag{1}$$

The likelihood as a function of the parameters will ordinarily have a maximum as the parameters are allowed to vary over the entire parameter space Ω; we shall denote this maximum value by $L(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ or more briefly by $L(\hat{\Omega})$. In the subspace ω, L will also have a maximum value which we shall denote by $L(\hat{\omega})$. The likelihood ratio is the quotient of these two maxima and is denoted by

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \tag{2}$$

This quantity is necessarily a positive fraction; L is positive because it is a product of density functions, and $L(\hat{\omega})$ will be smaller than or at most equal to $L(\hat{\Omega})$ because there is less freedom for maximizing L in ω than in Ω. The ratio λ is a function of the sample observations only; it does not involve any parameters. The range of the variate λ is zero to one. $0 \leq \lambda \leq 1$

An illustration will reveal the logic of using λ as a test criterion. Let the family of distributions be the one-parameter family of normal distributions with unit variance, and let the sample consist of n obser-

vations $x_1, x_2, \cdots, x_n$. We shall test the null hypothesis

$$H_0: \mu = 3$$

that the population mean is actually three. This point is ω while the whole μ axis is Ω. The likelihood is

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i-\mu)^2}$$

which may be written

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i-\bar{x})^2-(n/2)(\bar{x}-\mu)^2}$$

The maximum value of this quantity in Ω is, of course, given by putting $\mu = \bar{x}$ to obtain

$$L(\hat{\Omega}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i-\bar{x})^2}$$

Since in this example ω is a point (the null hypothesis is simple), there is no opportunity to vary μ and the largest value of L in ω is simply its only value:

$$L(\hat{\omega}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i-\bar{x})^2-(n/2)(\bar{x}-3)^2}$$

The likelihood ratio is then

$$\lambda = e^{-(n/2)(\bar{x}-3)^2}$$

If $\bar{x}$ happens to be quite near 3, then the sample is reasonably consistent with H_0, and λ will be near one. If $\bar{x}$ is much greater than or less than 3, the sample will not be consistent with H_0 and λ will be near zero.

Clearly the proper critical region for testing H_0 is an interval

$$0 < \lambda < A$$

where A is some number (less than one) chosen to give the desired control of the Type I error.

This example illustrates the general situation. If the maximum-likelihood estimates fall in or near ω, the sample will be considered consistent with H_0 and the ratio λ will be near one. If the estimate $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ is distant from ω, then the sample will not be in accord with H_0 and λ will ordinarily be small. The critical region for λ will always be an interval of the form $0 < \lambda < A$. The number A

will be determined by the distribution of λ and the desired probability of a Type I error. If that probability is to be .05, for example, and if the density of λ is $g(\lambda)$ when H_0 is true, then A is the number for which

$$\int_0^A g(\lambda)d\lambda = .05$$

In order to prescribe the critical region for λ, it is necessary to know the distribution of λ when H_0 is true. If H_0 is a simple hypothesis [ω is a point $(\theta_{10}, \theta_{20}, \cdots, \theta_{k0})$, for example], then there will be a unique distribution determined for λ. But if H_0 is a composite, there may or may not be a unique distribution for λ. It is quite possible that the distribution of λ may be different for different parameter points in ω, and in this case A will not be uniquely determined. To specify a test, it is necessary to add further arbitrary criteria into the method of constructing the test. We shall not investigate these problems; we merely wish to observe here that the likelihood-ratio method as far as we have described it does not always lead to a unique test.

As is usually the case for large samples, a very satisfactory solution to the problem of testing hypotheses exists when one is dealing with large samples. The solution is based on a theorem which we shall not be able to prove because of the advanced character of its proof:

If a density function $f(x; \theta_1, \theta_2, \cdots, \theta_k)$ satisfies conditions like those enumerated in Sec. 10.8, *if the dimensionality of Ω is k, and if the dimensionality of ω is $r < k$, then $-2 \log \lambda$ is approximately distributed like chi square with $k - r$ degrees of freedom for large samples when H_0 is true.* Since $-2 \log \lambda$ increases as λ decreases and approaches infinity as λ approaches zero, the critical region for $-2 \log \lambda$ is the right-hand tail of the chi-square distribution. Therefore if we are dealing with a large sample and wish to test a null hypothesis with probability .05 for a Type I error, for example, it is only necessary to compute $-2 \log \lambda$ and compare it with the .05 level of chi square; if $-2 \log \lambda$ exceeds the chi-square level, H_0 would be rejected; otherwise H_0 would be accepted.

12.6. Tests on the Mean of a Normal Population. The foregoing ideas are well illustrated by a very common practical problem—that of testing whether the mean of a normal population has a specified value. We shall suppose that we have a sample of n observations, $x_1, x_2, \cdots, x_n$, from a normal population with mean μ and variance σ^2. We wish to test the null hypothesis:

$$H_0: \mu = \mu_0 \tag{1}$$

where μ_0 is a given number. The parameter space Ω is the half plane

of Fig. 61. The subspace ω characterized by the null hypothesis is the vertical line $\mu = \mu_0$.

We shall test H_0 by means of the likelihood ratio. The likelihood is

$$L = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2}\Sigma[(x_i-\mu)/\sigma]^2} \tag{2}$$

We have already seen that the values of μ and σ^2 which maximize L in Ω are

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Substituting these values in L, we have

$$L(\hat{\Omega}) = \left[\frac{1}{(2\pi/n)\,\Sigma(x_i - \bar{x})^2}\right]^{n/2} e^{-(n/2)} \tag{3}$$

To maximize L in ω, we put $\mu = \mu_0$, and the only remaining parameter is σ^2; the value of σ^2 which then maximizes L is readily found to be

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \mu_0)^2$$

which gives

$$L(\hat{\omega}) = \left[\frac{1}{(2\pi/n)\,\Sigma(x_i - \mu_0)^2}\right]^{n/2} e^{-(n/2)} \tag{4}$$

The ratio of (4) to (3) is the likelihood ratio:

$$\lambda = \left[\frac{\Sigma(x_i - \bar{x})^2}{\Sigma(x_i - \mu_0)^2}\right]^{n/2} \tag{5}$$

Our next step is to obtain the distribution of λ under H_0 and use that distribution to determine a number A so that the critical region $0 < \lambda < A$ will give the desired probability, .05, for example, of rejecting H_0 when it is true.

It happens that the distribution of λ is easily obtained in this case. The sum of squares in the denominator of (5) may be put in the form

$$\Sigma(x_i - \mu_0)^2 = \Sigma(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2$$

so that λ may be written

$$\lambda = \left\{\frac{1}{1 + [n(\bar{x} - \mu_0)^2/\Sigma(x_i - \bar{x})^2]}\right\}^{n/2} \tag{6}$$

We may recall (Sec. 11.2) that the fraction in the denominator is just

$$\frac{t^2}{n-1}$$

where t has the t distribution with $n - 1$ degrees of freedom when H_0 is true. To obtain the distribution of λ, we need merely to transform the t distribution by the substitution

$$\lambda = \left\{\frac{1}{1 + [t^2/(n-1)]}\right\}^{n/2} \tag{7}$$

It is not necessary actually to obtain the distribution of λ, because it is a monotonic function of t^2 and the test can be done just as well with t^2 as a criterion as with λ. Since $t^2 = 0$ when $\lambda = 1$ and t^2 becomes infinite when λ approaches zero, a critical region of the form $0 < \lambda < A$ is equivalent to a critical region $t^2 > B$ where B may be determined from A by equation (7). The critical values of t are therefore the extreme values either positive or negative, and a .05 critical region for t is the pair of intervals

$$t < -t_{.05} \qquad \text{and} \qquad t > t_{.05}$$

where $t_{.05}$ is the number for which

$$\int_{t_{.05}}^{\infty} f(t; n-1)dt = .025 \tag{8}$$

$f(t; n - 1)$ representing the t distribution with $n - 1$ degrees of freedom. The test of H_0 may therefore be performed as follows: We compute the quantity

$$\frac{\sqrt{n(n-1)}(\bar{x} - \mu_0)}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{9}$$

If it lies between $-t_{.05}$ and $t_{.05}$, H_0 is accepted; otherwise H_0 is rejected.

It is worth while to observe the connection between this test and the confidence-interval estimate of the mean. Supposing the mean of the population sampled to be μ', a 95 per cent confidence interval for μ' is just the set of values μ for which

$$-t_{.05} < \frac{\sqrt{n(n-1)}\,(\bar{x} - \mu)}{\sqrt{\Sigma(x_i - \bar{x})^2}} < t_{.05} \tag{10}$$

Hence the test of H_0 is equivalent to the following test: Construct a confidence interval for the population mean. If μ_0 lies in the con-

fidence interval, accept H_0; if μ_0 does not lie in the confidence interval, reject H_0.

We may also observe that the theorem at the end of the preceding section gives the correct distribution of λ for large samples. Since

$$-2 \log \lambda = n \log \left(1 + \frac{t^2}{n-1}\right) \tag{11}$$

and since the series expansion of $\log (1 + x)$ is

$$\log (1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad \text{for } -1 < x \leq 1 \tag{12}$$

we have

$$-2 \log \lambda = \frac{n}{n-1} t^2 - \frac{n}{(n-1)^2} \frac{t^4}{2} + \frac{n}{(n-1)^3} \frac{t^6}{3} - \cdots$$

for any fixed value of t, however large, provided n is taken large enough to make $t^2/(n-1)$ less than one. The first term of this series is

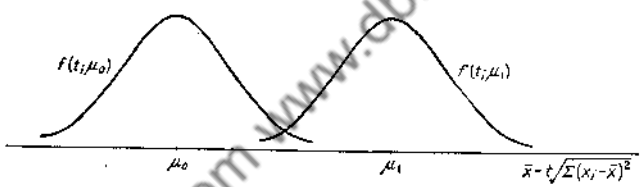


FIG. 62.

essentially t^2, and the others approach zero as n becomes large. Hence for large n, $-2 \log \lambda$ is approximately t^2. Furthermore t is approximately normally distributed for large samples (Sec. 11.7) with zero mean and unit variance if the true mean is μ_0; hence t^2 has approximately the chi-square distribution with one degree of freedom. This is in accord with the theorem, since Ω is a plane and has $k = 2$ dimensions, while ω is a line and has $r = 1$ dimension.

One-tailed Tests on the Mean. The test we have just constructed is called the two-tailed test of the mean, referring to the fact that the critical region is composed of both extremes of the t distribution. The test is not a uniformly most powerful test, and in fact there is no uniformly most powerful test for the given null hypothesis. If we consider a single one of the alternatives to μ_0, $\mu = \mu_1$, for example, where $\mu_1 > \mu_0$, the two t distributions are represented in Fig. 62. The best critical region for t, given a .05 probability of a Type I error, is obviously the interval $t > t_{.10}$, which cuts off 5 per cent of the area

under $f(t;\ \mu_0)$ on the right-hand tail. This will be the best critical region for any value of μ greater than μ_0. The power $P_1(\mu)$ of this test is plotted in Fig. 63 together with the power $P_2(\mu)$ of the two-tailed test. The one-tailed test is certainly better than the two-tailed test for alternatives $\mu > \mu_0$, and it is a uniformly most powerful test for those alternatives. But for alternatives $\mu < \mu_0$, the one-tailed test is no good at all; the power (probability of rejecting μ_0 when μ is the true mean) approaches zero as μ moves away from μ_0 towards the left.

There are many practical situations in which the one-tailed test should be employed. We may refer again to the light-bulb example used earlier in which the standard manufacturing process produced bulbs with a mean life of about 1400 hours. Any proposed new process

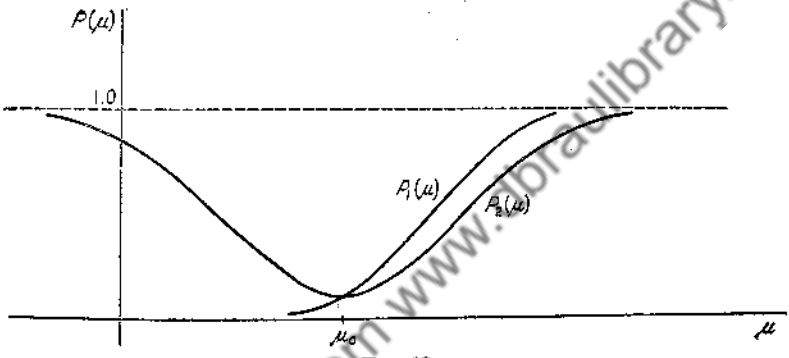


FIG. 63.

is of interest only if it produces bulbs with a greater mean life. One would test the null hypothesis $\mu = 1400$, and use the one-tailed test. Certainly no harm would be done by accepting $\mu = 1400$ if in fact μ were less than 1400, because the proposed process would simply be abandoned in either case. In other problems, the left-hand one-tailed test might be the appropriate test. For example, a new process might be thought to reduce the mean production cost per unit; one would test the null hypothesis that the mean cost θ for the new process was equal to the mean cost θ_0 for the standard process against the alternatives $\theta < \theta_0$. If one were comparing two proposed processes and wanted to choose the better for further research and development, then the two-tailed test would be appropriate.

12.7. The Difference between Means of Two Normal Populations. In many situations it is necessary to compare two means when neither is known; in the preceding section we assumed one was known. If, for example, one wished to compare two proposed new processes for manu-

facturing light bulbs, he would have to base the comparison on estimates of both process means. In comparing the yield of a new line of hybrid corn with that of a standard line, one would also have to use estimates of both mean yields because it is impossible to state the mean yield of the standard line for the given weather conditions under which the new line will be grown. It is necessary to compare the two lines by planting them in the same season and on the same soil type, and thereby obtain estimates of the mean yields for both lines under similar conditions. Of course the comparison is thus specialized; a complete comparison of the two lines would require tests over a period of years on a variety of soil types.

The general problem is this: We have two normal populations—one with variate x_1 which has mean μ_1 and variance σ_1^2, and one with variate x_2 which has mean μ_2 and variance σ_2^2. On the basis of two samples, one from each population, we wish to test the null hypothesis:

$$H_0: \mu_1 = \mu_2$$

The parameter space Ω here is four-dimensional; a joint distribution of x_1 and x_2 is specified when values are assigned to the four quantities $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. The subspace ω is three-dimensional because values for only three quantities $(\mu_2, \sigma_1^2, \sigma_2^2)$ need be specified in order to specify completely the joint distribution under the hypothesis that $\mu_1 = \mu_2$.

We shall suppose that there are m observations $(x_{11}, x_{12}, \cdots, x_{1m})$ in the sample from the first population and n observations $(x_{21}, x_{22}, \cdots, x_{2n})$ from the second. The likelihood is

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{m}{2}} e^{-\frac{1}{2}\sum_1^m\left(\frac{x_{1i}-\mu_1}{\sigma_1}\right)^2} \left(\frac{1}{2\pi\sigma_2^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_1^n\left(\frac{x_{2j}-\mu_2}{\sigma_2}\right)^2} \tag{1}$$

and its maximum in Ω is readily seen to be

$$L(\hat{\Omega}) = \left[\frac{m}{2\pi\sum_1^m (x_{1i} - \bar{x}_1)^2}\right]^{m/2} \left[\frac{n}{2\pi\sum_1^n (x_{2j} - \bar{x}_2)^2}\right]^{n/2} e^{-(m/2)} e^{-(n/2)} \tag{2}$$

If we put μ_1 and μ_2 equal to μ, say, and try to maximize L with respect to μ, σ_1^2, and σ_2^2, it will be found that the estimate of μ is given as the root of a cubic equation and will be a very complex function of the observations. The resulting likelihood ratio λ will therefore be a complicated function, and to find its distribution would be a tedious task indeed. No one has, in fact, worked out this distribution, and there is not much

incentive to do so because the distribution would very likely involve the ratio of the two variances. If it did involve this ratio, then it would be impossible to determine a critical region $0 < \lambda < A$ for given probability of a Type I error, because the ratio of the population variances is ordinarily unknown. A number of special devices can be employed to circumvent this difficulty, but we shall not pursue the problem further because statisticians are not yet agreed on what is the best procedure. Of course, for large samples this criterion may be used. The root of the cubic can be computed in any given instance by numerical methods, and λ can then be calculated. The quantity $-2 \log \lambda$ will have approximately the chi-square distribution with one degree of freedom.

When it can be assumed that the two populations have the same variance, the problem becomes relatively simple. The parameter space Ω is then three-dimensional with coordinates (μ_1, μ_2, σ^2), while ω, for the null hypothesis $\mu_1 = \mu_2$, has two coordinates, σ^2 and the common mean μ. In Ω we find

$$\hat{\mu}_1 = \bar{x}_1 \qquad \hat{\mu}_2 = \bar{x}_2$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[\sum_1^m (x_{1i} - \bar{x}_1)^2 + \sum_1^n (x_{2j} - \bar{x}_2)^2\right]$$

so that

$$L(\hat{\Omega}) = \left\{\frac{m+n}{2\pi[\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2]}\right\}^{(m+n)/2} e^{-[(m+n)/2]} \tag{3}$$

In ω

$$\hat{\mu} = \frac{1}{m+n}\left(\sum_1^m x_{1i} + \sum_1^n x_{2j}\right) = \frac{m\bar{x}_1 + n\bar{x}_2}{m+n}$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[\sum (x_{1i} - \hat{\mu})^2 + \sum (x_{2j} - \hat{\mu})^2\right]$$

$$= \frac{1}{m+n}\left[\sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2j} - \bar{x}_2)^2 + \frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2\right]$$

which gives

$$L(\hat{\omega}) =$$

$$\left\{\frac{m+n}{2\pi\left[\sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2j} - \bar{x}_2)^2 + \frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2\right]}\right\}^{(m+n)/2} e^{-[(m+n)/2]} \tag{4}$$

and finally

$$\lambda = \left[\frac{1}{1 + \dfrac{\dfrac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)^2}{\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2}}\right]^{(m+n)/2} \tag{5}$$

This last expression is very similar to the corresponding one obtained in the preceding section, and it turns out that this test can also be performed in terms of a quantity which has the t distribution. We know that $\bar{x}_1$ and $\bar{x}_2$ are independently normally distributed with means μ_1 and μ_2 and with variances σ^2/m and σ^2/n. Referring to Prob. 25 of Chap. 10, it is readily seen that $u = \bar{x}_1 - \bar{x}_2$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma^2[(1/m) + (1/n)]$. Under the null hypothesis the mean of u will be zero. The quantities $\Sigma(x_{1i} - \bar{x}_1)^2/\sigma^2$ and $\Sigma(x_{2j} - \bar{x}_2)^2/\sigma^2$ are independently distributed by chi-square laws with $m - 1$ and $n - 1$ degrees of freedom, respectively; hence their sum, say v, has the chi-square distribution with $m + n - 2$ degrees of freedom. Since under the null hypothesis

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{(1/m) + (1/n)}}$$

is normally distributed with zero mean and unit variance, the quantity

$$t = \frac{z}{\sqrt{v/(m + n - 2)}}$$

$$= \frac{\sqrt{mn/(m + n)}\,(\bar{x}_1 - \bar{x}_2)}{\sqrt{[\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2]/(m + n - 2)}} \tag{6}$$

has the t distribution with $m + n - 2$ degrees of freedom. The likelihood ratio is

$$\lambda = \left\{\frac{1}{1 + [t^2/(m + n - 2)]}\right\}^{(m+n)/2} \tag{7}$$

and its distribution is determined by the t distribution. The test would, of course, be done in terms of t rather than λ. Possible 5 per cent critical regions for t are again $t < -t_{.10}$, $t > t_{.10}$, or $t^2 > t^2_{.05}$, and the choice between these would depend on the problem at hand. If, for example, the first population referred to the yield of a variety of corn in common use, while the second referred to the yield of a proposed substitute, the critical region would be $t < -t_{.10}$. If one were comparing two proposed substitutes, the two-tailed test given by $t^2 > t^2_{.05}$ would be used.

We may observe here that it is possible to determine a confidence interval for the difference $\mu_1 - \mu_2$ of the population means by using the t distribution. When the two means are different, the quantity

$$y = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{(1/m) + (1/n)}}$$

is normally distributed with zero mean and unit variance so that

$$t = \frac{y}{\sqrt{v/(m + n - 2)}}$$

has the t distribution with $m + n - 2$ degrees of freedom. Since t does not involve σ^2 but only the parameter $\theta = \mu_1 - \mu_2$, a confidence interval for θ can be obtained. Upper and lower limits, for a 95 per cent confidence interval, for example, would be obtained by solving the equations

$$t = \pm t_{.05}$$

for θ.

12.8. Tests on the Variance of a Normal Distribution. To test the null hypothesis that the variance of a normal population has a specified value σ_0^2 on the basis of a sample of size n, we first maximize

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2}\Sigma[(x_i-\mu)/\sigma]^2} \tag{1}$$

in Ω, which has coordinates (μ, σ^2), and in ω, which is the line $\sigma^2 = \sigma_0^2$. The ratio of these maxima is readily found to be

$$\lambda = \left(\frac{u}{n}\right)^{n/2} e^{-\frac{1}{2}(u-n)} \tag{2}$$

where

$$u = \sum_1^n \frac{(x_i - \bar{x})^2}{\sigma_0^2} \tag{3}$$

Since u is known to have the chi-square distribution with $n - 1$ degrees of freedom, the distribution of λ could be found by transforming the chi-square distribution by (2). The test may, however, be done using u as a criterion. On plotting equation (2) (Fig. 64), it is seen that a critical region for λ of the form $0 < \lambda < A$ corresponds to the pair of intervals $0 < u < a$ and $b < u < \infty$ for u, where a and b are such that the ordinates of (2) are equal

It can be shown that the power of this test will be slightly improved if in the criterion (2), n is replaced by $n - 1$, i.e., if

$$\lambda' = \left(\frac{u}{n-1}\right)^{(n-1)/2} e^{-\frac{1}{2}(u-n+1)} \tag{4}$$

is used as the test criterion. We shall not prove this statement; it is an unimportant refinement unless n is small. Using λ', the critical region for u would be determined by numbers a' and b', say, such that the ordinates of (4) were equal at those points. Since the chi-square distribution is not tabulated in sufficient detail to determine these numbers, it is common practice to use $\chi^2_{.975} < u < \chi^2_{.025}$ for the accept-

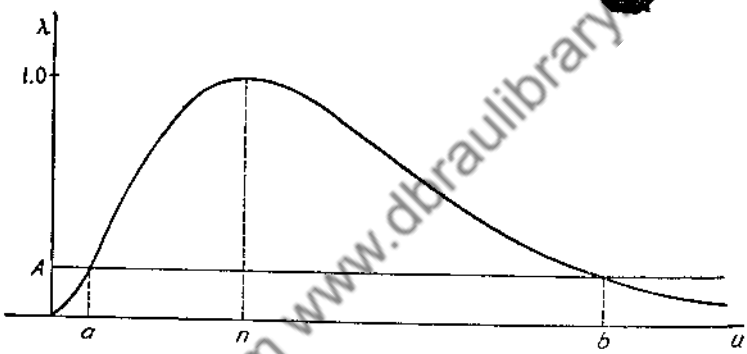


FIG. 64.

ance region rather than $a' < u < b'$, if, for example, the probability of a Type I error is specified to be .05. Here again there are some situations in which one of the one-tailed tests, $u < \chi^2_{.95}$ or $u > \chi^2_{.05}$, would be preferred over the two-tailed test.

Equality of Two Variances. Given samples from each of two normal populations with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2), we may test

$$H_0: \sigma_1^2 = \sigma_2^2$$

The likelihood ratio is found to be

$$\lambda = \frac{\left[\dfrac{m+n}{2\pi[\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2j} - \bar{x}_2)^2]}\right]^{(m+n)/2}}{\left[\dfrac{m}{2\pi\Sigma(x_{1i} - \bar{x}_1)^2}\right]^{m/2}\left[\dfrac{n}{2\pi\Sigma(x_{2j} - \bar{x}_2)^2}\right]^{n/2}} \tag{5}$$

where the notation is the same as that of the preceding section. This

criterion may be put in the form

$$\lambda = \frac{(m+n)^{(m+n)/2}}{m^{m/2}n^{n/2}} \frac{\left(\frac{m-1}{n-1}F\right)^{m/2}}{\left(1+\frac{m-1}{n-1}F\right)^{(m+n)/2}} \tag{6}$$

where F is the variance ratio:

$$F = \frac{(n-1)\Sigma(x_{1i}-\bar{x}_1)^2}{(m-1)\Sigma(x_{2j}-\bar{x}_2)^2} \tag{7}$$

which has the F distribution with $m-1$ and $n-1$ degrees of freedom when H_0 is true. On plotting λ as a function of F, it is apparent that the critical region $0 < \lambda < \Lambda$ corresponds to a two-tailed test on F. It is customary to make the two tails have equal areas (though this is not quite the best test) because the tabulations of F make this region easy to determine. Again one-tailed tests are often appropriate in problems of this kind.

Equality of Several Variances. A problem that frequently arises in applied statistics is that of testing whether several normal populations have the same variance. Let $x_{i1}, x_{i2}, \cdots, x_{in_i}$ be a sample of size n_i from a normal population with mean μ_i and variance σ_i^2, and let there be one such sample from each of k populations ($i = 1, 2, \cdots, k$). It is easily found that the likelihood ratio criterion for testing

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$$

is

$$\lambda = \frac{\prod_{i=1}^{k} (S_i/n_i)^{n_i/2}}{(\Sigma S_i/\Sigma n_i)^{\Sigma n_i/2}} \tag{8}$$

where

$$S_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Equation (8) is the direct generalization of (5). The distribution of λ is a complicated function, and from the applied point of view it is of no use because it would not be feasible to tabulate the function anyway. It contains k parameters $n_1, n_2, \cdots, n_k$ and would have to be tabulated for all possible combinations of values of these parameters for every value of k. When the n_i are large, the criterion does provide a test because $-2 \log \lambda$ will then have approximately the

chi-square distribution with $k - 1$ degrees of freedom under H_0. The number of degrees of freedom is $k - 1$ because Ω has $2k$ dimensions [the joint distribution of the x_{ij} is specified when $(\mu_1, \mu_2, \cdots, \mu_k, \sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2)$ are specified], while ω has $k + 1$ dimensions corresponding to the k means and the common variance.

It turns out that the test may be made even when the n_i are not large. The distribution of $-2 \log \lambda$ has been investigated and found to be well approximated by the chi-square distribution with $k - 1$ degrees of freedom in any case. The approximation is even better and the test somewhat improved if, instead of $-2 \log \lambda$, the criterion

$$u = \frac{-2 \log \lambda'}{1 + \dfrac{1}{3(k-1)}\left(\sum \dfrac{1}{n_i} - \dfrac{1}{\Sigma n_i}\right)} \tag{9}$$

is employed, where λ' represents the expression (8) with all the n_i replaced by $n_i - 1$. The quantity $-2 \log \lambda$ gives a slightly biased test, and u has been defined so as to make the test unbiased. The critical region for the test is, of course, the right-hand tail of the chi-square distribution; a two-tailed test is never appropriate here.

12.9. The Goodness-of-fit Test. If a population has the multinomial density

$$f(x_i; p_i) = \prod_{i=1}^{k} p_i^{x_i} \qquad x_i = 0, 1;\ \Sigma x_i = 1;\ \Sigma p_i = 1 \tag{1}$$

as would be the case in sampling with replacement from a population of individuals which could be classified into k classes, a common problem is that of testing whether the probabilities have specified numerical values. Thus the result of casting a die may be classified into one of six classes. On the basis of a sample of observations, we may wish to test whether the die is true, i.e., whether

$$p_i = \tfrac{1}{6} \qquad i = 1, 2, \cdots, 6$$

Let us suppose that n observations are drawn from a population with distribution (1) and that the number of observations that fall in the ith class is $n_i (\Sigma n_i = n)$. The likelihood of the sample is

$$L = \prod_{1}^{k} p_i^{n_i} \tag{2}$$

and we shall test the null hypothesis

$$H_0: p_i = p_{0i}$$

where the p_{0i} are given numbers. The parameter space Ω has $k - 1$ dimensions (given $k - 1$ of the p_i, the remaining one is determined by $\Sigma p_i = 1$), while ω is a point. It is readily found that L is maximized in Ω when

$$\hat{p}_i = \frac{n_i}{n} \tag{3}$$

hence

$$L(\hat{\Omega}) = \frac{1}{n^n} \prod_1^k n_i^{n_i} \tag{4}$$

In ω the maximum value of L is its only value

$$L(\hat{\omega}) = \prod_1^k p_{0i}^{n_i} \tag{5}$$

The likelihood ratio is

$$\lambda = n^n \prod_1^k \left(\frac{p_{0i}}{n_i}\right)^{n_i} \tag{6}$$

and the critical region is $0 < \lambda < A$, where A is chosen to give the desired probability of a Type I error. For small n, the distribution of λ may be tabulated directly in order to determine A; for large values of n, we may use the fact that $-2 \log \lambda$ has approximately the chi-square distribution with $k - 1$ degrees of freedom. The chi-square approximation is surprisingly good even if n is small provided that $k > 2$.

Another test commonly used for testing H_0 was proposed (by Karl Pearson) before the general theory of testing hypotheses was developed. This test criterion is

$$u = \sum \frac{(n_i - np_{0i})^2}{np_{0i}} \tag{7}$$

which in large samples has approximately the chi-square distribution with $k - 1$ degrees of freedom when H_0 is true. The argument for using (7) as a criterion is briefly this: The approximate large-sample distribution of the $\hat{p}_i = n_i/n$ $(i = 1, 2, \cdots, k - 1)$ is normal and is in fact

$$f(\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{k-1}) = \left(\frac{n}{2\pi}\right)^{\frac{k-1}{2}} \sqrt{\frac{1}{\prod_1^k p_i}}\, e^{-\frac{1}{2}\sum\sum n\left(\frac{\delta_{ij}}{p_i}+\frac{1}{p_k}\right)(\hat{p}_i - p_i)(\hat{p}_j - p_j)} \tag{8}$$

as follows from equation (10.9.18) on replacing k by $k - 1$. We have seen in Chap. 10 that the quadratic form of a multivariate normal distribution in $k - 1$ variates has the chi-square distribution with $k - 1$ degrees of freedom; hence

$$v = \sum_{1}^{k-1} \sum_{1}^{k-1} n \left(\frac{\delta_{ij}}{p_i} + \frac{1}{p_k} \right) (\hat{p}_i - p_i)(\hat{p}_j - p_j) \tag{9}$$

has that distribution approximately for large samples. On summing (9) with respect to j and remembering that

$$p_k = 1 - \sum_{1}^{k-1} p_i$$

we find

$$v = \sum_{i=1}^{k} \frac{n(\hat{p}_i - p_i)^2}{p_i} \tag{10}$$

or

$$v = \sum_{1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{11}$$

which is the same as (7) if the true values of the p_i are p_{0i}. But let us suppose the true p_i are p_{1i}, at least some of which are different from p_{0i}; then

$$u = \sum \frac{(n_i - np_{1i})^2}{np_{1i}} \tag{12}$$

has approximately the chi-square distribution with expected value $k - 1$. The quantity

$$u = \sum \frac{(n_i - np_{0i})^2}{np_{0i}} \tag{13}$$

is easily shown to have an expected value

$$E(u) = \sum \frac{1}{np_{0i}} [np_{1i}(1 - p_{1i}) + n^2(p_{1i} - p_{0i})^2] \tag{14}$$

which is certainly larger than $k - 1$ for sufficiently large n, and in fact is larger than $k - 1$ for any n, because if $E(u)$ is minimized with respect to the p_{0i}, it is found that the minimum occurs when $p_{0i} = p_{1i}$ and is therefore $k - 1$. The argument for using u as a test criterion is now evident. If the true p_i are p_{0i}, u will have the chi-square distribution approximately, while if the true p_i are not p_{0i}, u will be distributed with a larger mean value, and that mean value becomes infinite as n

becomes large. Hence it is reasonable to test H_0 by using u as a criterion and the right-hand tail of the distribution as the critical region.

We have discussed Pearson's chi-square criterion because of its historical interest and because it is still commonly used to test H_0. It is, in fact, equivalent to the likelihood-ratio test in large samples. Perhaps the easiest way to show this is to write λ in the form

$$\lambda = K \frac{n!}{\Pi n_i!} \prod p_{0i}^{n_i}$$

where

$$K = \frac{n^n}{\Pi n_i^{n_i}} \frac{\Pi n_i!}{n!}$$

If the variates of (8) are changed from $\hat{p}_i$ to n_i, the function will be unchanged except for the change in factor $n^{(k-1)/2}$ since $n\, d\hat{p}_i = dn_i$. It follows from Sec. 10.9 that $\lambda n^{k-1}/K$ approaches (8). By using Stirling's formula (Sec. 2.3) for the factorials in K, it can be shown that K/n^{k-1} just cancels the coefficient of the exponential in (8) to within terms of order $1/\sqrt{n}$; hence $-2 \log \lambda$ is asymptotically equivalent to u.

12.10. Tests of Independence in Contingency Tables. A contingency table is multiple classification. Thus in a public-opinion survey the individuals interviewed may be classified according to their attitude on a political proposal and according to sex, to obtain a table of the form:

	Favor	Oppose	Undecided
Men	1154	475	243
Women	1083	442	362

This is a 2×3 contingency table. The individuals are classified by two criteria, one having two categories and the other three categories. The six distinct classifications are called *cells*. A three-way contingency table would have been obtained had the individuals been further classified according to a third criterion, say according to annual income group. If there were five income groups set up (such as: under \$1000, \$1000 to \$3000, $\cdots$), the contingency table would be called a $2 \times 3 \times 5$ table and would have 30 cells into which a person might be put. It is often quite convenient to think of the cells as cubes in a block two units wide, three units long, and five units deep. If the

individuals were still further classified into eight geographical locations, one would have a four-way $2 \times 3 \times 5 \times 8$ contingency table with 240 cells in a four-dimensional block with edges 2, 3, 5, and 8 units long. The contingency table provides a technique for investigating suspected relationships. Thus one may suspect that men and women will react differently to a certain political proposal, in which case he would construct such a table as the one above and test the null hypothesis that their attitudes were independent of their sex. To consider another example, a geneticist may suspect that susceptibility to a certain disease is heritable. He would classify a sample of individuals according to (1) whether or not they ever had the disease, (2) whether or not their fathers had the disease, (3) whether or not their mothers had the disease. In the resulting $2 \times 2 \times 2$ contingency table he would test the null hypothesis that classification (1) was independent of (2) and (3). Again a medical research worker might suspect a certain environmental condition favored a given disease and classify individuals according to (1) whether or not they ever had the disease, (2) whether or not they were subject to the condition. An industrial engineer would use a contingency table to discover whether or not two kinds of defects in a manufactured product were due to the same underlying cause or to different causes. It is apparent that the technique can be a very useful tool in any field of research.

Two-way Contingency Tables. We shall suppose that n individuals or items are classified according to two criteria A and B, that there are r classifications $A_1, A_2, \cdots, A_r$ in A and s classifications $B_1, B_2, \cdots, B_s$ in B, and that the number of individuals belonging to A_i and B_j is n_{ij}. We have then an $r \times s$ contingency table with cell frequencies n_{ij} and $\Sigma n_{ij} = n$. As a further notation we shall denote the

	B_1	B_2	B_3	$\cdots$	B_s
A_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1s}
A_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2s}
A_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3s}
$\vdots$					
A_r	n_{r1}	n_{r2}	n_{r3}	$\cdots$	n_{rs}

(1)

row totals by $n_{i.}$ and the column totals by $n_{.j}$,

$$n_{i.} = \sum_j n_{ij} \qquad n_{.j} = \sum_i n_{ij}$$

Of course

$$\sum_i n_{i.} = \sum_j n_{.j} = n$$

We shall now set up a probability model for the problem with which we wish to deal. The n individuals will be regarded as a sample of size n from a multinomial population with probabilities p_{ij} ($i = 1, 2, \cdots, r; j = 1, 2, \cdots, s$). The probability distribution for a single observation is (Sec. 10.9)

$$f(x_{11}, x_{12}, \cdots, x_{rs}) = \prod_{ij} p_{ij}^{x_{ij}} \qquad x_{ij} = 0, 1; \sum_{i,j} x_{ij} = 1 \tag{2}$$

We wish to test the null hypothesis that the A and B classifications are independent, i.e., that the probability an individual falls in B_j is not affected by the A class to which the individual happens to belong. Using the symbolism of Chap. 2, we would write

$$P(B_j|A_i) = P(B_j) \qquad P(A_i|B_j) = P(A_i)$$

or

$$P(A_i, B_j) = P(A_i)P(B_j)$$

If we denote the marginal probabilities $P(A_i)$ by p_i ($i = 1, 2, \cdots, r$) and the marginal probabilities $P(B_j)$ by q_j, the null hypothesis is simply

$$H_0: p_{ij} = p_i q_j \qquad \Sigma p_i = 1, \ \Sigma q_j = 1 \tag{3}$$

When the null hypothesis is not true, there is said to be *interaction* between the two criteria of classification.

The complete parameter space Ω for the distribution (1) has $rs - 1$ dimensions (having specified all but one of the p_{ij}, the remaining one is fixed by $\sum_{ij} p_{ij} = 1$), while under H_0 we have a parameter space ω with $r - 1 + s - 1$ dimensions. The likelihood for a sample of size n is

$$L = \prod_{ij} p_{ij}^{n_{ij}} \tag{4}$$

and its maximum in Ω occurs when

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \tag{5}$$

In ω,

$$L = \prod_{ij} (p_i q_j)^{n_{ij}} = \left(\prod_i p_i^{n_{i\cdot}}\right)\left(\prod_j q_j^{n_{\cdot j}}\right) \tag{6}$$

and its maximum occurs at

$$\hat{p}_i = \frac{n_{i\cdot}}{n} \qquad \hat{q}_j = \frac{n_{\cdot j}}{n} \tag{7}$$

The likelihood ratio is therefore

$$\lambda = \frac{\left(\prod_i n_{i\cdot}^{n_{i\cdot}}\right)\left(\prod_j n_{\cdot j}^{n_{\cdot j}}\right)}{n^n \prod_{i,j} n_{ij}^{n_{ij}}} \tag{8}$$

The distribution of λ under the null hypothesis is not unique because the hypothesis is composite and the exact distribution of λ does involve the unknown parameters p_i and q_j. For large samples we do have a test, however, because $-2 \log \lambda$ is, in that case, approximately distributed by the chi-square law with

$$rs - 1 - (r + s - 2) = (r - 1)(s - 1)$$

degrees of freedom, and on the basis of this distribution a unique critical region for λ may be determined.

In casting about for a test which may be used when the sample is not large, we may inquire how it is that a test criterion comes to have a unique distribution for large samples when the distribution actually depends on unknown parameters which may have any values in certain ranges. The answer is that the parameters are not really unknown; they can be estimated, and their estimates approach their true values as the sample size increases. In the limit as n becomes infinite the parameters are known exactly, and it is at that point that the distribution of λ actually becomes unique. It is unique because a particular point in ω is selected as the true parameter point, so that the n_{ij} are given a unique distribution, and the distribution of λ is then determined by this distribution.

It would appear reasonable to employ a similar procedure to set up a test for small samples, i.e., to define a distribution for λ by using the estimates for the unknown parameters. In the present problem, since the estimates of the p_i and q_j are given by (7), we might just substitute those values in the distribution function of the n_{ij} and use that distribution to obtain a distribution for λ. However we should still be in trouble; the critical region would depend on the marginal totals $n_{i\cdot}$

and $n_{.j}$; hence the probability of a Type I error would vary from sample to sample for any fixed critical region $0 < \lambda < A$.

There is a way out of this difficulty which is well worth investigation because of its own interest and because the problem is important in applied statistics. Let us denote the joint density of all the n_{ij} briefly by $f(n_{ij})$, the marginal density of all the $n_{i.}$ and $n_{.j}$ by $g(n_{i.}, n_{.j})$, and the conditional density of the n_{ij}, given the marginal totals, by

$$f(n_{ij}|n_{i.}, n_{.j}) = \frac{f(n_{ij})}{g(n_{i.}, n_{.j})}$$

Under the null hypothesis, this conditional distribution happens to be independent of the unknown parameters (as we shall show presently); the estimators $n_{i.}/n$ and $n_{.j}/n$ form a sufficient set of statistics for the p_i and q_j. This fact will enable us to construct a test.

The joint density of the n_{ij} is simply the multinomial

$$f(n_{11}, n_{12}, \cdots, n_{rs}) = \frac{n!}{\prod_{ij} n_{ij}!} \prod p_{ij}^{n_{ij}} \tag{9}$$

in Ω, and in ω (we are interested in the distribution of λ under H_0) this becomes

$$f(n_{11}, n_{12}, \cdots, n_{rs}) = \frac{n!}{\prod_{ij} n_{ij}!} \left(\prod_i p_i^{n_{i.}}\right)\left(\prod_j q_j^{n_{.j}}\right) \tag{10}$$

To obtain the desired conditional distribution, we must first find the distribution of the $n_{i.}$ and $n_{.j}$, and this is accomplished by summing (10) over all sets of n_{ij} such that

$$\sum_i n_{ij} = n_{.j} \qquad \sum_j n_{ij} = n_{i.} \tag{11}$$

For fixed marginal totals, only the factor $1/\Pi n_{ij}!$ in (10) is involved in the sum, so we have in effect to sum that factor over all n_{ij} subject to (11). The desired sum is given by comparing the coefficients of $\prod_i x_i^{n_{i.}}$ in the expression

$$(x_1 + \cdots + x_r)^{n_{.1}}(x_1 + \cdots + x_r)^{n_{.2}} \cdots (x_1 + \cdots + x_r)^{n_{.s}} = (x_1 + \cdots + x_r)^n \tag{12}$$

On the right the coefficient of $\Pi x_i^{n_{i.}}$ is simply

$$\frac{n!}{\prod_i n_{i.}!} \tag{13}$$

On the left there are terms in $\Pi x_i^{n_{ij}}$ with coefficients of the form

$$\frac{n_{.1}!}{\prod_i n_{i1}!}\frac{n_{.2}!}{\prod_i n_{i2}!}\cdots\frac{n_{.s}!}{\prod_i n_{is}!} = \frac{\prod_j n_{.j}!}{\prod_{ij} n_{ij}!} \tag{14}$$

where n_{ij} is the exponent of x_i in the jth multinomial. In this expression the n_{ij} satisfy the conditions (11); the first condition is satisfied in view of the multinomial theorem (Sec. 2.4), while the second is satisfied because we require the power of x_i in these terms to be $n_{i.}$. The sum of all such coefficients (14) must equal (13); hence we may write

$$\sum \frac{1}{\Pi n_{ij}!} = \frac{n!}{\prod_i n_{i.}!\prod_j n_{.j}!}. \tag{15}$$

This is precisely the sum we require, because there is obviously one and only one coefficient of the form of (14) on the left of (12) for every possible contingency table (1) with given marginal totals. The distribution of the $n_{i.}$ and $n_{.j}$ is, therefore,

$$g(n_{i.}, n_{.j}) = \frac{(n!)^2}{(\Pi n_{i.}!)(\Pi n_{.j}!)}\left(\prod p_i^{n_{i.}}\right)\left(\prod q_j^{n_{.j}}\right) \tag{16}$$

which shows incidentally that the $n_{i.}$ are distributed independently of the $n_{.j}$; this is unexpected because $n_{1.}$ and $n_{.1}$, for example, have the variate n_{11} in common.

The conditional distribution of the n_{ij}, given the marginal totals, is obtained by dividing (10) by (16) to obtain

$$f(n_{11}, n_{12}, \cdots, n_{rs}|n_{1.}, n_{2.}, \cdots, n_{.s}) = \frac{(\Pi n_{i.}!)(\Pi n_{.j}!)}{n!\Pi n_{ij}!} \tag{17}$$

which, happily, does not involve the unknown parameters and shows that the estimators are sufficient.

To see how a test may be constructed, let us consider the general situation in which a criterion λ for some test has a distribution $u(\lambda; \theta)$ which involves an unknown parameter θ. If θ has a sufficient estimator $\hat{\theta}$, then the joint density of λ and $\hat{\theta}$ may be written

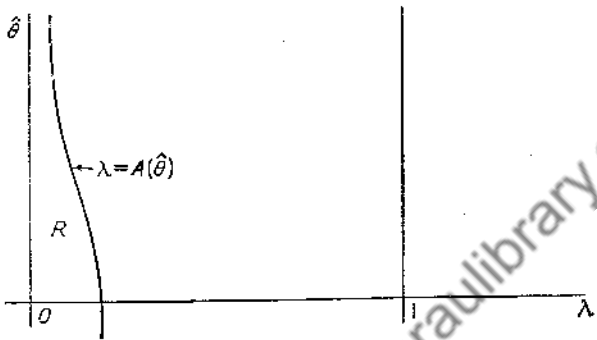
$$v(\lambda, \hat{\theta}; \theta) = v_1(\lambda|\hat{\theta})v_2(\hat{\theta}; \theta) \tag{18}$$

and the conditional density of λ given $\hat{\theta}$ will not involve θ. Using the

conditional distribution, we may find a number $A(\hat{\theta})$ for every $\hat{\theta}$ such that

$$\int_0^{A(\hat{\theta})} v_1(\lambda|\hat{\theta})d\lambda = .05$$

for example. In the λ, $\hat{\theta}$ plane the curve $\lambda = A(\hat{\theta})$ together with the line $\lambda = 0$ will determine a region R. The probability that a sample



will give rise to a pair of values $(\lambda, \hat{\theta})$ which correspond to a point in R is exactly .05 because

$$\begin{aligned} P[(\lambda, \hat{\theta}) \text{ in } R] &= \int_{-\infty}^{\infty} \int_0^{A(\hat{\theta})} v(\lambda, \hat{\theta}; \theta)d\lambda\, d\hat{\theta} \qquad (19) \\ &= \int_{-\infty}^{\infty} \left[\int_0^{A(\hat{\theta})} v_1(\lambda|\hat{\theta})d\lambda \right] v_2(\hat{\theta}; \theta)d\hat{\theta} \\ &= \int_{-\infty}^{\infty} .05 v_2(\hat{\theta}; \theta)d\hat{\theta} \\ &= .05 \end{aligned}$$

Hence we may test the hypothesis by using $\hat{\theta}$ in conjunction with λ. The critical region is a plane region instead of an interval $0 < \lambda < A$; it is such a region that whatever the unknown value of θ may be, the Type I error has a specified probability. The test in any given situation actually amounts to a conditional test; we observe $\hat{\theta}$ and test λ by an interval $0 < \lambda < A(\hat{\theta})$ using the conditional distribution of λ given $\hat{\theta}$. It is to be observed that this device cannot be employed unless θ has a sufficient estimator.

The above technique is obviously applicable when θ is a set of parameters rather than a single parameter and has a set of sufficient estimators $\hat{\theta}$. In particular the technique may be employed to test the criterion (8) for the null hypothesis of a two-way contingency table. One merely uses the conditional distribution (17) and determines an

interval $0 < \lambda < A(n_{i.}; n_{.j})$ which has the desired probability of a Type I error for the observed marginal totals.

In applications of this test one is confronted with a very tedious computation in determining the distribution of λ unless r, s, and the marginal totals are quite small. It can be shown, however, that the large-sample approximation may be used without appreciable error except when both r and s equal two. In the latter instance, other simplifying approximations have been developed (see, for example, Fisher and Yates, "Tables for Statisticians and Biometricians," Oliver & Boyd, Ltd., Edinburgh, 1938), but we shall not explore the problem that far.

If the distribution (17) is replaced by its multivariate normal approximation, it can be shown that the criterion

$$u = \sum_{i,j} \frac{[n_{ij} - (n_{i.}n_{.j}/n)]^2}{n_{i.}n_{.j}/n} \tag{20}$$

has approximately the chi-square distribution with $(r - 1)(s - 1)$ degrees of freedom and is a reasonable criterion for testing H_0 of (3). This is the criterion first proposed (by Karl Pearson) for testing the hypothesis, and it differs from $-2 \log \lambda$ by terms of order $1/\sqrt{n}$. The two criteria are therefore essentially equivalent unless n is small. The argument that u is a reasonable criterion is entirely analogous to that used to justify (7) in the preceding section.

Three-way Contingency Tables. If the elements of a population can be classified according to three criteria A, B, C with classifications A_i $(i = 1, 2, \cdots, s_1)$, B_j $(j = 1, 2, \cdots, s_2)$, and C_k $(k = 1, 2, \cdots, s_3)$, a sample of n individuals may be classified in a three-way $s_1 \times s_2 \times s_3$ contingency table. We shall let p_{ijk} represent the probabilities associated with the individual cells, n_{ijk} be the numbers of sample elements in the individual cells, and, as before, marginal totals will be indicated by replacing the summed index by a dot; thus

$$n_{i.k} = \sum_{j=1}^{s_2} n_{ijk} \qquad n_{..k} = \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} n_{ijk} \tag{21}$$

There are four hypotheses that may be tested in connection with this table. We may test whether all three criteria are mutually independent, in which case the null hypothesis is

$$p_{ijk} = p_i q_j r_k \tag{22}$$

or we may test whether any one of the three criteria is independent

of the other two. Thus to test whether the B classification was independent of A and C, we would set up the null hypothesis

$$p_{ijk} = p_{ik}q_j \tag{23}$$

The procedure for testing these hypotheses is entirely analogous to that for the two-way tables. The likelihood of the sample is

$$L = \prod_{ijk} p_{ijk}^{n_{ijk}} \qquad \sum_{ijk} p_{ijk} = 1 \qquad \sum_{ijk} n_{ijk} = n \tag{24}$$

In Ω the maximum of L occurs when

$$p_{ijk} = \frac{n_{ijk}}{n} \tag{25}$$

so that

$$L(\hat{\Omega}) = \frac{1}{n^n} \prod_{ijk} n_{ijk}^{n_{ijk}} \tag{26}$$

To test (23), for example, we would make the substitution (23) in (24) and maximize L with respect to the $p_{ik}q_j$ to find

$$\hat{p}_{ik} = \frac{n_{i.k}}{n} \qquad \hat{q}_j = \frac{n_{.j.}}{n} \tag{27}$$

and

$$L(\hat{\omega}) = \frac{1}{n^{2n}} \left(\prod_{ik} n_{i.k}^{n_{i.k}}\right) \left(\prod_j n_{.j.}^{n_{.j.}}\right) \tag{28}$$

The likelihood ratio λ is given by the quotient of (28) and (26), and in large samples $-2 \log \lambda$ has the chi-square distribution with

$$s_1s_2s_3 - 1 - [(s_1s_3 - 1) + s_2 - 1] = (s_1s_3 - 1)(s_2 - 1)$$

degrees of freedom. Again the large-sample distribution is quite adequate for all practical purposes unless the test has only one degree of freedom.

12.11. Notes and References. It is now apparent that the sampling distributions based on normal theory have an all-important role in statistical inference, both in estimation and in tests of hypotheses. We shall cite here the classic references.

The chi-square distribution is due to Karl Pearson [1], who was the first major contributor to the theory of statistics. Pearson published nearly one hundred papers from about 1895 to 1935 which laid a firm foundation for modern statistics. He formulated the basic problems and went far along the way to solving many of them. He is rightly regarded as the founder of the science of statistical inference.

We have already mentioned that W. S. Gosset first showed the way to make an exact inference. Before his paper [2] was published, the accepted method of making inferences was to substitute estimates for parameters in population distributions. Gosset was the second major contributor to the field of statistics; he published about twenty papers in this field between 1908 and 1931.

The F distribution was derived by R. A. Fisher [3], who also gave the first mathematical derivation of the t distribution [4]; Gosset had obtained it by heuristic methods. Fisher is the real giant in development of the theory of statistics. His first paper was published in 1912, and his work continues unabated today. Although hundreds of scholars have contributed to the science of statistics, this one man must be credited with at least half the essential and important developments as the theory now stands.

The general theory of testing hypotheses, as we have presented it, is due to J. Neyman and E. S. Pearson (the son of Karl Pearson), who published the theory in an important series of joint papers beginning in 1928 [5]. Many earlier workers, particularly Fisher, had carried this problem far, but one crucial ingredient of the theory (the power of a test) was missing until Neyman and Pearson supplied it.

1. Karl Pearson: "On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen in random sampling," *Philosophical Magazine*, Vol. 50 (1900), p. 157.
2. "Student" (W. S. Gosset): "The probable error of a mean," *Biometrika*, Vol. 6 (1908), p. 1.
3. R. A. Fisher: "The frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, Vol. 10 (1915), p. 507.
4. R. A. Fisher: "Applications of 'Student's' distribution," *Metron*, Vol. 5, No. 3 (1925), p. 90.
5. J. Neyman and E. S. Pearson: "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, Vol. 20A (1928), pp. 175 and 263.

12.12. Problems

1. Given the sample $(-0.2, -0.9, -0.6, 0.1)$ from a normal population with unit variance, test whether the population mean is zero at the .05 level of significance (i.e., with probability .05 of a Type I

error). Test whether the mean is zero at the .05 level relative to alternatives $\mu > 0$.

2. Given the sample $(-4.4, 4.0, 2.0, -4.8)$ from a normal population with variance four and the sample $(6.0, 1.0, 3.2, -0.4)$ from a normal population with variance five, test at the .01 level whether the means are equal relative to alternatives for which the mean of the first population is smaller than the mean of the second.

3. A metallurgist made four determinations of the melting point of manganese: 1269, 1271, 1263, 1265 degrees centigrade. Are these in accord with the published value of 1260 at the .05 level? (Assume normality.)

4. How would one make a two-sided test of $\mu = \mu_0$ for a normal population with known variance? Is this a uniformly most powerful test?

5. Plot the power function for two-sided tests of the null hypothesis $\mu = 0$ for a normal distribution with known variance using sample sizes 1, 4, 16, 64. (Use the standard deviation σ as the unit of measurement on the μ axis, and .05 probability of Type I error.)

6. What is the best critical region R in the sample space $(x_1, x_2, \cdots, x_n)$ for testing the null hypothesis that the mean is μ_0 against the alternative that the mean is μ_1 for a normal population?

7. Referring to Prob. 6, what would be the region for testing between two values of the variance, σ_0^2 and σ_1^2?

8. In testing between two values, μ_0 and μ_1, for the mean of a normal population, show that the probabilities for both types of error can be made arbitrarily small by taking a sufficiently large sample.

9. A cigarette manufacturer sent each of two laboratories presumably identical samples of tobacco. Each made five determinations of the nicotine content in milligrams as follows: (*a*) 24, 27, 26, 21, 24 and (*b*) 27, 28, 23, 31, 26. Were the two laboratories measuring the same thing? (Assume normality and a common variance.)

10. The metallurgist of Prob. 3, after assessing the magnitude of the various errors that might accrue in his experimental technique, decided that his measurements should have a standard deviation of about 2 degrees. Are the data consistent with this supposition at the .05 level? (Use a one-sided test, $\sigma > 2$.)

11. Test the hypothesis that the two samples of Prob. 9 came from populations with the same variance at the .05 level.

12. The power function for a test that the means of two normal populations are equal depends on the values of the two means, μ_1 and μ_2, and is therefore a surface. But the numerical value of the function

depends only on the difference $\theta = \mu_1 - \mu_2$, so that it can be adequately represented by a curve, say $P(\theta)$. Plot $P(\theta)$ when samples of four are drawn from one population with variance two, and samples of two are drawn from another population with variance three for tests at the .01 level.

13. Given the samples (1.8, 2.9, 1.4, 1.1), (5.0, 8.6, 9.2), (3.3, −4.1, 0.8) from normal populations, test whether the variances are equal at the .05 level.

14. Given a sample of size 100 with $\bar{x} = 2.7$ and $\Sigma(x_i - \bar{x})^2 = 225$, test the null hypothesis:

$$H_0: \mu = 3 \qquad \text{and} \qquad \sigma^2 = 2.5$$

at the .01 level assuming the population is normal.

15. Using the sample of Prob. 14, test the hypothesis that $\mu = \sigma^2$ at the .01 level.

16. Using the sample of Prob. 14, test at the .01 level whether the 95 per cent point α of the population distribution is three relative to alternatives $\alpha < 3$. The 95 per cent point is the number α such that $\int_{-\infty}^{\alpha} f(x)\,dx = .95$, where $f(x)$ is the population density; it is, of course, $\mu + 1.645\sigma$ in the present instance where the distribution is assumed to be normal.

17. Verify equations (8.5) and (8.6).

18. Verify equation (8.8).

19. Given the sample of Prob. 14 together with a sample from a second normal population of size 80 with $\bar{x} = 2.2$ and $\Sigma(x_i - \bar{x})^2 = 320$, test whether the means are equal at the .05 level. (The required root of the cubic equation encountered here is 2.56.)

20. In making two-sided tests of $\theta = \theta_0$, one does not ordinarily merely reject θ_0 when the test criterion falls in the critical region; he usually states that $\theta < \theta_0$ or that $\theta > \theta_0$ depending on which is indicated by the result of the test. In this situation there is a third error possible: one may declare $\theta < \theta_0$ when in fact $\theta > \theta_0$, or vice versa. Plot the probability of such a gross error as a function of $(\mu - \mu_0)/\sigma$ in the situation described in Prob. 4 for samples of size four and for probability .05 of a Type I error.

21. A sample of size n is drawn from each of k normal populations with the same variance. Derive the likelihood-ratio criterion for testing the hypothesis that the means are all zero. Show that criterion is a function of a ratio which has the F distribution.

22. Derive the likelihood-ratio criterion for testing whether the correlation of a bivariate normal distribution is zero.

23. If $x_1, x_2, \cdots, x_n$ are observations from normal populations with known variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$, how would one test whether their means were all equal?

24. A newspaper in a certain city observed that driving conditions were much improved in the city because the number of fatal automobile accidents in the past year was 9, whereas the average number per year over the past several years was 15. Is it possible that conditions were about the same as before? Assume number of accidents in a given year has a Poisson distribution.

25. Six 1-foot specimens of insulated wire were tested at high voltage for weak spots in the insulation. The numbers of such weak spots were found to be 2, 0, 1, 1, 3, 2. The manufacturer's quality standard states that there are less than 120 such defects per 100 feet. Does the batch from which these specimens were taken conform to the standard at the .05 level of significance? (Use the Poisson distribution.)

26. A psychiatrist newly employed by a medical clinic remarked at a staff meeting that about 40 per cent of all chronic headache sufferers were of the psychosomatic variety. His disbelieving colleagues mixed some pills of plain flour and water, giving them to all such patients on the clinic's rolls with the story that they were a new headache remedy and asking for comments. When the comments were all in, they could be fairly accurately classified as follows: (1) better than aspirin, 8; (2) about the same as aspirin, 3; (3) slower than aspirin, 1; (4) not worth the powder to blow them to hell, 29. While the doctors were somewhat surprised by these results, they nevertheless accused the psychiatrist of exaggeration. Did they have good grounds?

27. Supply the details of the argument in the last paragraph of Sec. 9.

28. A die was cast 300 times with the following results:

Occurrence	1	2	3	4	5	6
Frequency	43	49	56	45	66	41

Are the data consistent at the .05 level with the hypothesis that the die is true?

29. Of 64 offspring of a certain cross between guinea pigs, 34 were red, 10 were black, 20 were white. According to the genetic model these numbers should be in the ratio 9:3:4. Are the data consistent with the model at the .05 level?

30. A prominent baseball player's batting average dropped from .313 in one year to .280 in the following year. He was at bat 374 times during the first year and 268 times during the second. Is the hypothesis tenable at the .05 level that his hitting ability was the same during the two years?

31. Find the mean and variance of n_{ij} in the conditional distribution (10.17).

32. Show that the expected value of u defined by (10.20) is $n(r-1)(s-1)/(n-1)$ under the conditional distribution (10.17).

33. Using the data of Prob. 30, assume that one has a sample of 374 from one binomial population and 268 from another. Derive the λ criterion for testing whether the probability of a hit is the same for the two populations. How does this test compare with the ordinary test for a 2×2 contingency table?

34. The progeny of a certain mating were classified by a physical attribute into three groups, the numbers being 10, 53, 16. According to a genetic model the frequencies should be in the ratios $p^2 : 2p(1-p) : (1-p)^2$. Are the data consistent with the model at the .05 level?

35. A thousand individuals were classified according to sex and according to whether or not they were color-blind as follows:

	Male	Female
Normal	442	514
Color-blind	38	6

According to the genetic model these numbers should have relative frequencies given by

$$\begin{array}{ll} \frac{p}{2} & \frac{p^2}{2} + pq \\ \frac{q}{2} & \frac{q^2}{2} \end{array}$$

where $q = 1 - p$ is the proportion of defective genes in the population. Are the data consistent with the model?

36. Treating the table of Prob. 35 as a 2×2 contingency table, test the hypothesis that color blindness is independent of sex.

37. Gilby classified 1725 school children according to intelligence and apparent family economic level. A condensed classification follows:

	Dull	Intelligent	Very capable
Very well clothed	81	322	233
Well clothed	141	457	153
Poorly clothed	127	163	48

Test for independence at the .01 level.

38. A serum supposed to have some effect in preventing colds was tested on 500 individuals, and their records for 1 year were compared with the records of 500 untreated individuals as follows:

	No colds	One cold	More than one cold
Treated	252	145	103
Untreated	224	136	140

Test at the .05 level whether the sets of probabilities for the two trinomial populations may be regarded as the same.

39. Derive the general λ criterion for testing for independence in an $r \times s$ table when one set of marginal totals (the row totals, for example) are fixed in advance as in Prob. 38. Each row is regarded as a sample from an s-fold multinomial population with probabilities p_{ij} such that $\sum_j p_{ij} = 1$ for all i. The hypothesis of independence becomes: $p_{1j} = p_{2j} = p_{3j} = \cdots = p_{rj}$ for all j. How many degrees of freedom does $-2 \log \lambda$ have?

40. According to the genetic model the proportion of individuals having the four blood types should be related by:

$$\begin{aligned} \text{O}&: q^2 \\ \text{A}&: p^2 + 2pq \\ \text{B}&: r^2 + 2qr \\ \text{AB}&: 2pr \end{aligned}$$

where $p + q + r = 1$. Given the sample: O, 374; A, 436; B, 132; AB, 58; how would you test the correctness of the model?

41. Given cell frequencies n_{ijk} ($i = 1, 2, \cdots, r$; $j = 1, 2, \cdots, s$; $k = 1, 2, \cdots, t$) in a three-way classification, derive the criterion for testing whether all three criteria of classification are independent. How many degrees of freedom does $-2 \log \lambda$ have?

42. Galton investigated 78 families classifying children according to whether or not they were light-eyed, whether or not they had a light-eyed parent, whether or not they had a light-eyed grandparent. The following $2 \times 2 \times 2$ table resulted:

		Grandparent			
		Light		Not	
		Parent			
		Light	Not	Light	Not
Child	Light	1928	552	596	508
	Not	303	395	225	591

Test for complete independence at the .01 level. Test whether the child classification is independent of the other two classifications at the .01 level.

43. Derive the λ criterion for testing whether the i classification is independent of the jk classification in a three-way contingency table when the marginal totals $n_{i..}$ are fixed in advance. The probabilities satisfy the relations $\sum_{jk} p_{ijk} = 1$ for all i, and the null hypothesis is

$$p_{1jk} = p_{2jk} = \cdots = p_{rjk} \qquad \text{or simply} \qquad p_{ijk} = p_{jk}$$

How many degrees of freedom does $-2 \log \lambda$ have?

44. Derive the test for complete independence in the situation described in Prob. 43. The null hypothesis is $p_{ijk} = p_{ij}p_{k}$. How many degrees of freedom does $-2 \log \lambda$ have? How does this test compare with that for the case in which the $n_{i..}$ are not fixed in advance?

45. Compute the exact distribution of λ for a 2×2 contingency table with marginal totals $n_{1.} = 4$; $n_{2.} = 7$; $n_{.1} = 6$; $n_{.2} = 5$. What is the exact probability that $-2 \log \lambda$ exceeds 3.84, the .05 level of chi square for one degree of freedom?

CHAPTER 13

REGRESSION AND LINEAR HYPOTHESES

13.1. Families of Populations. In this chapter we shall study special cases of a situation which may be described as follows: A family of populations has a set of variates (which may be symbolized by x whether or not there is only one variate), a set of parameters θ, which are in general unknown, and a set of parameters z, which are usually observable and known for a given sample. The parameters θ may or may not be functions of the parameters z. If they are functions of z, the functions will in general be unknown. We shall consider the problem of making inferences about the parameters θ on the basis of samples drawn from populations with different values of z. The family of density functions may be represented by

$$f(x;\ \theta,\ z)$$

We shall select populations with known values of z and draw samples from each of these populations. Thus we shall deal with collections of samples: x_{1j} $(j = 1, 2, \cdots, n_1)$ for $z = z_1$; x_{2j} $(j = 1, 2, \cdots, n_2)$ for $z = z_2$; $\cdots$; x_{mj} $(j = 1, 2, \cdots, n_m)$ for $z = z_m$. We may, of course, draw only one observation from each population, in which case the observations could be represented by (x_1, z_1), (x_2, z_2), $\cdots$, (x_m, z_m). On the basis of such collections of observations on x and z, we may estimate certain of the parameters θ or test hypotheses about the parameters θ.

This general problem may be illustrated by considering the distribution of heights of individuals. A person's height may be expected to be related to his father's height z and his mother's height z'. Let us assume that for parents with given heights, childrens' adult heights will be normally distributed with means $\mu(z, z')$ and variances σ^2 independent of the z's, i.e., that heights x have densities

$$f(x;\ \mu,\ \sigma^2,\ z,\ z') = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(1/2\sigma^2)[x-\mu(z,z')]^2} \tag{1}$$

Here we have a variate x, a pair of parameters z, z' which can be observed, and a pair of unknown parameters μ and σ^2, one of which is

regarded as a function of the observable parameters. By measuring the heights of children and parents in several families with more than one child, we may, for example, test the hypothesis that the function form of $\mu(z, z')$ is

$$\mu = a + bz + cz' \tag{2}$$

where the a, b, c are unknown constants. If this hypothesis is acceptable, we may further wish to estimate the unknown parameters a, b, c.

To consider another example, the velocity of an object falling from rest in air may be expected to depend on the length of time t it has been falling, on its weight w, and on certain other parameters s specifying its size and shape. Again the distribution of velocities might be assumed to be normal with mean μ and variance σ^2, both of which may be functions of the observable parameters t, w, s. On the basis of a sample of observed velocities together with the corresponding values of the observable parameters, one might, for example, test certain hypotheses about the forms of the unknown functions $\mu(t, w, s)$ and $\sigma^2(t, w, s)$.

These problems are *regression* problems. They are sometimes referred to as prediction problems. Thus in the first example, after the parameters a, b, c, and σ^2 are estimated, one may predict with about 95 per cent certainty that the children of a couple with given heights z_0, z_0' would have heights between

$$a + bz_0 + cz_0' - 1.96\sigma \qquad \text{and} \qquad a + bz_0 + cz_0' + 1.96\sigma$$

if the estimates were based on a large sample. The accuracy of a prediction depends largely on the size of the prediction interval which in the present instance depends on the error variance σ^2. In the case of a falling body, the error variance is so small under certain conditions that the velocity can be predicted almost exactly (the length of a 95 per cent prediction interval is small enough to be negligible for most practical purposes). In the case of predicting heights of children, the prediction interval would not be small relative to the mean $a + bz_0 + cz_0'$.

Regression problems occur in great variety in all sciences, both natural and social. In fact, from one point of view the whole aim of science in general is to predict (on the basis of past experimental work) what will happen in a given circumstance.

We shall be concerned with a special case of the general regression problem which, however, has very wide application. We shall deal with normal distributions in which the mean is a function of the

observable parameters. The variance of the normal distribution will be assumed to be independent of the observable parameters. The mean $\mu(z)$, where z is the set of observable parameters, is called the *regression function;* the function would represent a curve if z consisted of one parameter, a surface if z consisted of two parameters, a hypersurface for more than two parameters.

13.2. Simple Linear Normal Regression. A variate x is normally distributed about a regression function which is linear in a single observable parameter; the variance is independent of that parameter. The density is

$$f(x;\, \alpha,\, \beta,\, \sigma^2,\, z) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(1/2\sigma^2)[x-(\alpha+\beta z)]^2} \tag{1}$$

We shall deal with the one-parameter family of normal distributions for which α, β, σ^2 are fixed. The family is represented in Fig. 65; for any given value of z, x is normally distributed with mean $\alpha + \beta z$ and variance σ^2.

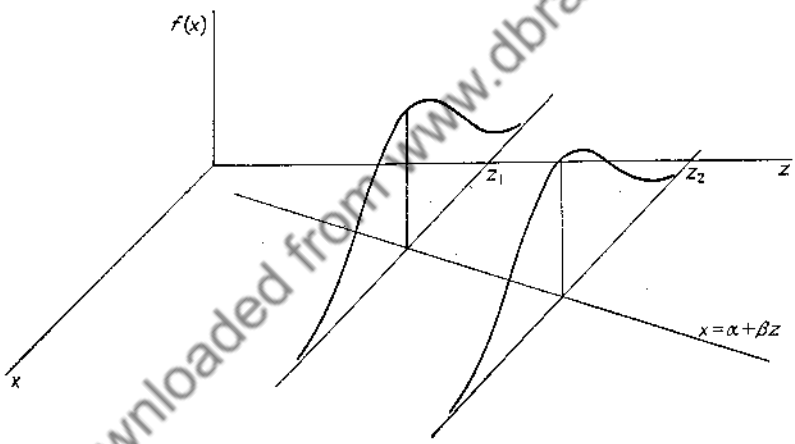


FIG. 65.

We shall consider first the estimation of α, β, σ^2. Let (x_i, z_i), $i = 1, 2, \cdots, n$, be a sample of x's together with the corresponding values of z. Some of the z_i may be equal, as would be the case if more than one x value were drawn from any specific distribution. It is convenient to label the z's differently even when some of them are the same. It is necessary that there be at least two different values of z however. Obviously one cannot expect to estimate α and β from a sample drawn from a single member of the family of distribu-

tions. The method of maximum likelihood will be employed to estimate the parameters. The likelihood is

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2\sigma^2)[x_i-(\alpha+\beta z_i)]^2} \tag{2}$$

and its logarithm is

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_i [x_i - (\alpha + \beta z_i)]^2 \tag{3}$$

On putting the derivatives of this expression with respect to α, β, σ^2 equal to zero, we obtain the relations

$$n\sigma^2 = \Sigma(x_i - \alpha - \beta z_i)^2 \tag{4}$$

$$\Sigma(x_i - \alpha - \beta z_i) = 0 \tag{5}$$

$$\Sigma z_i(x_i - \alpha - \beta z_i) = 0 \tag{6}$$

which must be solved for the unknown parameters. The last two equations are called the *normal equations* which determine the coefficients in a linear regression function. They are linear in α and β and therefore readily solved. We shall let

$$\bar{x} = \frac{1}{n}\sum x_i \qquad \bar{z} = \frac{1}{n}\sum z_i \tag{7}$$

The solutions of (5), (6), and (7) may then be written

$$\hat{\beta} = \frac{\Sigma(x_i - \bar{x})(z_i - \bar{z})}{\Sigma(z_i - \bar{z})^2} \tag{8}$$

$$\hat{\alpha} = \bar{x} - \beta\bar{z} \tag{9}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \hat{\alpha} - \hat{\beta} z_i)^2 \tag{10}$$

which are the required point estimators of the unknown parameters. We notice that the solution could not be carried through if all the z_i were equal because the denominator of (8) would vanish.

Distribution of the Estimators. Since $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the x_i (which are normally distributed), it follows that $\hat{\alpha}$ and $\hat{\beta}$ must themselves have a bivariate normal distribution. One could specify that distribution by simply finding the means, variances, and covariance of the $\hat{\alpha}$ and $\hat{\beta}$. We shall, however, find the distribution another way. The main objective is to show that $\hat{\alpha}$ and $\hat{\beta}$ are distributed

independently of $\hat{\sigma}^2$, and in doing this their distribution will fall out incidentally.

We shall evaluate the joint-moment generating function:

$$m(s_1, s_2, s_3) = E\left(e^{s_1\frac{\hat{\alpha}-\alpha}{\sigma}+s_2\frac{\hat{\beta}-\beta}{\sigma}+s_3\frac{n\hat{\sigma}^2}{\sigma^2}}\right) \tag{11}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{s_1\frac{\hat{\alpha}-\alpha}{\sigma}+s_2\frac{\hat{\beta}-\beta}{\sigma}+s_3\frac{n\hat{\sigma}^2}{\sigma^2}-\frac{1}{2\sigma^2}\sum(x_i-\alpha-\beta z_i)^2} \prod dx_i \tag{12}$$

for the three variates $(\hat{\alpha} - \alpha)/\sigma$, $(\hat{\beta} - \beta)/\sigma$, and $n\hat{\sigma}^2/\sigma^2$. The first step in evaluating the integral is to transform the variates x_i to

$$y_i = \frac{1}{\sigma}(x_i - \alpha - \beta z_i) \tag{13}$$

this removes the factor $1/\sigma^n$ from the integrand and changes the exponent in the integrand of (12) to

$$\sum_{i=1}^{n} c_i y_i - \tfrac{1}{2} \sum_{i,j=1}^{n} \sigma^{ij} y_i y_j \tag{14}$$

where

$$c_i = \frac{s_1[(\Sigma z_i^2/n) - \bar{z}z_i] + s_2(z_i - \bar{z})}{\Sigma(z_i - \bar{z})^2} = a_i s_1 + b_i s_2 \tag{15}$$

and

$$\sigma^{ij} = \delta_{ij}(1 - 2s_3) + 2s_3[na_i a_j + n\bar{z}(a_i b_j + a_j b_i) + b_i b_j \Sigma z_i^2] \tag{16}$$

where the a_i and b_i are defined by (15) and δ_{ij} is one or zero according as i is or is not equal to j. We have then to evaluate an integral of the form

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{2\pi}\right)^{n/2} e^{\Sigma c_i y_i - \frac{1}{2}\Sigma\Sigma\sigma^{ij} y_i y_j} \prod dy_i \tag{17}$$

which, apart from a factor $\sqrt{|\sigma^{ij}|}$, is just the integral in equation (9.6.1) with the ξ's put equal to zero. The value of that integral is given in (9.6.4), and it follows that

$$m(s_1, s_2, s_3) = e^{\frac{1}{2}\Sigma\Sigma\sigma_{ij}c_i c_j} \sqrt{|\sigma^{ij}|} \tag{18}$$

The algebraic reduction of (18) may be accomplished as follows: Since

$$a_i + \bar{z}b_i = \frac{1}{n}$$

equation (16) may be written

$$\sigma^{ij} = \delta_{ij}(1 - 2s_3) + 2s_3\left[b_i b_j \sum (z_i - \bar{z})^2 + \frac{1}{n}\right] \tag{19}$$

or

$$\sigma^{ij} = \delta_{ij}(1 - 2s_3) + 2s_3\left(d_i d_j + \frac{1}{n}\right) \tag{20}$$

where

$$d_i = \frac{z_i - \bar{z}}{\sqrt{\Sigma(z_i - \bar{z})^2}} \tag{21}$$

so that $\Sigma d_i = 0$ and $\Sigma d_i^2 = 1$. It is not difficult to verify then that

$$|\sigma^{ij}| = (1 - 2s_3)^{n-2} \tag{22}$$

and that the elements of the inverse of the matrix $\|\sigma^{ij}\|$ are

$$\sigma^{ij} = \frac{\delta_{ij}}{1 - 2s_3} - \frac{2s_3}{1 - 2s_3}\left(d_i d_j + \frac{1}{n}\right) \tag{23}$$

These last two relations enable one to put (18) in the form

$$m(s_1, s_2, s_3) = \frac{e^{[1/2\Sigma(z_i-\bar{z})^2](s_1^2(1/n)\Sigma z_i^2 - 2\bar{z}s_1s_2 + s_2^2)}}{(1 - 2s_3)^{(n-2)/2}} \tag{24}$$

The form of the moment generating function (24) enables one to draw several important conclusions. Remembering that s_1 is associated with $(\hat{\alpha} - \alpha)/\sigma$, s_2 with $(\hat{\beta} - \beta)/\sigma$, s_3 with $n\hat{\sigma}^2/\sigma^2$, we observe

1. That the pair of variates $\hat{\alpha}$ and $\hat{\beta}$ are distributed independently of $\hat{\sigma}^2$ because $m(s_1, s_2, s_3)$ factors into a function of s_3 alone and a function of s_1 and s_2 alone (see Sec. 10.4). We shall let

$$m(s_1, s_2, s_3) = m_1(s_1, s_2)m_2(s_3) \tag{25}$$

2. That the functional form of $m_1(s_1, s_2)$ is that of the moment generating function for a bivariate normal distribution (Sec. 9.6); hence $\hat{\alpha}$ and $\hat{\beta}$ are jointly normally distributed with means α and β, respectively, and variances and covariances

$$\sigma_{\hat{\alpha}}^2 = \frac{\sigma^2 \Sigma z_i^2}{n\Sigma(z_i - \bar{z})^2} \tag{26}$$

$$\sigma_{\hat{\beta}}^2 = \frac{\sigma^2}{\Sigma(z_i - \bar{z})^2} \tag{27}$$

$$\text{cov}(\alpha, \beta) = -\frac{\sigma^2 \bar{z}}{\Sigma(z_i - \bar{z})^2} \tag{28}$$

The inverse of the matrix of these variances and covariances is

$$\begin{Vmatrix} n/\sigma^2 & n\bar{z}/\sigma^2 \\ n\bar{z}/\sigma^2 & \Sigma z_i^2/\sigma^2 \end{Vmatrix} \tag{29}$$

which are the coefficients of the quadratic form in the distribution of $(\hat{\alpha} - \alpha)$ and $(\hat{\beta} - \beta)$.

3. That $\hat{\alpha}$ and $\hat{\beta}$ will be independently distributed if the z_i are chosen so that $\bar{z} = 0$.

4. That the quadratic form of the joint distribution of $\hat{\alpha}$ and $\hat{\beta}$,

$$Q = \frac{1}{\sigma^2}\left[n(\hat{\alpha} - \alpha)^2 + 2n\bar{z}(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) + \sum z_i^2(\hat{\beta} - \beta)^2\right] \tag{30}$$

has the chi-square distribution with two degrees of freedom.

5. That $m_2(s_3)$ is the moment generating function for a chi-square distribution with $n - 2$ degrees of freedom, hence that $n\hat{\sigma}^2/\sigma^2$ has that distribution (Sec. 10.3).

Confidence Regions and Tests of Hypothesis. In regression problems the main interest is usually in the regression coefficients α and β. Of course there is no trouble in estimating σ^2 or in testing hypotheses about σ^2, because the chi-square distribution of 5 above provides confidence intervals and tests directly.

To obtain a confidence interval for α, we need only to observe that the marginal distribution of $\hat{\alpha}$ is normal with mean α and variance given by (26); hence

$$u = \frac{(\hat{\alpha} - \alpha)}{\sigma}\sqrt{\frac{n\,\Sigma(z_i - \bar{z})^2}{\Sigma z_i^2}}$$

has a normal distribution with zero mean and unit variance. Since u and $n\hat{\sigma}^2/\sigma^2$ are independently distributed, it follows from Sec. 10.6 that

$$\begin{aligned} t &= \frac{u}{\sqrt{n}\,\hat{\sigma}/\sqrt{n-2}} \\ &= (\hat{\alpha} - \alpha)\sqrt{\frac{n(n-2)\,\Sigma(z_i - \bar{z})^2}{\Sigma z_i^2\,\Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2}} \end{aligned} \tag{31}$$

has the t distribution with $n - 2$ degrees of freedom. Since α is the only unknown quantity in this expression, the inequalities in

$$P(-t_\epsilon < t < t_\epsilon) = 1 - \epsilon$$

may be converted to obtain a confidence interval with fiducial probability $1 - \epsilon$ for α. The quantity t also provides a test criterion for

testing hypotheses about α in just the same way it does for the mean of a normal distribution (Sec. 12.6). Thus to test whether the regression line $x = \alpha + \beta z$ passes through the origin in the x, z plane, we should simply put $\alpha = 0$ in (31) and observe whether $|t| < t_\epsilon$ if the level of significance is to be ϵ. One-tailed tests may also be made.

Confidence intervals for β and tests on β may be made in a quite similar way. It is readily seen that

$$t = (\hat{\beta} - \beta)\sqrt{\frac{(n-2)\Sigma(z_i - \bar{z})^2}{\Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2}} \tag{32}$$

also has the t distribution with $n - 2$ degrees of freedom and involves only the unknown parameter β. To test, for example, whether the means of the family of normal distributions under consideration were independent of the observable parameter, one would put $\beta = 0$ in (32) and observe whether $|t| < t_\epsilon$, where ϵ is the chosen significance level.

For simultaneous estimation of α and β, we may use the fact that

$$F = \frac{Q}{n\hat{\sigma}^2/\sigma^2} \tag{33}$$

where Q is defined by (30), has the F distribution with 2 and $n - 2$ degrees of freedom (section 10.5), and involves only the unknown parameters α and β. The inequality in

$$P(F < F_\epsilon) = 1 - \epsilon$$

is readily seen to define an elliptical confidence region in the α, β plane for α and β. To test whether α and β had certain specified values α_0 and β_0, one would put $\alpha = \alpha_0$ and $\beta = \beta_0$ in (33) and observe whether or not the resulting value of F exceeded F_ϵ.

All these tests on α and β could have been obtained by the likelihood-ratio method.

It is worth observing that the accuracy of the estimation of α and β depends on the choice of the z_i. Thus the variance of $\hat{\alpha}$ will be as small as possible when the z_i are chosen so that $\bar{z} = 0$. For, since

$$\Sigma(z_i - \bar{z})^2 = \Sigma z_i^2 + n\bar{z}^2$$

the least possible value for $\sigma_{\hat{\alpha}}^2$ (equation 26) is σ^2/n and occurs when $\bar{z} = 0$. Evidently the confidence interval for α will be shortest on the average for given n when $\bar{z} = 0$. The variance of $\hat{\beta}$ (equation 27) can evidently be made small by choosing widely separated values for

the z_i. In fact, if z_1 is the smallest practicable value of z and z_2 is the largest, then β will be best estimated when all the sampling is done at those two values of z. It often happens in practice, however, that there is some doubt about the linearity of the regression function $\alpha + \beta z$, and it is desired to test for linearity. In this case it is necessary to have observations for more than two values of z. A test for linearity will be described in Sec. 14.2.

13.3. Prediction. Let us suppose that a linear regression function $x = \alpha + \beta z$ has been estimated by $x = \hat{\alpha} + \hat{\beta} z$ on the basis of a sample of n observations. We now wish to predict the value of x for some specified value of z, say z_0. Thus if x is son's adult height and z is father's height, a sample of observations will provide estimates $\hat{\alpha}$ and $\hat{\beta}$ for a linear regression function. A prospective father of height z_0 may wish to predict his son's height. The predicted height is, of course, $x_0 = \hat{\alpha} + \hat{\beta} z_0$. Or to consider a different problem: Let x be the demand for some commodity, and let z be the wholesale price of the item two months earlier, or the wholesale price of some ingredient or part of the item two months earlier. It is desired to predict the demand two months in advance of the present. From past records one may collect a set of pairs of observations (x_i, z_i), where x_i is the demand at a given time and z_i is the wholesale price two months previous to that time, and estimate coefficients α and β of a linear regression. If z_0 is the present wholesale price, then the predicted demand two months hence is $x_0 = \hat{\alpha} + \hat{\beta} z_0$.

The worth of a prediction depends on the magnitude of its possible error, and we shall take account of that error by obtaining a *prediction interval* which is analogous to a confidence interval. The variate x is a random variable with a normal distribution having mean $\alpha + \beta z_0$ and variance σ^2. The predicted value $x_0 = \hat{\alpha} + \hat{\beta} z_0$ has two sources of error: in the first place $\hat{\alpha} + \hat{\beta} z_0$ is merely an estimate of the mean of x, and the actual value of x may, of course, deviate from its mean; in the second place the estimated mean is subject to the random sampling errors inherent in $\hat{\alpha}$ and $\hat{\beta}$. If α, β, and σ were exactly known, then a 95 per cent prediction interval for x would simply be

$$\alpha + \beta z_0 - 1.96\sigma \qquad \text{to} \qquad \alpha + \beta z_0 + 1.96\sigma$$

since the probability that x will fall within 1.96σ of its mean is .95 for a normal distribution. Since all these parameters except z_0 are unknown, we must attempt to set up an interval in terms of their estimates.

The variate

$$u = x - \hat{\alpha} - \hat{\beta}z_0 \tag{1}$$

is necessarily normally distributed since it is a linear function of the normal variates x, $\hat{\alpha}$, $\hat{\beta}$. The distribution of u is therefore known when its mean and variance are given. Since

$$E(x) = \alpha + \beta z_0 \qquad E(\hat{\alpha}) = \alpha \qquad E(\hat{\beta}) = \beta$$

we have

$$E(u) = 0$$

The variance of u is therefore

$$\begin{aligned} \sigma_u^2 &= E(u^2) \\ &= E(x - \hat{\alpha} - \hat{\beta}z_0)^2 \\ &= \sigma_x^2 + \sigma_{\hat{\alpha}}^2 + z_0^2\sigma_{\hat{\beta}}^2 + 2z_0E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)] \end{aligned} \tag{2}$$

remembering that x is independent of $\hat{\alpha}$ and $\hat{\beta}$. σ_x^2 is simply σ^2, the variance of the normal distribution, and the other terms in (34) are given by (26), (27), (28), so that

$$\begin{aligned} \sigma_u^2 &= \sigma^2\left[1 + \frac{(1/n)\Sigma z_i^2 + z_0^2 - 2z_0\bar{z}}{\Sigma(z_i - \bar{z})^2}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(z_0 - \bar{z})^2}{\Sigma(z_i - \bar{z})^2}\right] \\ &= \sigma^2\left[\frac{n+1}{n} + \frac{(z_0 - \bar{z})^2}{\Sigma(z_i - \bar{z})^2}\right] \end{aligned} \tag{3}$$

A 95 per cent prediction interval for u is just $-1.96\sigma_u$ to $1.96\sigma_u$, but this still involves one unknown parameter σ which appears in σ_u. We can eliminate σ by using the t distribution. The variate u/σ_u is normally distributed with zero mean and unit variance and is distributed independently of $n\hat{\sigma}^2/\sigma^2$; hence

$$t = \frac{u/\sigma_u}{\sqrt{n\hat{\sigma}^2/(n-2)\sigma^2}} \tag{4}$$

has the t distribution with $n - 2$ degrees of freedom and involves no unknown parameters. The inequalities in

$$P(-t_\epsilon < t < t_\epsilon) = 1 - \epsilon$$

may be converted to determine a $100(1 - \epsilon)$ per cent prediction inter-

val for x. The interval is given by

$$P(\hat{\alpha} + \hat{\beta}z_0 - A < x < \hat{\alpha} + \hat{\beta}z_0 + A) = 1 - \epsilon \tag{5}$$

where

$$A = t_\epsilon \sigma \sqrt{\frac{n}{n-2}\left[\frac{n+1}{n} + \frac{(z_0 - \bar{z})^2}{\Sigma(z_i - \bar{z})^2}\right]} \tag{6}$$

Several properties of the prediction interval should be observed:

1. The length of the interval is greater than $2t_\epsilon\sigma$ on the average regardless of how large a sample was used to estimate α and β. This is entirely reasonable because we are predicting a single observation x which is normally distributed with standard deviation σ.

2. The average length of the prediction interval increases as z_0 moves away from $\bar{z}$. If it is possible, the values z_i chosen for obtaining observations to estimate the parameters should be selected so as to have a mean value near z_0.

3. The relation (5) holds only for a single prediction based on the estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$. One cannot use the estimated regression to make several predictions and expect (5) to remain true. The relation has meaning only if α, β, σ are reestimated each time a prediction on x is made. The probability statement takes account of sampling variation in the estimates as well as in x, and if the original estimates are used repeatedly (not allowed to vary), the statement cannot be effective.

It is easy to generalize the above technique to take account of prediction of the mean of a sample of size m observed for $z = z_0$. Let x_1', x_2', $\cdots$, x_m' be a sample of m observations at z_0 with mean $\bar{x}'$. The mean of

$$v = \bar{x}' - \hat{\alpha} - \hat{\beta}z_0$$

is zero, and its variance σ^2 is the same as (3) except that $(n + 1)/n$ is replaced by $(1/m) + (1/n)$. The variate

$$t = \frac{v/\sigma_v}{\sqrt{n\hat{\sigma}^2/(n-2)\sigma^2}}$$

has the t distribution with $n - 2$ degrees of freedom and involves no unknown parameters; hence it may be employed to construct a prediction interval for $\bar{x}'$.

13.4. Discrimination. The discrimination problem is an estimation problem and is in a sense the reverse of the prediction problem. In prediction one wishes to predict x knowing z_0 on the basis of estimates of α, β, σ. In discrimination one wishes to estimate z_0 having observed

x. The general class of biological assay problems are of this character. Thus, for example, the concentration of a certain vitamin may be measured by observing the gain in weight of a week-old chick when its diet is augmented by daily doses of the vitamin for several days. A manufacturer of the vitamin might determine the strength of a new batch as follows: Let x be the gain in weight and let z be the concentration. Using material of known concentration, he would feed several chicks with different concentrations z_i $(i = 1, 2, \cdots, n)$ and observe their gains in weight x_i. At the same time other chicks would receive their vitamins from the batch with unknown concentration z_0, and their gains in weight, say x'_j $(j = 1, 2, \cdots, m)$, would be observed. On the basis of these data it is desired to estimate the parameter z_0.

The general problem of classification is a discrimination problem. Anthropologists, for example, make measurements x on skulls of known age z, then estimate the age z_0 of a skull of unknown age with measurements x'. Taxonomists use the technique to discriminate between varieties of plants with quite similar appearance.

Using the notation of the first paragraph and the model of Sec. 2, the likelihood of the observations $x_1, x_2, \cdots, x_n$ and $x'_1, x'_2, \cdots, x'_m$ is

$$L = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{m+n} e^{-(1/2\sigma^2)\Sigma(x_i-\alpha-\beta z_i)^2-(1/2\sigma^2)\Sigma(x_j'-\alpha-\beta z_0)^2} \tag{1}$$

and on differentiating the logarithm of this expression with respect to σ^2, α, β, z_0 in turn, one can readily determine the maximum-likelihood estimates of these parameters; they are

$$\hat{\beta} = \frac{\Sigma(x_i - \bar{x})(z_i - \bar{z})}{\Sigma(z_i - \bar{z})^2} \tag{2}$$

$$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{z} \tag{3}$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[\sum (x_i - \hat{\alpha} - \hat{\beta}z_i)^2 + \sum (x'_j - \bar{x}')^2\right] \tag{4}$$

$$z_0 = \frac{\bar{x}' - \hat{\alpha}}{\hat{\beta}} \tag{5}$$

where

$$\bar{x} = \frac{1}{n}\sum x_i \qquad \bar{z} = \frac{1}{n}\sum z_i \qquad \bar{x}' = \frac{1}{m}\sum x'_j$$

Equations (2) and (3) are the same as (2.8) and (2.9); equation (5) gives the desired point estimate of z_0.

A confidence interval for z_0 is also easily set up. The quantity

$$v = \bar{x}' - \hat{\alpha} - \hat{\beta} z_0 \tag{6}$$

is normally distributed since it is a linear function of normal variates; its mean is zero, and its variance is

$$\sigma_v^2 = \sigma^2 \left[\frac{1}{m} + \frac{1}{n} + \frac{(z_0 - \bar{z})^2}{\Sigma(z_i - \bar{z})^2} \right] \tag{7}$$

just as was found in Sec. 3. The two sums in (4) both have chi-square distributions when they are divided by σ^2, the first with $n - 2$ and the second with $m - 1$ degrees of freedom. The two chi squares are independent since they are functions of independent samples; hence their sum has the chi-square distribution with $m + n - 3$ degrees of freedom. Furthermore the two chi squares are obviously independent of v. It follows then that

$$t = \frac{v/\sigma_v}{\sqrt{(m + n)\hat{\sigma}^2/(m + n - 3)\sigma^2}} \tag{8}$$

has the t distribution and will provide a confidence interval for z_0 since that is the only unknown parameter which appears in (8).

We have considered a very much simplified discrimination problem, but it is one which occurs frequently in practice. The more general problem has to do with the case in which each observation consists of several components $(x_1, x_2, \cdots, x_k)$ which have a multivariate normal distribution with means $\alpha_1 + \beta_1 z$, $\alpha_2 + \beta_2 z$, $\cdots$, $\alpha_k + \beta_k z$. Given estimates of the α's and β's on the basis of a sample of observations $(x_{1i}, x_{2i}, \cdots, x_{ki})$, one wishes to estimate z_0 for an observation $(x_{10}, x_{20}, \cdots, x_{k0})$. We shall have to omit this problem because it is very cumbersome to handle by elementary methods.

13.5. Multiple Regression. We shall consider now a variate x which is normally distributed with variance σ^2 and with a mean of the form $\alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_k z_k$; the z's are observable parameters, and we are concerned with the other parameters (the α's and σ^2). We may wish to estimate the parameters or test certain hypotheses about the parameters. The density for a sample of size n is

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-(1/2\sigma^2)\sum_i \left(x_i - \sum_p \alpha_p z_{pi}\right)^2} \tag{1}$$

and the logarithm of the likelihood is

$$L = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i\left(x_i - \sum_p \alpha_p z_{pi}\right)^2 \tag{2}$$

We shall let the indices i and j run from 1 to n, and the indices p, q, r, and s run from 1 to k. On differentiating L with respect to α_q, we find that the $\hat{\alpha}$'s are determined by the following set of k *normal equations* (there being an equation for each value of q):

$$\sum_i z_{qi}\left(x_i - \sum_p \hat{\alpha}_p z_{pi}\right) = 0 \tag{3}$$

If we define a_{pq} and y_q by the relations

$$a_{pq} = \sum_i z_{pi}z_{qi}$$

$$y_q = \sum_i x_i z_{qi}$$

the normal equations may be written

$$\sum_p a_{pq}\hat{\alpha}_p = y_q \tag{4}$$

The matrix of coefficients $\|a_{pq}\|$ may be inverted if its determinant does not vanish, and letting a^{pq} represent the elements of the inverse matrix, the solution of (4) for the $\hat{\alpha}$'s may be written

$$\hat{\alpha}_p = \sum_q a^{pq}y_q \tag{5}$$

as follows by multiplying both sides of (4) by a^{qr} and summing on q (see Sec. 9.2). The maximum-likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i\left(x_i - \sum_p \hat{\alpha}_p z_{pi}\right)^2 \tag{6}$$

as follows from putting the derivative of L with respect to σ^2 equal to zero and substituting the $\hat{\alpha}$'s for the α's.

Distributions and Confidence Regions. In considering the distribution of the estimators, we observe that the a_{pq} are not functions of the random variables x_i and that the y_q are linear functions of normally distributed variates and must therefore be normally distributed. We may determine the distribution of the $\hat{\alpha}_p$ by simply finding their means

variances and covariances. The mean is

$$\begin{aligned} E(\hat{\alpha}_p) &= E\left(\sum_q a^{pq} y_q\right) \\ &= \sum_q a^{pq} \sum_i z_{qi} E(x_i) \\ &= \sum_q a^{pq} \sum_i z_{qi} \sum_r \alpha_r z_{ri} \\ &= \sum_q \sum_r a^{pq} a_{qr} \alpha_r \\ &= \sum_r \delta_{pr} \alpha_r \\ &= \alpha_p \end{aligned} \tag{7}$$

The covariance of $\hat{\alpha}_p$ and $\hat{\alpha}_q$ is

$$\begin{aligned} E(\hat{\alpha}_p - \alpha_p)(\hat{\alpha}_q - \alpha_q) &= E(\hat{\alpha}_p \hat{\alpha}_q) - \alpha_p \alpha_q \\ &= E\left(\sum_{r,i} a^{pr} z_{ri} x_i\right)\left(\sum_{s,j} a^{qs} z_{sj} x_j\right) - \alpha_p \alpha_q \\ &= E \sum_{i,j} \left(\sum_r a^{pr} z_{ri}\right)\left(\sum_s a^{qs} z_{sj}\right) E(x_i x_j) - \alpha_p \alpha_q \end{aligned} \tag{8}$$

When $i \neq j$,

$$E(x_i x_j) = \left(\sum_u \alpha_u z_{ui}\right)\left(\sum_v \alpha_v z_{vj}\right)$$

where u and v run from 1 to k, and when $i = j$,

$$E(x_i^2) = \left(\sum_u \alpha_u z_{ui}\right)^2 + \sigma^2$$

On substituting these values in (8) and making reductions similar to those employed to obtain (7), one finds

$$E[(\hat{\alpha}_p - \alpha_p)(\hat{\alpha}_q - \alpha_q)] = a^{pq}\sigma^2 \tag{9}$$

The inverse of the matrix $\|a^{pq}\sigma^2\|$ is $\|a_{pq}/\sigma^2\|$; hence the $\hat{\alpha}$'s have the density

$$\left(\frac{1}{2\pi}\right)^{k/2} \left|\frac{a_{pq}}{\sigma^2}\right|^{1/2} e^{-(1/2\sigma^2)\sum_{p,q} a_{pq}(\hat{\alpha}_p - \alpha_p)(\hat{\alpha}_q - \alpha_q)} \tag{10}$$

It can also be shown that $n\hat{\sigma}^2/\sigma^2$ has the chi-square distribution with $n - k$ degrees of freedom and further that $n\hat{\sigma}^2/\sigma^2$ is distributed independently of the $\hat{\alpha}$'s. We shall omit the argument, which is somewhat complicated but entirely analogous to that used in Sec. 2 to obtain

the joint distribution of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$ in the case $k = 2$. From these facts it follows that any particular regression coefficient α_p may be estimated by a confidence interval using the t distribution; $\hat{\alpha}_p - \alpha_p$ is normally distributed with zero mean and variance $a^{pp}\sigma^2$; hence

$$t = \frac{\hat{\alpha}_p - \alpha_p}{\sqrt{a^{pp}n\hat{\sigma}^2/(n-k)}} \tag{11}$$

has the t distribution with $n - k$ degrees of freedom and involves no unknown parameters except α_p. A confidence region for the whole set of regression coefficients, $\alpha_1, \alpha_2, \cdots, \alpha_k$, in a k-dimensional space may be determined by the inequality in

$$P(F < F_{1-\beta}) = \beta$$

where $F_{1-\beta}$ is the critical level for the F distribution with k and $n - k$ degrees of freedom. The quadratic form in the exponent of (10) has the chi-square distribution with k degrees of freedom and is distributed independently of $n\hat{\sigma}^2/\sigma^2$; hence

$$F = \frac{(n-k)\Sigma a_{pq}(\hat{\alpha}_p - \alpha_p)(\hat{\alpha}_q - \alpha_q)}{kn\hat{\sigma}^2} \tag{12}$$

has the F distribution with k and $n - k$ degrees of freedom.

It may be instructive to compare the results obtained thus far in this section with those of Sec. 2 by putting $k = 2$, $z_1 = 1$, and identifying α_1, α_2, z_2 with α, β, z, respectively.

Prediction. Given estimates of the parameters α_p and σ in (1), one may predict the value of x corresponding to a given set of values, z_{0p}, of the observable parameters. The predicted value would of course be

$$x_0 = \sum_p \hat{\alpha}_p z_{0p} \tag{13}$$

The prediction interval is set up by considering the variate

$$u = x - \Sigma\hat{\alpha}_p z_{0p}$$

which is normally distributed as it is a linear function of normally distributed variates. The mean of u is zero since both x and $\Sigma\hat{\alpha}_p z_{0p}$ have expected value $\Sigma\alpha_p z_{0p}$. The variance of u is

$$\sigma_u^2 = E(u^2) \tag{14}$$

$$= E(x - \Sigma\alpha_p z_{0p})^2 + E[\Sigma(\hat{\alpha}_p - \alpha_p)z_{0p}]^2 \tag{15}$$

since x is independent of the $\hat{\alpha}$'s. The first term on the right of (15) is σ^2, and the second term is readily evaluated by means of (9). One finds

$$\sigma_u^2 = \sigma^2 \left(1 + \sum_{p,q} a^{pq} z_{0p} z_{0q}\right) \tag{16}$$

Thus the variate

$$t = \frac{u/\sigma_u}{\sqrt{n\hat{\sigma}^2/(n-k)\sigma^2}} \tag{17}$$

has the t distribution with $n - k$ degrees of freedom (u being independent of $\hat{\sigma}^2$) and may be employed to define a prediction interval for x since it involves no unknown parameters.

13.6. Linear Hypotheses. Referring to the multiple regression let us consider how we might test the hypothesis that the regression coefficients α_p have certain specified values α_{0p}. The null hypothesis is

$$H_\omega\colon \alpha_p = \alpha_{0p} \ (p = 1, 2, \cdots, k) \qquad \text{and} \qquad \sigma^2 > 0 \tag{1}$$

and the alternatives are

$$H_\Omega\colon -\infty < \alpha_p < \infty \ (p = 1, 2, \cdots, k) \qquad \text{and} \qquad \sigma^2 > 0 \tag{2}$$

The subspace ω has one dimension, while Ω has $k + 1$ dimensions. If the likelihood (5.1) is maximized in ω and in Ω, one finds the λ criterion, after considerable algebraic reduction, to be

$$\lambda = \frac{1}{[1 + [k/(n-k)]F]^{n/2}} \tag{3}$$

where F is the quantity in (5.12) with the α_p replaced by α_{0p}. Hence the λ test is equivalent to an F test, and large values of F correspond to small values of λ; the null hypothesis would be tested by using the right-hand tail of the F distribution for the critical region. When the α_{0p} are zero, as is often the case, the double sum in the numerator of F may be reduced to the simple form, $\Sigma\hat{\alpha}_p y_p$, by substituting for $\hat{\alpha}_q$ from (5.5).

A more commonly desired test is one which tests some but not all the regression coefficients. Let us suppose that we wish to test whether the coefficients $\alpha_1, \alpha_2, \cdots, \alpha_m$ $(m \leq k)$ have specified values α_{0u} $(u = 1, 2, \cdots, m)$ whatever the values of the last $k - m$ of the α's. The null hypothesis is now

$$H_\omega\colon -\infty < \alpha_r < \infty \ (r = m + 1, \cdots, k)$$
$$\alpha_u = \alpha_{0u} \ (u = 1, \cdots, m) \qquad \sigma^2 > 0 \tag{4}$$

while H_0 is as specified in (2). We shall merely present the test without the derivation because of the complex algebraic reduction required. It becomes plausible if one considers the marginal distribution of the α_u $(u = 1, \cdots, m)$; this distribution is obtained by integrating out $\alpha_{m+1}, \alpha_{m+2}, \cdots, \alpha_k$ from (5.10). After the integration there will remain a multivariate normal distribution, and the coefficients of the quadratic form will be, say, b_{uv}/σ^2. The b_{uv} are obtained (Sec. 9.2) by striking out the last $k - m$ rows and columns of a^{pq} and inverting the result; i.e.,

$$\|b_{uv}\| = \|a^{uv}\|^{-1} \qquad u, v = 1, 2, \cdots, m$$

The quadratic form of the marginal distribution is

$$Q = \frac{\sum_{u,v} b_{uv}(\hat{\alpha}_u - \alpha_u)(\hat{\alpha}_v - \alpha_v)}{\sigma^2} \tag{5}$$

and it has the chi-square distribution with m degrees of freedom. Since Q is distributed independently of $\hat{\sigma}^2$, the quantity

$$F' = \frac{Q/m}{n\hat{\sigma}^2/(n-k)\sigma^2} \tag{6}$$

has the F distribution with m and $n - k$ degrees of freedom. The λ criterion for testing (4) turns out to be

$$\lambda = \frac{1}{\{1 + [m/(n-k)]F'\}^{n/2}} \tag{7}$$

with the α_0's substituted for the α's in Q; hence F' provides an equivalent test.

We are now in a position to consider what is called the *general linear hypothesis* of normal regression theory. The problem is to test the hypothesis that the coefficients α_p satisfy certain linear relations, say:

$$\begin{aligned} c_{11}\alpha_1 + c_{12}\alpha_2 + \cdots + c_{1k}\alpha_k &= c_{01} \\ c_{21}\alpha_1 + c_{22}\alpha_2 + \cdots + c_{2k}\alpha_k &= c_{02} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ c_{m1}\alpha_1 + c_{m2}\alpha_2 + \cdots + c_{mk}\alpha_k &= c_{0m} \end{aligned}$$

where $m \leq k$ and the c's are given numbers. These equations may be written

$$\sum_p c_{up}\alpha_p = c_{0u} \qquad p = 1, 2, \cdots, k \qquad u = 1, 2, \cdots, m \tag{8}$$

We suppose that these m relations are independent, i.e., that it is not possible to obtain one of them by adding chosen multiples of the others.

The null hypothesis that (8) is true may be reduced to the form of (4) by recasting the problem in terms of new parameters, say $\beta_1, \beta_2, \cdots, \beta_k$, and new observable parameters, say $w_1, w_2, \cdots, w_k$. The first m of the β's are defined by putting

$$\sum_p c_{up}\alpha_p = \beta_u \tag{9}$$

The independence of the relations (8) ensures that m of the α's can be solved for in terms of the remaining α's and the β_u. Supposing the equations can be solved for the first m of the α's, the solutions are simply

$$\alpha_u = \sum_v c^{uv}\left(\beta_v - \sum_r c_{vr}\alpha_r\right) \tag{10}$$

where u and v run from 1 to m and r runs from $m + 1$ to k, and where the c^{uv} are the elements of the inverse of $\|c_{uv}\|$. The remaining β's may be put equal to the remaining α's:

$$\alpha_r = \beta_r \qquad r = m + 1, \cdots, k \tag{11}$$

These new parameters β_p are now substituted for the α_p in the mean of x:

$$\sum_p \alpha_p z_p = \sum_u \left[\sum_v c^{uv}\left(\beta_v - \sum_r c_{vr}\beta_r\right)\right] z_u + \sum_r \beta_r z_r \tag{12}$$

The new observable parameters are then taken to be the coefficients of the β's in (12); i.e., w_p is the coefficient of β_p in (12):

$$\begin{aligned} w_p &= \sum_u c^{up} z_u & p &= 1, 2, \cdots, m \\ &= z_p - \sum_{u,v} c^{uv} c_{vp} z_u & p &= m + 1, \cdots, k \end{aligned} \tag{13}$$

The mean of x is now expressed in the form $\Sigma\beta_p w_p$. The null hypothesis becomes simply $\beta_u = c_{0u}$ $(u = 1, 2, \cdots, m)$, the one already discussed as (4).

13.7. Applications of Normal Regression Theory. The estimation and test procedures we have just developed have a very wide range of application. The reason for this is the completely arbitrary nature of what we have called the observable parameters. The z_p may, for example, be artificial code variables. Thus, suppose in a fertilizer experiment to investigate the effect of nitrogen and potash on a given

crop, the crop is grown on plots with different fertilizer treatments. We may express the mean yield in the form $\sum_1^4 \alpha_\mu z_\mu$. Let $z_1 = 1$ for all plots; let z_2 be zero for those plots with no nitrogen and one for all plots with a given application of nitrogen; let z_3 be zero for plots with no potash and one for those with potash; and let z_4 be zero for all plots except those treated with both fertilizers. Now α_1 represents the yield with no fertilizer, α_2 the added yield due to nitrogen, α_3 the added yield due to potash, $\alpha_2 + \alpha_3 + \alpha_4$ the added yield due to both fertilizers. Having performed the experiment, we may estimate the α's, and we may test various hypotheses. Thus to test whether potash has any effect, we set up the null hypothesis that it does not and test whether α_3 and α_4 are both zero. To test whether effects of nitrogen and potash are strictly additive (that there is no *interaction* between nitrogen and potash), we would test whether $\alpha_4 = 0$.

In another instance the z_μ may represent functions of some variable. As an example, we may consider a time series. The average monthly prices of some agricultural product, eggs, for example, if plotted against time over a period of years, will show rather erratic looking fluctuations but will have certain inherent regularities. There will be a *trend* of some kind—a smooth curve which may be thought of as representing the general character of the variation of price with time apart from any fluctuations. Also there will be an annual cycle of sorts; the prices in a given year will usually be higher during the winter months than the summer months. A firm which stores eggs in large quantity may wish to know, for example, whether the amplitude of the cycle is independent of the average price level from year to year. This question might be studied as follows: Let x be the price, and let t represent time in months. The data consist of prices $x_1, x_2, \cdots, x_n$ at times $t = 1, 2, \cdots, n$. Over the period of time included, let us suppose it is apparent that a quadratic function will fit the trend quite well enough. Then the following regression function might reasonably represent the trend and cycle if the null hypothesis (that the amplitude is constant) is true:

$$\alpha_1 + \alpha_2 t + \alpha_3 t^2 + \alpha_4 \sin \frac{2\pi t}{12} + \alpha_5 \cos \frac{2\pi t}{12}$$

If the null hypothesis is not true, the amplitude might reasonably be supposed to be proportional to the general price level given by the trend, or more generally, to be some linear or quadratic function of the

time. To take account of this possibility, terms like

$$\alpha_6 t \sin \frac{2\pi t}{12} + \alpha_7 t \cos \frac{2\pi t}{12} + \alpha_8 t^2 \sin \frac{2\pi t}{12} + \alpha_9 t^2 \cos \frac{2\pi t}{12}$$

would be added to the function given above. The z_p are now defined by $z_1 = 1$, $z_2 = t$, $\cdots$, $z_9 = t^2 \cos (2\pi t/12)$. The null hypothesis would be tested by testing whether the last four regression coefficients were zero.

The observable parameters may be any functions of any number of variables. Thus, for example, a variate x may be known to be some function of two variables u and v, but the form of the function, say $f(u, v)$, may be unknown, and the purpose of the experiment may be to investigate the form of the function in the neighborhood of some point (u_0, v_0). It may be reasonable to suppose that the function can be adequately represented in this neighborhood by a quadratic function, i.e., by the first six terms of its series expansion:

$$\begin{aligned} f(u_0, v_0) &+ f_u(u_0, v_0)(u - u_0) + f_v(u_0, v_0)(v - v_0) \\ &+ \tfrac{1}{2}[f_{uu}(u_0, v_0)(u - u_0)^2 + 2f_{uv}(u_0, v_0)(u - u_0)(v - v_0) \\ &\qquad + f_{vv}(u_0, v_0)(v - v_0)^2] \end{aligned}$$

where the subscripts indicate partial differentiation. One would merely estimate the α's in $\sum_1^6 \alpha_p z_p$, where $z_1 = 1$, $z_2 = u - u_0$, $z_3 = v - v_0$, $z_4 = (u - u_0)^2$, $z_5 = (u - u_0)(v - v_0)$, $z_6 = (v - v_0)^2$. If one wished to test the adequacy of the quadratic representation, cubic terms might be included in the regression function.

13.8. The Method of Least Squares. There is a general problem of curve fitting which is entirely unrelated to normal regression theory but which may be solved by formulas identical with those we have obtained for estimating regression coefficients.

Suppose some variable x is a function $f(z)$ of another variable z and that the function has been investigated by measuring x for certain chosen values of z. The result might be as shown in Fig. 66. There may be no question of random variation. The value x_1 measured at z_1 might be exactly the same if it were determined a second time. The function is simply not smooth. But for purposes for which the function is to be used, one may wish to approximate it by a smooth function, say a straight line. How might such an approximating line be drawn? One might simply lay a transparent ruler along the points and draw a line which fits pretty well, and this method may be as good

as any for the purposes at hand. Or one might divide the points into two groups, the left-hand four and right-hand four, and compute the averages of the x and z values for the two groups. The averages $\bar{x}$ and $\bar{z}$ for one group will determine one point, and the averages $\bar{\bar{x}}$ and $\bar{\bar{z}}$ of the other group will determine a second point which, together with the first, determines an approximating line. There are many possibilities.

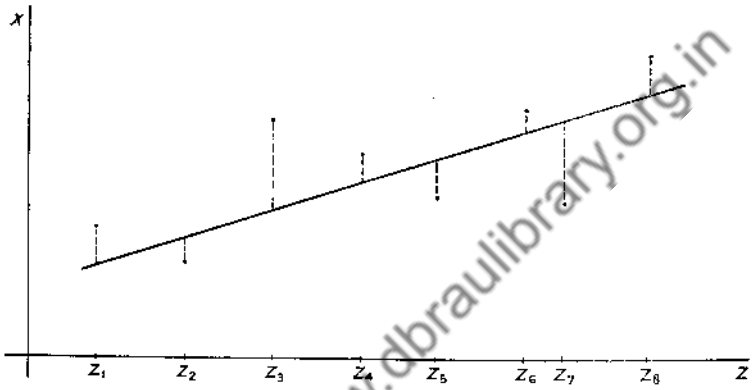


FIG. 66.

The problem is generally solved by what is called the *method of least squares*. This method chooses that line, $x = \alpha + \beta z$, which minimizes the sum of squares of the vertical deviations of the points from the line. Supposing now that there are n points (z_i, x_i) $(i = 1, 2, \cdots, n)$ and that we denote the ordinate of the point on the line at z_i by x_i', the vertical deviations are $x_i - x_i'$ and their sum of squares is, say,

$$S = \sum_i (x_i - x_i')^2 = \sum_i (x_i - \alpha - \beta z_i)^2$$

We wish to fix the line (determine α and β) so that S will be minimized. This would be done by setting the partial derivatives of S with respect to α and β equal to zero and solving for α and β. The resulting equations are the same as (2.5) and (2.6).

More generally, any empirical function $x_i = f(u_i, v_i, \cdots, w_i)$ $(i = 1, 2, \cdots, n)$ may be approximated by any linear combination $\sum_{p=1} \alpha_p z_p$ of known functions z_p of the variates $u, v, \cdots, w$ by the method of least squares. One would choose the α's so as to minimize the sum of squares of the deviations of the x_i from $x_i' = \sum_p \alpha_p z_{pi}$; i.e.,

one would minimize

$$S = \sum_i \left(x_i - \sum_p \alpha_p z_{pi}\right)^2$$

with respect to the α's and find that they were determined by the relations (5.3).

The primary reason that the method of least squares is commonly used for curve fitting is merely that it leads to a simple linear system of equations for determining the coefficients. To determine the coefficients by minimizing, say, the sum of the absolute deviations, or the sum of the fourth powers of the deviations, would ordinarily be much more troublesome. It just happens that the form of the normal distribution is such that the sum of squares of deviations from the regression function is to be minimized to determine the coefficients in the regression function. If, for example, the points in Fig. 66 were supposed to be deviations from a regression line with a probability distribution other than a normal distribution, then it would be appropriate to determine estimates of α and β by maximizing the likelihood defined by that distribution. Even here, though, the method of least squares is commonly used in practice to avoid algebraic and arithmetic difficulties, and this is, of course, good and sufficient reason. The theoretical advantages of the principle of maximum likelihood over the principle of least squares may become unimportant when it comes to a matter of choosing, say, between a 40-hour and a 10-hour computation.

13.9. Notes and References. A more complete account of the theory of regression may be found in Chap. VIII of Wilks' book [1]. In particular, the proof of the important result that $\hat{\sigma}^2$ is distributed independently of the $\hat{\alpha}$'s is given there. The notation of Secs. 5 and 6 has been made quite similar to that of Wilks in order to facilitate reference to that proof and to others which are omitted here.

There is a great body of literature on a subject which we have omitted entirely. A special case of normal regression theory of particular interest arises if one considers the conditional distribution of, say, x_1 in a k-variate normal distribution; it is normal with a mean which is a linear function of the other variates, $x_2, x_3, \cdots, x_k$. The coefficients of these variates (corresponding to what we have called α_p) are certain functions of the variances and covariances of the original multivariate normal distribution. Estimation of these coefficients implies estimation of certain correlations and partial correlations. There is an elaborate theory associated with this sort of correlation analysis which was once regarded as a very essential part of statistics.

In recent years it has come to be realized that most (though not all) correlation problems which arise in practice can be handled more appropriately by regression methods. The latter require only the assumption that deviations from the regression function be normal, whereas the correlation analysis requires that the variate and what we have called the observable parameters all be jointly normally distributed. A good account of correlation analysis is given by Kendall [2].

A rather complete treatment of the theory of least squares and its various applications may be found in [3]. In [4] are treated a great variety of practical problems in regression and correlation analysis.

1. S. S. Wilks: "Mathematical Statistics," Princeton University Press, Princeton, N. J., 1943.
2. M. G. Kendall: "Advanced Theory of Statistics," Vol. 1, Charles Griffin & Co., Ltd., London, 1944.
3. W. E. Deming: "Statistical Adjustment of Data," John Wiley & Sons, Inc., New York, 1943.
4. M. Ezekiel: "Methods of Correlation Analysis," John Wiley & Sons, Inc., New York, 1930.

13.10. Problems

1. Verify equations (2.22) and (2.23).

2. Derive the likelihood-ratio criterion for testing the null hypothesis that the parameter α of Sec. 2 has the value α_0.

3. Verify equations (3.3) and (3.6).

4. Verify equations (2), (3), (4), and (5) of Sec. 4.

5. Verify equation (5.9).

6. Verify equation (6.3).

7. Verify equation (6.7).

8. Given the data:

x	−6.1	−0.5	7.2	6.9	−0.2	−2.1	−3.9	3.8	−7.5	−2.1
z	−2.0	0.6	1.4	1.3	0.0	−1.6	−1.7	0.7	−1.8	−1.1

fit a regression line assuming x is normally distributed about a linear function of z, and find a 95 per cent confidence interval for the coefficient of z.

9. Plot the regression line of Prob. 8 and plot two curves showing the 95 per cent limits of prediction intervals for x in the range $-3 < z < 3$.

10. Plot a 95 per cent confidence region for the two regression parameters of Prob. 8.

11. Given the data:

x	12.1	11.9	10.2	8.0	7.7	5.3	7.9	7.8	5.5	2.6
z_1	0	1	2	3	4	5	6	7	8	9
z_2	7	4	4	6	4	2	1	1	1	0

fit a regression plane, and find a 95 per cent confidence interval for σ^2.

12. Find a 95 per cent confidence interval for α_1 of Prob. 11.

13. Test the null hypothesis that α_2 of Prob. 11 is zero.

14. What is the 95 per cent prediction interval for x at $z_1 = 2.5$, $z_2 = 2.5$ in Prob. 11?

15. Test the null hypothesis that $\alpha_1 + 10\alpha_2 = 0$ in Prob. 11.

16. Using only the first two rows of the data of Prob. 11, fit a regression function of the form

$$\alpha_0 + \alpha_1 z_1 + \alpha_2 z_1^2$$

and test the null hypothesis that $\alpha_2 = 0$.

17. The fitting of polynomials such as the quadratic of Prob. 16 is much simplified when the values are equally spaced by using *orthogonal polynomials*. Let $z = 0, 1, \cdots, n$. The first three orthogonal polynomials are

$$P_1 = z - \frac{n}{2}$$

$$P_2 = \left(z - \frac{n}{2}\right)^2 - \frac{n(n+2)}{12}$$

$$P_3 = \left(z - \frac{n}{2}\right)^3 - \frac{2n(n+2) - 4}{20}\left(z - \frac{n}{2}\right)$$

Show that

$$\sum_z P_1 P_2 = \sum_z P_1 P_3 = \sum_z P_2 P_3 = 0$$

18. Rework Prob. 16, fitting instead the regression function

$$\alpha_0 + \alpha_1 P_1 + \alpha_2 P_2$$

where P_1 and P_2 are defined in Prob. 17.

19. If x_1 and x_2 have a bivariate normal distribution, what are the

coefficients (in terms of σ_{11}, σ_{22}, and ρ) of the regression function for the conditional distribution of x_1? For the conditional distribution of x_2? If the two regression lines were estimated from the same sample, would they, in general, be different?

20. If x_1, x_2, x_3 have a trivariate normal distribution, what are the coefficients of the regression function for the conditional distribution of x_1, given x_2 and x_3, in terms of the variances and correlations?

21. If the correlation ρ of a bivariate normal distribution is zero, show that its estimator $\hat{\rho}$ has the density

$$\frac{[(n-3)/2]!(1-\hat{\rho}^2)^{(n-4)/2}}{\sqrt{2\pi}\,[(n-4)/2]!}$$

for samples of size n.

22. Referring to Prob. 21, transform $\hat{\rho}$ to a new variate

$$t = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}}$$

showing that it has "Student's" distribution with $n-2$ degrees of freedom so that the t tables may be used for testing the null hypothesis $\rho = 0$.

23. Assume that the data of Prob. 8 are from a bivariate normal population and test the null hypothesis that $\rho = 0$.

24. When ρ is not zero, the distribution of $\hat{\rho}$ is not a simple function, but it has been tabulated for n, the sample size, less than 25. For larger n, Fisher has shown that

$$z = \frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}}$$

is approximately normally distributed with mean

$$\xi = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$$

and variance $1/(n-3)$. Using this result, estimate roughly a 95 per cent confidence interval for ρ of Prob. 23.

25. Derive the λ criterion given in equation (6.3).

26. What is the maximum-likelihood estimator of the multiple correlation coefficient $R_{1.23}$ (defined in Prob. 27 of Chap. 9).

27. A variate x is distributed about a linear regression function, $\alpha + \beta z$, by the density

$$f(x) = 1 \qquad \alpha + \beta z - \tfrac{1}{2} < x < \alpha + \beta z + \tfrac{1}{2}$$

Find the maximum-likelihood estimate of the regression function, given the sample of four points (x, z): $(0.3, 1)$, $(-0.6, 2)$, $(-1.7, 3)$, $(-1.8, 4)$. Compare it with the least-squares line.

28. A variate x is distributed about $\alpha + \beta z$ by the density

$$\begin{aligned} f(x) &= \tfrac{1}{2}e^{-(x-\alpha-\beta z)} & x > \alpha + \beta z \\ &= \tfrac{1}{2}e^{x-\alpha-\beta z} & x < \alpha + \beta z \end{aligned}$$

Estimate the regression function given the sample of four points (x, z): $(3.4, 1)$, $(7.1, 2)$, $(12.4, 3)$, $(15.5, 4)$. Compare it with the least-squares line.

29. A normal variate x has mean $\alpha + \beta z$ and variance σ^2. The parameter z can take only the values zero and one. Set up a test of the hypothesis that $\beta = 0$ and compare it with the test of the equality of means of two normal populations with the same variance. (If the two means are μ_1 and μ_2, let $\alpha = \mu_1$ and $\beta = \mu_2 - \mu_1$.)

30. Referring to the situation described in the first paragraph of Sec. 7, set up a test for the null hypothesis $\alpha_2 = 0$. Assume that there are $4n$ observations, there being n for each of four treatments: no fertilizer, nitrogen, potash, both nitrogen and potash.

CHAPTER 14

EXPERIMENTAL DESIGNS AND THE ANALYSIS OF VARIANCE

14.1. Experimental Design. The general subject of experimental design is too broad to be included with any degree of completeness in this book. It comprises the processes of planning experiments, analyzing the results, and interpreting the results. We are primarily concerned with the last-mentioned problem, which, in so far as statistics is involved, is a matter of statistical inference. The technique for making inferences is known as the *analysis of variance*, and it is that technique which will be studied in this chapter. In order to motivate the study, it will be instructive, however, to consider briefly some of the general aspects of experimental design.

An experiment is intended to find out something about the relation between two or more variables. For example, one may wish to discover the effect of carbon content (one variable) on the hardness (second variable) of steel; the effect of a drug in preventing colds; the value of paint in preserving wood; the effect on flavor of meat caused by cold storage; and so on. Any experiment may be thought of as an investigation of a function of two or more variables. As we have noted in the first chapter, some variables may be entirely unwanted but must in the nature of things be involved in the experiment. In the terminology of experimental design, one variable may be called the *subject* of the experiment while the other variables are called *factors*. Thus, carbon content is a factor which affects the hardness of steel (the subject of the experiment); freezing (a factor) affects the flavor of meat (the subject).

In planning experiments, one has on the one hand certain *principles of experimental design*, and on the other a large class of geometrical configurations, specific *experimental designs*. In accordance with the principles, one fits a specific design to the projected experiment.

In the course of this chapter we shall illustrate some of the principles and give examples of a few very simple designs. But first we may observe two important principles of design which are largely matters of common sense and experience. The first is: every possible outcome of the experiment must be anticipated and a conclusion decided upon

for each possible outcome in advance of performing the experiment. For example, suppose a man claims he can read his wife's mind to the extent that he can very often tell whether she is looking at a red or black playing card. To test this contention, the following experiment is to be performed: His wife is to look at cards drawn one by one from an ordinary deck, and the man is to say in each instance whether it is red or black. If the whole deck is to be used, there are 53 possible outcomes; he may call 0, 1, 2, $\cdots$, 52 of the cards correctly. And let us suppose it is agreed to accept his claim if 40 or more are called correctly and to reject the claim if 39 or less are called correctly. This simple experiment is now completely designed in the sense that the conclusion is only a matter of performing the experiment, observing the number correctly called, and adopting the appropriate conclusion. If it turned out, for example, that 30 cards were called correctly, among them 12 of the spades, the man might argue that he had demonstrated his ability because the probability of calling 12 spades correctly under the assumption of random calling is so very small as to make that assumption absurd. This argument is not valid because any set of 30 cards can be found to have some peculiarity which would make it highly improbable under random sampling. (In particular, of course, the probability of drawing any specified set of 30 cards is $1\Big/\binom{52}{30} \cong 10^{-14}$ for random selection of 30 cards without replacement.) Any inference from experimental data cannot be supported by a fiducial probability statement unless that inference was taken account of in advance of the performance of the experiment. Any seemingly significant but unforeseen inference can only suggest a new experiment. It follows, of course, that an experimenter who does not anticipate any inferences at all but merely waits to see what will turn up in the data, cannot support any conclusion whatever by a fiducial probability statement.

The second broad principle we wish to mention specifically is this: there must be an element of randomization in the experiment. An experiment is performed to test a hypothesis, or to estimate a parameter or a set of parameters. The hypothesis adopted is supported by odds based on a computation which assumes random sampling under a null hypothesis. The parameter is estimated by a confidence interval with a fiducial probability determined by the assumption of randomness. It is quite evident that the results of an experiment cannot be supported by probability statements unless the sampling was in fact random. Referring to the card-calling experiment described above,

the null hypothesis is that the man has not any ability to call the cards correctly. The probability of calling 40 or more cards correctly is roughly .0001 under the assumption of random calling, and the null hypothesis would be emphatically rejected if 40 or more were called correctly, provided random sampling is operative under the null hypothesis. The proper condition obtains if the cards are presented in a random order (by thoroughly shuffling the deck, for example), for then the result of the experiment will have a random sampling distribution under any system of calling which is independent of the actual sequence of colors of the cards. (It is tacitly assumed here that red and black will be called in about equal numbers, that one will not call all 52 cards black, for example.) One could, of course, present the cards in some order particularly devised perhaps to confuse the caller, and the caller might nevertheless be quite successful and establish his ability beyond reasonable doubt, but one could not measure his success in probability terms. Statistical inference is impossible in nonrandomized experiments.

It has been found in practice that persons cannot be relied upon to write down random sets of numbers at will. Randomization in experimental design must be carried out by actually tossing coins, casting dice, drawing numbered chips from a bowl, or the like. Specially prepared tables of random numbers have been published to save experimenters the trouble of performing these operations.

14.2. Analysis of Variance in Regression. The analysis of variance is a technique for testing linear hypotheses, and basically it is just the technique described in the preceding chapter. All we shall do in this chapter is study that technique in more detail and investigate simplifications that can be made in applying the technique to certain special problems that arise frequently in practice. The point of view, however, will be somewhat different, and to illustrate it, we return to the simple linear regression problem.

Let us suppose that a variate x is normally distributed about a regression function $\alpha + \beta z$ with variance σ^2. A sample of size n is observed: $(x_1, z_1), (x_2, z_2), \cdots, (x_n, z_n)$. Let $\hat{\alpha}$ and $\hat{\beta}$ be defined by equations (13.2.8) and (13.2.9). The sum of squares of deviations from the true regression will be divided into two parts as follows:

$$\begin{aligned}\Sigma(x_i - \alpha - \beta z_i)^2 &= \Sigma(x_i - \hat{\alpha} - \hat{\beta} z_i + \hat{\alpha} + \hat{\beta} z_i - \alpha - \beta z_i)^2 \\ &= \Sigma(x_i - \hat{\alpha} - \hat{\beta} z_i)^2 \\ &\quad + 2\Sigma(x_i - \hat{\alpha} - \hat{\beta} z_i)(\hat{\alpha} + \hat{\beta} z_i - \alpha - \beta z_i) \\ &\quad + \Sigma(\hat{\alpha} + \hat{\beta} z_i - \alpha - \beta z_i)^2 \qquad (1)\end{aligned}$$

The middle sum on the right of (1) vanishes identically, as may be seen by performing the summation and using the definitions of $\hat{\alpha}$ and $\hat{\beta}$. The first sum on the right of (1) is the sum of squares of deviations from the estimated regression function; it is just $n\hat{\sigma}^2$ where $\hat{\sigma}^2$ is the maximum-likelihood estimate of σ^2 defined in Sec. 13.2. The third sum on the right of (1) is, apart from a division σ^2, the quadratic form (13.2.30) in the distribution of $\hat{\alpha}$ and $\hat{\beta}$. The total sum of squares on the left of (1), on division by σ^2, has the chi-square distribution with n degrees of freedom; it has been partitioned into two parts which are independently distributed by chi-square distributions—one with $n - 2$ degrees of freedom and the other with two degrees of freedom.

The third sum on the right of (1) may be further partitioned into two parts each of which are independently distributed by chi-square laws with one degree of freedom. It is apparent from (13.2.30) that $\hat{\alpha}$ and $\hat{\beta}$ are not independently distributed except in the special case in which $\bar{z} = 0$. However, $\bar{x}$ and $\hat{\beta}$ are independently distributed, as may be seen by changing the variable $\hat{\alpha}$ to $\bar{x}$ using the substitution

$$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{z} \tag{2}$$

in the joint distribution of $\hat{\alpha}$ and $\hat{\beta}$. In fact, $\bar{x}$ and $\hat{\beta}$ are independently normally distributed. In terms of these variables, the third sum of (1) is

$$\begin{aligned}
\Sigma(\hat{\alpha} + \hat{\beta}z_i - \alpha - \beta z_i)^2 &= \Sigma(\bar{x} - \hat{\beta}\bar{z} + \hat{\beta}z_i - \alpha - \beta z_i)^2 \\
&= \Sigma[(\bar{x} - \alpha - \beta\bar{z}) + (\hat{\beta} - \beta)(z_i - \bar{z})]^2 \\
&= \Sigma(\bar{x} - \alpha - \beta\bar{z})^2 + \Sigma[(\hat{\beta} - \beta)(z_i - \bar{z})]^2 \qquad (3) \\
&= n(\bar{x} - \alpha - \beta\bar{z})^2 + (\hat{\beta} - \beta)^2\Sigma(z_i - \bar{z})^2 \qquad (4)
\end{aligned}$$

The sum of cross products has been omitted in (3) because it is readily seen to vanish since $\Sigma(z_i - \bar{z}) = 0$. The two terms on the right of (4), apart from a factor $-2\sigma^2$, are just the exponents in the univariate normal distributions of $\bar{x}$ and $\hat{\beta}$; hence they are independently distributed by chi-square laws with one degree of freedom.

The total sum of squares of deviations has now been partitioned into three parts:

$$\begin{aligned}
\Sigma(x_i - \alpha - \beta z_i)^2 = \Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2 &+ (\hat{\beta} - \beta)^2\Sigma(z_i - \bar{z})^2 \\
&+ n(\bar{x} - \alpha - \beta\bar{z})^2 \qquad (5)
\end{aligned}$$

each of which is independently distributed by chi-square laws. We turn now to the question of testing whether α and β differ from zero.

If, in particular, α and β are put equal to zero throughout (5), we have

$$\Sigma x_i^2 = \Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2 + \hat{\beta}^2\Sigma(z_i - \bar{z})^2 + n\bar{x}^2 \tag{6}$$

All these terms are directly calculable from the data, and in the analysis of variance, this partition of the sum of squares is usually exhibited in a table such as the one given here. In a particular problem, the entries in the table would all be numerical.

ANALYSIS OF VARIANCE FOR SIMPLE LINEAR REGRESSION

Source	Sum of squares	Degrees of freedom	Mean square	F ratio
Mean	$n\bar{x}^2$	1	$n\bar{x}^2$	$\dfrac{n(n-2)\bar{x}^2}{\Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2}$
Slope	$\hat{\beta}^2\Sigma(z_i - \bar{z})^2$	1	$\hat{\beta}^2\Sigma(z_i - \bar{z})^2$	$\dfrac{(n-2)\hat{\beta}^2\Sigma(z_i - \bar{z})^2}{\Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2}$
Deviations	$\Sigma(x_i - \hat{\alpha} - \hat{\beta}z_i)^2$	$n - 2$	$\dfrac{1}{n-2}\sum(x_i - \hat{\alpha} - \hat{\beta}z_i)^2$	
Total	Σx_i^2	n		

Now let us consider the null hypothesis that $\beta = 0$. If it is true, then the sums of squares in the second and third lines of the table are independently distributed by chi-square laws with 1 and $n - 2$ degrees of freedom (on division by σ^2), and the ratio of the mean squares will have the F distribution with 1 and $n - 2$ degrees of freedom. This is exactly the test given by (13.2.32) because the square of a t variate with k degrees of freedom has the F distribution with 1 and k degrees of freedom (Sec. 10.6). The sum-of-squares entry in the second line of the table is said to be the portion of the total sum of squares Σx_i^2 associated with β.

Now let us turn to the first line of the table. The F ratio in the first line provides a test for the null hypothesis, $\alpha = 0$, *only if it is assumed that* $\beta = 0$ (unless $\bar{z}$ happens to be zero). Thus the two F tests indicated in the right-hand column of the table are of two different kinds. The second one tests

$$\beta = 0, \text{ whatever } \alpha \text{ may be}$$

the first one tests

$$\alpha = 0, \text{ provided } \beta \text{ is actually zero}$$

These statements are evident on comparing (5) and (6). The first term on the right of (6) has the chi-square distribution whatever α and β may be; the second term has the chi-square distribution whatever α may be provided only that $\beta = 0$; the third term has the chi-square distribution only if $\alpha + \beta\bar{z} = 0$.

The two tests on α and β are said to be *nonorthogonal*. If it had been possible to partition the two degrees of freedom for α and β into two single degrees of freedom, one involving α only and one involving β only in such a way that they were independently distributed, then we should have had *orthogonal* tests of α and β and could test $\alpha = 0$ whatever β might be.

If in collecting the data, the values of z are chosen so that $\bar{z} = 0$, then orthogonal tests of α and β are available. For then $\hat{\alpha}$ becomes equal to $\bar{x}$, and in fact the F test indicated in the first line of the table becomes equivalent to the t test given by equation (13.2.31). It is to be recalled, of course, that we can test $\alpha = 0$ without assuming $\beta = 0$ by using that t test.

The condition of orthogonality is regarded as desirable because it provides a partial measure of statistical independence in tests. Suppose $\bar{z} = 0$; then the two tests of $\alpha = 0$ and $\beta = 0$ are still not statistically independent because the two F ratios have the same denominator. If one worked out the joint distribution of the two ratios, he would find that they are not independently distributed. But the fact that the two numerators of the F ratios are independently distributed has some intuitive appeal. It is usually impossible to design experiments so as to get completely independent tests, but it is often possible to design them so as to get orthogonal tests. Thus in the present example, one can investigate a regression function $\alpha + \beta z$ by means of the two t tests described in Sec. 13.2, and these tests are nonorthogonal in general; it may be possible, however, to select z values so that $\bar{z} = 0$ and thus obtain orthogonal tests.

From the practical point of view, orthogonality is quite desirable because the analysis of data is usually very much simpler for orthogonal than for nonorthogonal designs.

Test of Linearity. Before leaving the linear regression problem we shall consider one other test which is quite useful when the data are such that it is feasible. Suppose that for one or more of the z values there are two or more x observations. More precisely, let there be k

distinct values of z—$z_1, z_2, \cdots, z_k$—and let the x observations be denoted by x_{st} where $s = 1, 2, \cdots, k$ and $t = 1, 2, \cdots, n_s$. Corresponding to z_s, there are thus n_s x observations, and we assume that not all the n_s are one. Letting $n = \sum_s n_s$, we may relabel the x_{st}, calling them $x_1, x_2, \cdots, x_n$ and perform the analysis already described. The deviations from the fitted regression may be written

$$\sum_i (x_i - \hat{\alpha} - \hat{\beta}z_i)^2 = \sum_{st} (x_{st} - \hat{\alpha} - \hat{\beta}z_s)^2 \tag{7}$$

in the x_{st} notation; the z_s are all distinct, while the z_i are not, with the data under present consideration.

The right-hand side of (7) will now be partitioned into two parts as follows:

$$\begin{aligned}\sum_{st} (x_{st} - \hat{\alpha} - \hat{\beta}z_s)^2 &= \sum_{st} (x_{st} - \bar{x}_s + \bar{x}_s - \hat{\alpha} - \hat{\beta}z_s)^2 \\ &= \sum_{st} (x_{st} - \bar{x}_s)^2 + \sum_{st} (\bar{x}_s - \hat{\alpha} - \hat{\beta}z_s)^2 \\ &= \sum_{st} (x_{st} - \bar{x}_s)^2 + \sum_s n_s(\bar{x}_s - \hat{\alpha} - \hat{\beta}z_s)^2 \end{aligned} \tag{8}$$

where $\bar{x}_s = \sum_t x_{st}/n_s$. The first sum on the right has the chi-square distribution with $\sum_s (n_s - 1) = n - k$ degrees of freedom, *whatever the regression function may be.* For, for fixed z, x is normally distributed with variance σ^2, and the sample $x_{11}, x_{12}, \cdots, x_{1n_1}$ of n_1 observations for $z = z_1$ provides a sum of squares $\sum_t (x_{1t} - \bar{x}_1)^2$, which on division by σ^2 has the chi-square distribution with $n_1 - 1$ degrees of freedom. The first sum on the right of (8) is simply the sum of all such chi-squares for the various values of z. The second sum of squares on the right of (8) has the chi-square distribution (with $k - 2$ degrees of freedom) only if the regression function is in fact of the form $\alpha + \beta z$. Thus

$$F = \frac{\Sigma n_s(\bar{x}_s - \hat{\alpha} - \hat{\beta}z_s)^2/(k - 2)}{\sum_{st} (x_{st} - \bar{x}_s)^2/(n - k)} \tag{9}$$

provides a test for the hypothesis that the regression function is of the form $\alpha + \beta z$, and the critical region is the right-hand tail of the F distribution since a regression function different from $\alpha + \beta z$ would tend to increase the deviations of $\bar{x}_s$ from $\hat{\alpha} + \hat{\beta}z_s$.

Though this device is called a test for linearity, the same technique could obviously be used to test the validity of any specified regression function provided the function was linear in the unknown coefficients and there were fewer coefficients than distinct values of z.

14.3. One-factor Experiments. As an illustration, let us suppose that a factory manager wishes to buy machines to perform a certain operation in a production process. There are four companies which make such machines, and he obtains one on trial from each company with a view to determining which of the four is best suited to his purposes. Suppose also that a machine is operated by one man. The manager intends to have several of his men operate the machines for a few days in order to discover which of the four produces the most items per day. In this simple experiment the subject is the number of items produced, and the single factor is type of machine.

Let us suppose that twenty men are to be used in the experiment, five being assigned at random to each machine, and that each man will work one day on the particular machine he was assigned to. There will then be five observations for each of the four machines, each observation being the amount produced by the machine in one day. The data might be such as appear in the accompanying table. The question of interest is whether or not the machines are different with respect to number of items produced; i.e., is the subject of the experiment affected by the factor being investigated?

Machine number			
1	2	3	4
64	41	65	45
39	48	57	51
65	41	56	55
46	49	72	48
63	57	64	47

In order to analyze these data, the following assumptions will be made: the five observations for machine 1 constitute a random sample from a normal population with mean ξ_1 and variance σ^2; the observations for the second machine are an independent random sample from a normal population with mean ξ_2 and the same variance σ^2; and similarly for the other two machines. The assumptions are thus:

1. The samples are random.
2. The samples are independent.

3. The populations are normal.

4. The populations all have the same variance (often called the assumption of homoscedasticity).

In the general one-factor experiment, the factor will appear at k levels (instead of four); the observations will be denoted by x_{ij}, with $i = 1, 2, \cdots, k$, and $j = 1, 2, \cdots, n_i$, allowing for the possibility that there may be different numbers of observations at each level. The joint density of the x_{ij} is the product of the individual densities:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-(1/2\sigma^2)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\xi_i)^2} \tag{1}$$

where n represents Σn_i. The null hypothesis to be tested is that $\xi_1 = \xi_2 = \xi_3 = \cdots = \xi_k$. One could obtain a test by the likelihood-ratio method, i.e., by maximizing (1) with respect to all the parameters, then with all ξ_i made equal, and using the ratio as a test criterion. We shall, however, proceed differently.

The average of all the population means will be denoted by ξ,

$$\xi = \frac{1}{n}\sum_{ij}\xi_i = \frac{1}{n}\sum_i n_i\xi_i \tag{2}$$

and the deviations of the ξ_i from ξ will be denoted by

$$\alpha_i = \xi_i - \xi \qquad \Sigma n_i\alpha_i = 0 \tag{3}$$

The α_i are called the *effects* of the factor; the effects are zero under the null hypothesis. Also we shall denote the cell means by

$$\bar{x}_i = \frac{1}{n_i}\sum_j x_{ij} \tag{4}$$

and the mean of the whole set of observations by

$$\bar{x} = \frac{1}{n}\sum_{ij} x_{ij} = \frac{1}{n}\sum_i n_i\bar{x}_i \tag{5}$$

The sum of squares of deviations from the population mean for the observations in any one cell may be partitioned as follows:

$$\begin{aligned}\sum_j (x_{ij} - \xi_i)^2 &= \sum_j (x_{ij} - \bar{x}_i + \bar{x}_i - \xi_i)^2 \\ &= \sum_j (x_{ij} - \bar{x}_i)^2 + n_i(\bar{x}_i - \xi_i)^2\end{aligned} \tag{6}$$

and the two terms on the right of (6) (on division by σ^2) have inde-

pendent chi-square distributions with $k - 1$ and 1 degrees of freedom, as follows from Sec. 10.4. On summing (6) over i, the total sum of squares is partitioned into two parts:

$$\sum_{ij} (x_{ij} - \xi_i)^2 = \sum_{ij} (x_{ij} - \bar{x}_i)^2 + \sum_i n_i(\bar{x}_i - \xi_i)^2 \tag{7}$$

independently distributed by chi-square laws with $n - k$ and k degrees of freedom. The second term on the right of (7) may be further partitioned:

$$\begin{aligned}\sum_i n_i(\bar{x}_i - \xi_i)^2 &= \sum_i n_i(\bar{x}_i - \bar{x} - \alpha_i + \bar{x} - \xi)^2 \\ &= \sum_i n_i(\bar{x}_i - \bar{x} - \alpha_i)^2 + n(\bar{x} - \xi)^2 \end{aligned} \tag{8}$$

The two terms on the right of (8) are independently distributed by chi-square laws with $k - 1$ and 1 degrees of freedom, as may be shown by an argument entirely analogous to that employed in Sec. 10.4. We have then

$$\sum_{ij} (x_{ij} - \xi_i)^2 = \sum_{ij} (x_{ij} - \bar{x}_i)^2 + \sum_i n_i(\bar{x}_i - \bar{x} - \alpha_i)^2 + n(\bar{x} - \xi)^2 \tag{9}$$

and this partition is usually exhibited in an analysis-of-variance table such as the accompanying one with the parameters put equal to zero.

ANALYSIS OF VARIANCE FOR ONE-FACTOR EXPERIMENTS

Source	Sum of squares	Degrees of freedom	Mean square	F ratio
Mean	$n\bar{x}^2$	1		
Effects	$\Sigma\, n_i(\bar{x}_i - \bar{x})^2$	$k - 1$	$\dfrac{\sum_i n_i(\bar{x}_i - \bar{x})^2}{k - 1}$	$\dfrac{\sum n_i(\bar{x}_i - \bar{x})^2/(k - 1)}{\sum_{ij} (x_{ij} - \bar{x}_i)^2/(n - k)}$
Deviations	$\sum_{ij} (x_{ij} - \bar{x}_i)^2$	$n - k$	$\dfrac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{n - k}$	
Total	$\sum_{ij} x_{ij}^2$	n		

The ratio in the right-hand column obviously has the F distribution with $k - 1$ and $n - k$ degrees of freedom under the null hypothesis, $\alpha_i = 0$, and thus provides a test criterion for that hypothesis. Ordinarily there is no interest in testing $\xi = 0$, but if there is, the quantity $n\bar{x}^2$ divided by the mean-square deviations will have the F distribution with 1 and $n - k$ degrees of freedom under that null hypothesis. This latter test, incidentally, is orthogonal to the test on α_i, for the term on the right of (9) does not involve the α_i.

14.4. An Application of Normal Regression Theory. The foregoing analysis of the one-factor experiment is somewhat artificial in that the partition of the sum of squares seems to have no particular motivation. How would one know to embark on such an analysis in the first place? Having developed a logical theory of testing linear hypotheses in the preceding chapter, why not apply it here? The answer is that the foregoing analysis is relatively simple, whereas the application of the general theory involves some troublesome algebraic manipulation. As experiments become more complicated, the algebra of the general method becomes quite complex, involving, as it does, the inversion of large matrices. With experience, one can develop a facility for partitioning the sum of squares appropriately and thus save himself a great deal of mathematical analysis.

The simple partitioning of the sum of squares happens to give the correct tests when tests are orthogonal, but it does not prove, without advanced mathematical arguments unavailable to us here, that the tests are correct. A rigorous derivation of the tests does require application of the general theory, and we shall illustrate such an application for the one-factor experiment.

The k normal populations of Sec. 3 may be combined into a normal regression system with mean

$$\mu = \sum_i \delta_i \xi_i \tag{1}$$

where δ_i is an observable parameter defined to be one when an observation is drawn from the ith population and zero otherwise. The means ξ_i thus become coefficients of a linear regression function. It is simpler, however, to set up the regression function in terms of the α's so that the null hypothesis is in the form $\alpha_i = 0$ rather than $\xi_1 = \xi_2 = \cdots = \xi_k$, i.e., in the form of (13.6.4) rather than (13.6.8). To this end, we write (1) as

$$\mu = \xi + \Sigma \delta_i \alpha_i \tag{2}$$

but now we have one too many parameters because the α_i are connected by $\Sigma n_i\alpha_i = 0$. We shall eliminate α_k from (2) by the substitution

$$\alpha_k = -\frac{1}{n_k}\sum_1^{k-1} n_i\alpha_i \tag{3}$$

and get

$$\begin{aligned}\mu &= \xi + \sum_1^{k-1} \delta_i\alpha_i - \frac{1}{n_k}\delta_k \sum_1^{k-1} n_i\alpha_i \\ &= \sum_1^{k-1} \alpha_i\left(\delta_i - \frac{n_i}{n_k}\delta_k\right) + \xi \end{aligned} \tag{4}$$

Now we define new observable parameters z_p by

$$z_p = \delta_p - \frac{n_p}{n_k}\delta_k \qquad p = 1, 2, \cdots, k-1 \tag{5}$$

$$= 1 \qquad p = k \tag{6}$$

or, for $p = 1, 2, \cdots, k-1$,

$$\begin{aligned} z_p &= 1 && \text{if } x_{ij} \text{ has } i = p \\ &= -\frac{n_p}{n_k} && \text{if } x_{ij} \text{ has } i = k \\ &= 0 && \text{otherwise} \end{aligned} \tag{7}$$

The regression function is now

$$\mu = \sum_1^{k-1} \alpha_p z_p + \xi z_k \tag{8}$$

and is of the form discussed in Sec. 13.5, where ξ is to be identified with the α_k of that section.

Since, obviously,

$$\hat{\xi}_i = \bar{x}_i \tag{9}$$

we have at once the estimators

$$\hat{\alpha}_i = \bar{x}_i - \bar{x} \qquad i = 1, 2, \cdots, k-1 \tag{10}$$

$$\hat{\xi} = \bar{x} \tag{11}$$

$$\hat{\sigma}^2 = {}_n\sum (x_{ij} - \bar{x}_i)^2 \tag{12}$$

The test of the null hypothesis, $\alpha_i = 0$, is given by (13.6.6) so that we

must only evaluate Q, which is defined by (13.6.5). To this end, we must examine the matrix $\|a_{pq}\|$ defined by

$$a_{pq} = \sum_{ij} z_{pij} z_{qij} \tag{13}$$

since the sum on i in Sec. 13.5 refers to the sum over all sample observations and becomes the total sum over i and j in the present example. z_{pij} is, of course, the value of z_p for the observation x_{ij}. It follows readily from equation (7) that the matrix is

$$\|a_{pq}\| = \begin{Vmatrix} n_1 + \frac{n_1^2}{n_k} & \frac{n_1 n_2}{n_k} & \frac{n_1 n_3}{n_k} & \cdots & \frac{n_1 n_{k-1}}{n_k} & 0 \\ \frac{n_1 n_2}{n_k} & n_2 + \frac{n_2^2}{n_k} & \frac{n_2 n_3}{n_k} & \cdots & \frac{n_2 n_{k-1}}{n_k} & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \frac{n_1 n_{k-1}}{n_k} & \frac{n_2 n_{k-1}}{n_k} & \frac{n_3 n_{k-1}}{n_k} & \cdots & n_{k-1} + \frac{n_{k-1}^2}{n_k} & 0 \\ 0 & 0 & 0 & \cdots & 0 & n \end{Vmatrix} \tag{14}$$

To obtain the coefficients b_{uv} which appear in Q, one would ordinarily invert (14), then strike out the last row and column (m being $k - 1$ in the present example), then invert the result. This work is not necessary in the instance at hand, for Q is the quadratic form of the marginal distribution of the $\hat{\alpha}_u$ ($u = 1, 2, \cdots, k - 1$), and it is apparent from the form of (14) that the $\hat{\alpha}_u$ and $\hat{\xi}$ are independently distributed. That is, because of the zeros in $\|a_{pq}\|$, the joint distribution of the $\hat{\alpha}_u$ and $\hat{\xi}$ may be written as the product of a function of the $\hat{\alpha}_u$ alone and a function of $\hat{\xi}$ alone. It is evident then that

$$b_{uv} = a_{uv} \qquad u, v = 1, 2, \cdots, k - 1 \tag{15}$$

hence that

$$\sigma^2 Q = \sum_{u,v=1}^{k-1} \left(n_u \delta_{uv} + \frac{n_u n_v}{n_k} \right) (\hat{\alpha}_u - \alpha_u)(\hat{\alpha}_v - \alpha_v) \tag{16}$$

In this expression we put the α's equal to zero, and we may substitute from (10) for the $\hat{\alpha}$'s to obtain

$$\sigma^2 Q = \sum n_u (\bar{x}_u - \bar{x})^2 + \frac{1}{n_k} \sum_{u,v} n_u n_v (\bar{x}_u - \bar{x})(\bar{x}_v - \bar{x}) \tag{17}$$

The second term is simply

$$\frac{1}{n_k}\left[\sum_u n_u(\bar{x}_u - \bar{x})\right]^2 = \frac{1}{n_k}\left[\sum_1^{k-1}(n_u\bar{x}_u - (n - n_k)\bar{x}\right]^2$$

$$= \frac{1}{n_k}[n\bar{x} - n_k\bar{x}_k - (n - n_k)\bar{x}]^2$$

$$= n_k(\bar{x} - \bar{x}_k)^2$$

Thus Q becomes

$$Q = \frac{1}{\sigma^2}\sum_1^k n_i(\bar{x}_i - \bar{x})^2 \tag{18}$$

and the F ratio (13.6.6) is

$$F = \frac{\Sigma n_i(\bar{x}_i - \bar{x})^2/(k-1)}{\Sigma(x_{ij} - \bar{x}_i)^2/(n-k)} \tag{19}$$

the same as appears in the analysis-of-variance table of the preceding section.

We have shown, incidentally, in this section that the two terms of equation (3.8) are independently distributed by chi-square laws.

14.5. Two-factor Experiments with One Observation per Cell. It may have been noticed that the experiment described in Sec. 3 was very poorly designed. The trouble is that there is an extraneous factor, ability of the various workmen, which must necessarily enter into the experiment. If, in the experiment of Sec. 3, the production from one machine turned out to be relatively large, was it due to the machine, or to the excellence of the particular group of workmen assigned to it? There is no way to tell from that experiment. In the language of experimental design, the effects due to machines and the effects due to groups of workmen are completely *confounded;* there is no way to differentiate the two factors.

The difficulty is removed by redesigning the experiment as a two-factor experiment. Let, for example, only five men be involved in the experiment and let each of the five men work one day on each of the four machines. The order in which a given man works on the four machines would be assigned by a random process. The data are now classified in a two-way table in accordance with the two factors and might appear as in the table on page 330. When a two-factor experiment is used to control an extraneous factor as in the case here, the design is referred to as a *randomized block design.* The factor of interest is compared in blocks (men, in the present instance) so that conditions of the comparison are homogeneous within each block though they differ from block to block.

		Machine			
		1	2	3	4
Man	1	53	47	57	45
	2	56	50	63	52
	3	45	47	54	42
	4	52	47	57	41
	5	49	53	58	48

In general there will be, say, r rows and c columns for a two-factor experiment, one factor being examined at r levels, $A_1, A_2, \cdots, A_r$ and the other factor B at c levels, $B_1, B_2, \cdots, B_c$. The observations may be denoted by x_{ij}, $i = 1, 2, \cdots, r$, and $j = 1, 2, \cdots, c$. It is assumed that the x_{ij} are random independent observations from normal populations with the same variances. It is further assumed that the effects of the two factors are additive. This last assumption will be discussed further in Secs. 6 and 9. Analytically it states that the means of the normal population associated with the individual cells are assumed to be of the form

$$\xi_{ij} = \xi + \alpha_i + \beta_j \tag{1}$$

with

$$\Sigma\alpha_i = 0 \qquad \Sigma\beta_j = 0 \tag{2}$$

The parameter ξ is the average of all the population means. In terms of the illustrative example, the most skilled workman will have a positive α associated with him, and the assumption (1) states that whatever machine he works on his production will be exactly α (in the population mean) larger than the mean production of all workers on that machine. Or in other terms, if one workman is 10 units better than another on one machine, he will be ten units better than the other on all machines. Similarly if one machine is 10 units better than another, that margin is assumed to be the same (in the population means) regardless of whether a workman is good or bad.

In the general two-factor experiment, the two null hypotheses of interest are $\alpha_i = 0$ and $\beta_j = 0$. (In the illustration we are using, there is, of course, little interest in the α's.) We shall therefore try to partition the total sum of squares into parts, one of which involves the

α_i, another the β_j, and another ξ. The proper procedure is suggested by the estimators of these parameters, which are readily found to be

$$\hat{\xi} = \bar{x} = \frac{1}{rc} \sum_{ij} x_{ij} \tag{3}$$

$$\hat{\alpha}_i = \bar{x}_{i.} - \bar{x} = \frac{1}{c} \sum_{j} x_{ij} - \bar{x} \tag{4}$$

$$\hat{\beta}_j = \bar{x}_{.j} - \bar{x} = \frac{1}{r} \sum_{i} x_{ij} - \bar{x} \tag{5}$$

$\bar{x}_{i.}$ being the mean of the observations in the ith row and $\bar{x}_{.j}$ the mean of those in the jth column. The total sum of squares may be partitioned as follows:

$$\sum_{ij} (x_{ij} - \xi - \alpha_i - \beta_j)^2 = \sum_{ij} [(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}) + (\bar{x}_{i.} - \bar{x} - \alpha_i) + (\bar{x}_{.j} - \bar{x} - \beta_j) + (\bar{x} - \xi)]^2 \tag{6}$$

$$= \sum_{ij} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 + c \sum_{i} (x_{i.} - \bar{x} - \alpha_i)^2 + r \sum_{j} (x_{.j} - \bar{x} - \beta_j)^2 + rc(\bar{x} - \xi)^2 \tag{7}$$

Equation (7) is obtained by squaring the expression in (6), using the grouping indicated by the parentheses; then it is easily seen that the cross-product terms sum to zero.

ANALYSIS OF VARIANCE FOR TWO-FACTOR EXPERIMENTS WITH ONE OBSERVATION PER CELL

Source	Sum of squares	Degrees of freedom	Mean square	F ratio
Mean	$rc\bar{x}^2 = S_1$	1	$S_1 = s_1$	$\frac{s_1}{s_4}$
A effect	$c \sum_i (\bar{x}_{i.} - \bar{x})^2 = S_2$	$r - 1$	$\frac{S_2}{r-1} = s_2$	$\frac{s_2}{s_4}$
B effect	$r \sum_j (\bar{x}_{.j} - \bar{x})^2 = S_3$	$c - 1$	$\frac{S_3}{c-1} = s_3$	$\frac{s_3}{s_4}$
Deviations	$\sum_{ij} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_4$	$(r-1)(c-1)$	$\frac{S_4}{(r-1)(c-1)} = s_4$	
Total	$\sum_{ij} x_{ij}^2$	rc		

If we had some relatively advanced techniques at our disposal, the analysis would now be virtually complete, for then it would be possible to argue that the four terms on the right of (7) are each independently distributed by chi-square laws (on division by σ^2)—the first with $(r-1)(c-1)$ degrees of freedom, the second with $r-1$, the third with $c-1$, and the fourth with one degree of freedom. Assuming the truth of this statement for the moment, we may construct the table shown on page 331. The F ratios in the final column give orthogonal tests of three null hypotheses: $\xi = 0$, $\alpha_i = 0$, $\beta_j = 0$.

To demonstrate the validity of the above analysis, we must investigate the tests more formally. Again the general theory of testing linear hypotheses will be employed. Equation (1) may be put in the form of a linear regression function by defining observable parameters δ_i and ϵ_j so that

$$\begin{aligned} \delta_i &= 1 \qquad \text{if } x_{ij} \text{ has } i = i' \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{8}$$

$$\begin{aligned} \epsilon_j &= 1 \qquad \text{if } x_{ij} \text{ has } j = j' \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{9}$$

Then (1) becomes

$$\xi_{ij} = \xi + \sum_i \delta_i\alpha_i + \sum_j \epsilon_j\beta_j \tag{10}$$

This relation involves only $r + c - 1$ parameters in view of conditions (2), so we shall eliminate α_r and β_c from (10) to get

$$\xi_{ij} = \xi + \Sigma(\delta_i - \delta_r)\alpha_i + \Sigma(\epsilon_j - \epsilon_c)\beta_j \tag{11}$$

and as in Sec. 4, new observable parameters are defined by

$$\left.\begin{aligned} z_p &= 1 \text{ if } x_{ij} \text{ has } i = p \\ &= -1 \text{ if } x_{ij} \text{ has } i = r \\ &= 0 \text{ otherwise} \end{aligned}\right\} \qquad p = 1, 2, \cdots, r-1 \tag{12}$$

$$\left.\begin{aligned} z_p &= 1 \text{ if } x_{ij} \text{ has } j = p - r + 1 \\ &= -1 \text{ if } x_{ij} \text{ has } j = c \\ &= 0 \text{ otherwise} \end{aligned}\right\} p = r, r+1, \cdots, r + c - 2 \tag{13}$$

$$z_{r+c-1} = 1 \tag{14}$$

There are thus $r + c - 1$ observable parameters; the first $r - 1$ are associated with the α's, the next $c - 1$ with the β's, and the last one

with ξ. The population mean is now of the form

$$\sum_{p=1}^{r+c-1} \alpha_p z_p \tag{15}$$

if we redefine

$$\beta_j = \alpha_{r-1+j} \qquad j = 1, 2, \cdots, c-1 \tag{16}$$

$$\xi = \alpha_{r+c-1} \tag{17}$$

The $\hat{\alpha}_p$ are given by equations (3), (4), and (5), and $\hat{\sigma}^2$ is readily seen to be

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{rc}\sum_{ij}\left(x_{ij} - \sum \hat{\alpha}_p z_p\right)^2 \\ &= \frac{1}{rc}\sum\left(x_{ij} - \sum \hat{\alpha}_{i.}\delta_i - \sum \hat{\beta}_{j.}e_j - \hat{\xi}\right)^2 \\ &= \frac{1}{rc}\sum (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2\end{aligned} \tag{18}$$

The joint distribution of the $\hat{\alpha}_p$ is normal, with the matrix of the quadratic form defined by

$$a_{pq} = \sum_{ij} z_{pij} z_{qij} \tag{19}$$

and on evaluating these sums using (12), (13), and (14), it is easily found that

$$\|a_{pq}\| = \left\|\begin{array}{cccccc|cccccc|c}
2c & c & c & \cdots & & c & 0 & 0 & 0 & \cdots & & 0 & 0 \\
c & 2c & c & \cdots & & c & 0 & 0 & 0 & \cdots & & 0 & 0 \\
c & c & 2c & \cdots & & c & 0 & 0 & 0 & \cdots & & 0 & 0 \\
\vdots & \vdots & \vdots & & & \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\
c & c & c & \cdots & & 2c & 0 & 0 & 0 & \cdots & & 0 & 0 \\
\hline
0 & 0 & 0 & \cdots & & 0 & 2r & r & r & \cdots & & r & 0 \\
0 & 0 & 0 & \cdots & & 0 & r & 2r & r & \cdots & & r & 0 \\
0 & 0 & 0 & \cdots & & 0 & r & r & 2r & \cdots & & r & 0 \\
\vdots & \vdots & \vdots & & & \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & & 0 & r & r & r & \cdots & & 2r & 0 \\
\hline
0 & 0 & 0 & & & 0 & 0 & 0 & 0 & \cdots & & 0 & rc
\end{array}\right\| \tag{20}$$

There are $r - 1$ rows and columns in the upper left-hand block, and $c - 1$ rows and columns in the block completely enclosed by dashed lines. The form of $\|a_{pq}\|$ shows at once that the three sets of parameters $(\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_{r-1})$, $(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_{c-1})$, and $(\hat{\xi})$ are independently distributed; hence their quadratic forms

$$\sum_{i,i'=1}^{r-1} a_{ii'}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_{i'} - \alpha_{i'}) \tag{21}$$

$$\sum_{j,j'=1}^{c-1} a_{r-1+j,\, r-1+j'}(\hat{\beta}_j - \beta_j)(\hat{\beta}_{j'} - \beta_{j'}) \tag{22}$$

$$rc(\hat{\xi} - \xi)^2 \tag{23}$$

are independently distributed by chi-square laws with $r - 1$, $c - 1$, and one degrees of freedom, respectively. All three of them are distributed independently of

$$rc\hat{\sigma}^2 \text{ [which has } rc - (r - 1) - (c - 1) - 1 = (r - 1)(c - 1) \text{ degrees of freedom]}$$

in view of the results of Sec. 13.5. These three forms reduce directly to the last three terms of (7); hence the F ratios of the analysis-of-variance table are all of the form (13.6.6).

14.6. Two-factor Experiments with Several Observations per Cell. To continue the illustration that has already been used, suppose again that there are four kinds of machines to be tested with five men and also that instead of one each of the machines, there are three each. Every man works one day with all twelve machines, and the data are classified again in a 4×5 array, but now there are three observations in each cell corresponding to the three machines of each type.

In general, we shall suppose that there are r rows and c columns and that there are m observations in each cell. There will then be rcm observations altogether which will be denoted by x_{ijk} $(i = 1, 2, \cdots, r;$ $j = 1, 2, \cdots, c;$ $k = 1, 2, \cdots, m)$. The observations in the (i, j) cell are assumed to be a random sample from a normal population with mean ξ_{ij} and variance σ^2, the same for all cells; the cell populations differ only in their means. The numbers ξ_{ij} may be put in the form

$$\xi_{ij} = \xi + \alpha_i + \beta_j + \gamma_{ij} \tag{1}$$

with

$$\sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0 \qquad \sum_i \gamma_{ij} = 0 \qquad \sum_j \gamma_{ij} = 0 \tag{2}$$

To do this, one first computes

$$\xi = \frac{1}{rc}\sum_{ij} \xi_{ij}$$

then

$$\alpha_i = \frac{1}{c}\sum_j \xi_{ij} - \xi \qquad \beta_j = \frac{1}{r}\sum_i \xi_{ij} - \xi$$

and finally, the γ_{ij}, using (1). ξ is called the *mean effect;* the α_i are called the *main effects due to rows*, or briefly the *row effects;* the β_j are called *column effects;* and the γ_{ij} are called the *row-column interaction effects*, or simply the *interactions*. When the interactions are all zero, the means ξ_{ij} are said to be additive (see preceding section).

We shall now partition the sum of squares into parts suitable for constructing tests on the main effect, row effects, column effects, and interactions. Considering first the observations in a single cell, the sum of squares may be divided into two parts just as was done in equation (3.6):

$$\sum_k (x_{ijk} - \xi_{ij})^2 = \sum_k (x_{ijk} - \bar{x}_{ij.})^2 + k(\bar{x}_{ij.} - \xi_{ij})^2 \tag{3}$$

where $\bar{x}_{ij.}$ is the cell mean and is the estimator of ξ_{ij}. The sum on the right of (3) has $k - 1$ degrees of freedom, and the other term on the right is independently distributed of the sum with one degree of freedom. Summing (3) over all cells

$$\sum_{ijk} (x_{ijk} - \xi_{ij})^2 = \sum_{ijk} (x_{ijk} - \bar{x}_{ij.})^2 + m\sum_{ij} (\bar{x}_{ij.} - \xi_{ij})^2 \tag{4}$$

the total sum of squares is divided into two parts independently distributed by chi-square laws (on division by σ^2), the first with $rc(m - 1)$ degrees of freedom and the second with rc degrees of freedom. The second sum of squares for the cell means may be partitioned into four parts just as was done in equations (5.6) and (5.7) for the case of a single observation in a two-way table. The result is

$$m\sum_{ij} (\bar{x}_{ij.} - \xi - \alpha_i - \beta_j - \gamma_{ij})^2 = m\sum_{ij} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x} - \gamma_{ij})^2$$
$$+ mc\sum_i (\bar{x}_{i..} - \bar{x} - \alpha_i)^2 + mr\sum_j (\bar{x}_{.j.} - \bar{x} - \beta_j)^2 + mrc(\bar{x} - \xi)^2 \tag{5}$$

which differs from (5.7) only in the appearance of m and γ_{ij}. The

symbols $\bar{x}_{i..}$ and $\bar{x}_{.j.}$ are the row and column means

$$\bar{x}_{i..} = \frac{1}{mc}\sum_{jk} x_{ijk} = \frac{1}{c}\sum_{j} \bar{x}_{ij.}$$

$$\bar{x}_{.j.} = \frac{1}{mr}\sum_{ik} x_{ijk} = \frac{1}{r}\sum_{i} \bar{x}_{ij.}$$

while $\bar{x}$ represents the mean of all the observations.

It is now apparent why the population means were assumed to be additive in Sec. 5; the first term on the right of (5) corresponds to the deviation sum of squares in (5.7), and if the γ_{ij} were not zero, it would be impossible to carry out the tests described there, because the γ_{ij} are usually unknown parameters. However an alternative model to be described in Sec. 9 allows the tests of Sec. 5 to be made in any case.

Returning to the present problem, the total sum of squares has been partitioned into parts which may be exhibited as in the accompanying table. The degree of freedom corresponding to ξ has been omitted, as it often is in such tables, because there is practically never any interest in testing the null hypothesis that $\xi = 0$. The three F ratios in the final column of the table may be used to test the three null hypotheses: $\alpha_i = 0$, $\beta_j = 0$, $\gamma_{ij} = 0$. These are the appropriate tests

Source	Sum of squares	Degrees of freedom	Mean square	F
Row	$mc\sum_i (\bar{x}_{i..} - \bar{x})^2 = S_1$	$r - 1$	$\frac{S_1}{r-1} = s_1$	$\frac{s_1}{s_4}$
Column	$mr\sum_j (\bar{x}_{.j.} - \bar{x})^2 = S_2$	$c - 1$	$\frac{S_2}{c-1} = s_2$	$\frac{s_2}{s_4}$
Interaction	$m\sum_{ij} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2 = S_3$	$(r-1)(c-1)$	$\frac{S_3}{(r-1)(c-1)} = s_3$	$\frac{s_3}{s_4}$
Deviations	$\sum_{ijk} (x_{ijk} - \bar{x}_{ij.})^2 = S_4$	$rc(m-1)$	$\frac{S_4}{rc(m-1)} = s_4$	
Total	$\sum_{ijk} (x_{ijk} - \bar{x})^2$	$rcm - 1$		

for these three hypotheses under the theoretical model used here. Actually in practice the row effects and column effects are rarely tested in this manner. Ordinarily the two sets of main effects are tested by the ratios s_1/s_3 and s_2/s_3. These tests do not make sense in

theory with the present model if the γ_{ij} are not zero, for then it is

$$m \sum_{ij} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x} - \gamma_{ij})^2$$

which has the chi-square distribution, not the quantity S_3 in which the γ_{ij} have been put equal to zero.

The rationale for comparing main effects with interaction rather than deviations in an F test may be indicated as follows from a purely practical standpoint: Using the men and machines illustration, suppose the null hypothesis, $\gamma_{ij} = 0$, is rejected. The implication is that while one man does better on one machine than another man, he may not do as much better than the other on a second machine or he may even do worse. Suppose these interactions between men and machines are of the order of 3 or 4 units produced per day. It would be quite surprising, in view of such interactions, if the main effects were not at least of this order (3 or 4 units per day). In fact, the vanishing of the α_i or β_j in the face of nonvanishing γ_{ij} would rightly be regarded as a pathological case. Suppose the β_j (the main effects due to machines) are, in truth, of the same order of magnitude as the interactions. Then certainly the differences between machines are of no practical consequence, for one might purchase what appears to be the best machine only to have it operated by a man who does not happen to work so well with that machine, and better production might have resulted had another machine been purchased. Obviously machine differences are important only if they are large relative to the men-machines interactions.

Arguing very crudely now, the sum of squares S_2 in the table is a measure of the "variance" of the β_j since

$$\hat{\beta}_j = \bar{x}_{.j.} - \bar{x}$$

and S_3 is a measure of the "variance" of the γ_{ij} since

$$\hat{\gamma}_{ij} = \bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}$$

The ratio s_2/s_3 measures the relative sizes of these "variances," and if the ratio is large (relative to unity), the machine differences are important in relation to the interactions. These rough considerations will be made more precise in Sec. 9.

14.7. Three-factor Experiments. To augment our illustrative example, the products of the machines in question may be made in several different sizes, and for purposes of the experiment three sizes may have been selected for inclusion. There would then be three

Source	Sum of squares	Degrees of freedom
A effect	$mr_2r_3\sum_h(\bar{x}_{h\ldots}-\bar{x})^2$	r_1-1
B effect	$mr_1r_3\sum_i(\bar{x}_{.i.}-\bar{x})^2$	r_2-1
C effect	$mr_1r_2\sum_j(\bar{x}_{..j}-\bar{x})^2$	r_3-1
AB interaction	$mr_3\sum_{hi}(\bar{x}_{hi.}-\bar{x}_{h\ldots}-\bar{x}_{.i.}+\bar{x})^2$	$(r_1-1)(r_2-1)$
AC interaction	$mr_2\sum_{hj}(\bar{x}_{h.j}-\bar{x}_{h\ldots}-\bar{x}_{..j}+\bar{x})^2$	$(r_1-1)(r_3-1)$
BC interaction	$mr_1\sum_{ij}(\bar{x}_{.ij}-\bar{x}_{.i.}-\bar{x}_{..j}+\bar{x})^2$	$(r_2-1)(r_3-1)$
ABC interaction	$m\sum_{hij}(\bar{x}_{hij}-\bar{x}_{hi.}-\bar{x}_{h.j}-\bar{x}_{.ij}+\bar{x}_{h..}+\bar{x}_{.i.}+\bar{x}_{..j}-\bar{x})^2$	$(r_1-1)(r_2-1)(r_3-1)$
Deviations	$\sum_{hijk}(x_{hijk}-\bar{x}_{hij.})^2$	$r_1r_2r_3(m-1)$
Total	$\sum_{hijk}(x_{hijk}-\bar{x})^2$	$r_1r_2r_3m-1$

factors: machines at four levels, men at five levels, sizes of product at three levels. The observations might then be arranged in a three-dimensional table with $4 \times 5 \times 3 = 60$ cells, and if there were three machines of each type, there would again be three observations per cell or 180 observations in all.

In general, let there be three factors A, B, and C with levels r_1, r_2, r_3, respectively, and let there be m observations per cell. The observations may be denoted by x_{hijk}, where $h = 1, 2, \cdots, r_1$; $i = 1, 2, \cdots, r_2$; $j = 1, 2, \cdots, r_3$; $k = 1, 2, \cdots, m$. The observations are assumed to come from normal populations with means ξ_{hij} and variances σ^2. The means may be written in the form

$$\xi_{hij} = \xi + \alpha_h + \beta_i + \gamma_j + \delta_{hi} + \epsilon_{hj} + \zeta_{ij} + \eta_{hij} \tag{1}$$

where any letter on the right sums to zero on any one of its indexes. The δ_{hi}, ϵ_{hj}, ζ_{ij} are called *two-factor interactions*, or *first-order interactions;* the η_{hij} are called *three-factor interactions*, or *second-order interactions*. The details of partitioning the sum of squares are so similar to those of the preceding section that we shall merely present the resulting analysis-of-variance table here. The mean squares are obtained by dividing the sums of squares by their corresponding degrees of freedom. The various null hypotheses ($\alpha_h = 0$, $\delta_{hi} = 0$, etc.) are tested by dividing the appropriate mean square by the deviation mean square and comparing the result with the critical F value. Here again, most of these tests would be pointless in many practical situations if some of the interactions were nonvanishing.

If there is only one observation per cell, there will be no deviation sum of squares, and it is necessary to use the three-factor-interaction sum of squares in its place. With the present model this substitution requires the assumption that the η_{hij} are zero.

14.8. Latin and Greco-Latin Squares. Latin and Greco-Latin squares are devices for reducing the scope of experiments which involve several factors and for performing experiments when it is impossible to obtain observations for all combinations of all levels of the factors. As an illustration of the latter case, we may alter the example already used. Suppose four kinds of machines (one of each kind) must be tested in one day and that a man must work at least 2 hours on a machine in order to get an adequate measure of his production on that machine. The 8-hour working day will be divided into four 2-hour periods, but now a third factor has entered the experiment because the time periods differ, at least to the extent that the workmen may be expected to be less efficient toward the end of the day due to fatigue.

We have then three factors: machine, men, and time periods. But it is impossible to obtain observations for all combinations of all levels since, for example, all men cannot work on the first machine during the first time period. The difficulty is met by setting up the experiment so that all factors appear at the same number of levels. Thus, since there are four machines and four time periods, we should use four men in the experiment.

The experiment is performed by arranging the levels of one factor in a Latin square which is simply a square array of letters such that every letter appears once and only once in every row and column.

A	B	C	D
C	D	A	B
B	C	D	A
D	A	B	C

We may identify the four letters with the four machines. The rows and columns are assigned to the other two factors. Thus if rows refer to men and columns to time periods, then the second man works on the first machine (A) during the third time period. Of course this design could be used for any three-factor experiment where all factors are at four levels each. Such an experiment would require 64 observations for all combinations, whereas with the Latin square it can be done with 16 observations; but of course this reduction in size of the experiment is at the expense of precision in the results.

In general, let us suppose that the three factors of a Latin square have r levels each and that the observations are $x_{ij(k)}$ where $i, j, k = 1, 2, \cdots, r$, and where i refers to rows, j to columns, and k to letters in the square. The (k) is enclosed in parentheses to indicate that it is not independent of i and j. The observations are assumed to come from normal populations with the same variance σ^2 and with means

$$\xi_{ij(k)} = \xi + \alpha_i + \beta_j + \gamma_k \tag{1}$$

in which $\Sigma\alpha_i = 0$, $\Sigma\beta_j = 0$, $\Sigma\gamma_k = 0$. All interactions are assumed to be zero in this model.

If we denote the row means by $\bar{x}_{i.}$, the column means by $\bar{x}_{.j}$, and the means of observations associated with the kth letter in the square (the kth level of the third factor) by $\bar{x}_{(k)}$, the sum of squares may easily

be partitioned as follows:

$$\sum_{ij} (x_{ij(k)} - \xi_{ij(k)})^2 = r \sum_i (\bar{x}_{i.} - \bar{x} - \alpha_i)^2 + r \sum_j (\bar{x}_{.j} - \bar{x} - \beta_j)^2$$
$$+ r \sum_k (\bar{x}_{(k)} - \bar{x} - \gamma_k)^2 + \sum_{ij} (x_{ij(k)} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_{(k)} + 2\bar{x})^2$$
$$+ r^2(\bar{x} - \xi)^2 \quad (2)$$

All these sums on the right are independently distributed by chi-square laws (on division by σ^2); the various sums have degrees of freedom indicated in the accompanying analysis-of-variance table. The degree of freedom for the mean has been omitted from the table.

Source	Sum of squares	Degrees of freedom
Rows	$r\Sigma(\bar{x}_{i.} - \bar{x})^2$	$r - 1$
Columns	$r\Sigma(\bar{x}_{.j} - \bar{x})^2$	$r - 1$
Letters	$r\Sigma(\bar{x}_{(k)} - \bar{x})^2$	$r - 1$
Deviations	$\Sigma(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_{(k)} + 2\bar{x})^2$	$(r - 1)(r - 2)$
Total	$\Sigma(x_{ij} - \bar{x})^2$	$r^2 - 1$

The three null hypotheses—$\alpha_i = 0$, $\beta_j = 0$, $\gamma_{(k)} = 0$—are tested by dividing the appropriate mean square by the deviation mean square and using the F distribution.

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
$B\gamma$	$A\delta$	$D\alpha$	$C\beta$
$C\delta$	$D\gamma$	$A\beta$	$B\alpha$
$D\beta$	$C\alpha$	$B\delta$	$A\gamma$

If the number of levels of the factors is a prime number or a power of a prime number, then it is possible to test more than three factors without increasing the number of observations. A Greco-Latin square is an arrangement of r Greek and r Latin letters in an $r \times r$ square so that each Greek and each Latin letter appears once and only once

in every row and column and such that every Greek letter appears once and only once with each Latin letter. With such an arrangement, four factors may be tested at r levels, each using only r^2 observations, while the complete experiment would require r^4 observations. The analysis-of-variance table would be similar to the one above for Latin squares; there would be an extra line for Greek letters having sum of squares $r\Sigma(\bar{x}_{(h)} - \bar{x})^2$ with $r - 1$ degrees of freedom, and the error sum of squares would become

$$\sum_{ij} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_{(k)} - \bar{x}_{(h)} + 3\bar{x})^2$$

with $(r - 1)(r - 3)$ degrees of freedom. The $\bar{x}_{(h)}$ represents the mean of those observations associated with the hth Greek letter.

More generally, when r is a prime or a power of a prime, it is possible to arrange $r - 1$ sets of r letters in an $r \times r$ square so that each letter of every set occurs once in every row and column and once with each letter of every other set. By means of such arrangements, many factors may be studied in one experiment with relatively few observations.

14.9. Components-of-variance Models. In this section we shall consider an alternative mathematical model for analyzing factorial experiments. To introduce the ideas, we shall consider a two-factor experiment with one observation per cell, the same situation as was discussed in Sec. 5. The observations are again denoted by x_{ij} with $i = 1, 2, \cdots, r_1$ and $j = 1, 2, \cdots, r_2$. In this model the row effects, the column effects, and the interaction effects are all assumed to be random variables. Specifically it is assumed that

$$x_{ij} = u_i + v_j + w_{ij} \tag{1}$$

where $u_1, u_2, \cdots, u_{r_1}$ is a random sample from a normal population with mean ξ_u; the v_j are an independent random sample from a normal population with mean ξ_v; and the w_{ij} are an independent random sample from a third normal population with mean ξ_w.

Altering the circumstances of the experiment in Sec. 5 slightly, let us suppose that there are a large number of manufacturers of the machines in question and that four particular makes were chosen at random. Also the five men chosen to participate in the experiment were chosen at random from some large group of men. It is assumed then that these five men have production abilities $u_1, u_2, \cdots, u_5$ which constitute five observations from a normal population. Similarly the four machines have productive capacities v_1, v_2, v_3, v_4 which

constitute an independently drawn sample from a second population. The variables w_{ij} may be looked upon as a sum of two variables, say $y_{ij} + z_{ij}$, with the y_{ij} interpreted as the interactions between men and machines and the z_{ij} consisting of miscellaneous minor effects which influence the final observations. These two variables y and z are assumed to be normal random variates, and their sum w will then be a normal random variate.

Referring back to equation (1), if we let

$$\xi = \xi_u + \xi_v + \xi_w, \qquad a_i = u_i - \xi_u, \qquad b_j = v_j - \xi_v, \qquad c_{ij} = w_{ij} - \xi_w$$

then the equation may be written in the form

$$x_{ij} = \xi + a_i + b_j + c_{ij} \tag{2}$$

where the three variates a, b, c now have zero means. We shall denote their variances by σ_a^2, σ_b^2, σ_c^2, respectively. Clearly the mean of any x_{ij} is ξ, and the variance of any x_{ij} is

$$\sigma_x^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 \tag{3}$$

since the three variates are assumed to be independent. It is to be observed that the x_{ij} themselves are not independent if they fall in the same row or column. Thus, for example,

$$E[(x_{11} - \xi)(x_{12} - \xi)] = E[(a_1 + b_1 + c_{11})(a_1 + b_2 + c_{12})] \tag{4}$$
$$= \sigma_a^2 \tag{5}$$

which arises from the a_1^2 term on the right of (4). Similarly the covariance between two observations in the same column is σ_b^2.

With the present model, the null hypothesis that the row effects are identical takes the form $\sigma_a^2 = 0$. This is to say that the a_i, which have mean zero, are actually identically zero; their distribution is concentrated at a point (zero), which is the only way σ_a^2 can be zero. Similarly the null hypothesis that the column effects are all the same takes the form $\sigma_b^2 = 0$.

To test these two hypotheses, the sum of squares is partitioned just as before:

$$\sum_{ij} (x_{ij} - \bar{x})^2 = \sum_{ij} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 + r_2 \sum_i (\bar{x}_{i.} - \bar{x})^2 + r_1 \sum_j (\bar{x}_{.j} - \bar{x})^2 \tag{6}$$

in which the degree of freedom for $\bar{x}$ has been omitted. If we substitute for the x's on the right in terms of a, b, and c, the result is

$$\sum_{ij} (x_{ij} - \bar{x})^2 = \sum_{ij} (c_{ij} - \bar{c}_{i.} - \bar{c}_{.j} + \bar{c})^2 + r_2 \sum_i (a_i + \bar{c}_{i.} - \bar{a} - \bar{c})^2 + r_1 \sum_j (b_j + \bar{c}_{.j} - \bar{b} - \bar{c})^2 \quad (7)$$

where the $\bar{c}$'s are defined in the same way as the $\bar{x}$'s and $\bar{a} = \Sigma a_i/r_2$, $\bar{b} = \Sigma b_j/c$. It is easily shown that the three sums on the right are independently distributed by chi-square laws by using the results of Sec. 5. The results of the latter part of that section may be applied to the c_{ij} since these variables are independently normally distributed. (Since the c_{ij} all have zero means, the means α_i, β_j, ξ of Sec. 5 are all replaced by zero.) It follows from Sec. 13.5 and equations (5.20), (5.21), (5.22) that $\Sigma(c_{ij} - \bar{c}_{i.} - \bar{c}_{.j} + \bar{c})^2$ is distributed independently of the deviations $\bar{c}_{i.} - \bar{c}$ and $\bar{c}_{.j} - \bar{c}$; further, the set of deviations $c_{i.} - \bar{c}$ is distributed independently of the set $\bar{c}_{.j} - \bar{c}$, as follows from (5.20). Also the sum in question when divided by σ_c^2 has the chi-square distribution with $(r_1 - 1)(r_2 - 1)$ degrees of freedom.

Since, by assumption, the c's are independent of the a's and b's, it follows that the first sum on the right of (7) is distributed independently of the other two sums. These other two sums are also distributed independently, since the variables a_i and the variables $\bar{c}_{i.} - \bar{c}$ are independent of the b_j (by hypothesis) and the $\bar{c}_{.j} - \bar{c}$ (by equation 20 of Sec. 5). Furthermore, these two sums are distributed by chi-square laws. For considering the sum $\Sigma(a_i + \bar{c}_{i.} - \bar{a} - \bar{c})^2$, we may let

$$y_i = a_i + \bar{c}_{i.}$$

and we know that y is a normally distributed variate with mean zero and variance $\sigma_a^2 + (\sigma_c^2/r_2)$. Thus

$$\frac{\Sigma(y_i - \bar{y})^2}{\sigma_a^2 + (\sigma_c^2/r_2)}$$

has the chi-square distribution with $r_1 - 1$ degrees of freedom; hence it follows that the second sum on the right of (7), when divided by $(r_2\sigma_a^2 + \sigma_c^2)$, has the chi-square distribution with $r_1 - 1$ degrees of freedom. In the same vein, the third sum on the right of (7), when divided by $(r_1\sigma_b^2 + \sigma_c^2)$, has the chi-square distribution with $r_2 - 1$ degrees of freedom. All these results may be summarized in the accompanying table. The final column provides the divisors which make the corresponding sums of squares chi-square variates.

Source	Sum of squares	Degrees of freedom	Expected mean square
Rows	$r_2\Sigma(\bar{x}_{i.} - \bar{x})^2$	$r_1 - 1$	$\sigma_e^2 + r_2\sigma_a^2$
Columns	$r_1\Sigma(\bar{x}_{.j} - \bar{x})^2$	$r_2 - 1$	$\sigma_e^2 + r_1\sigma_b^2$
Deviations	$\Sigma(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$	$(r_1 - 1)(r_2 - 1)$	σ_e^2
Total	$\Sigma(x_{ij} - \bar{x})^2$	$r_1r_2 - 1$	

To test the null hypothesis $\sigma_a^2 = 0$, one again uses the ratio of the row mean square to the error mean square and compares that ratio with the critical value of the F distribution for $r_1 - 1$ and

$$(r_1 - 1)(r_2 - 1)$$

degrees of freedom. For under the null hypothesis these two sums of squares have the same divisor σ_e^2; hence that unknown parameter cancels out in the ratio of the two chi squares, and the ratio of the mean squares has the F distribution. The tests for the row and column effects thus take exactly the same form as those of Sec. 5, but here no assumption of additivity is required.

14.10. Components of Variance for Two-factor and Three-factor Experiments. For a two-factor experiment with m observations per cell, the observations are assumed to be of the form

$$x_{ijk} = \xi + a_i + b_j + c_{ij} + e_{ijk} \tag{1}$$

where the a's, b's, c's, and e's are normally distributed with zero means. The a's are associated with row effects, the b's with column effects, the c's with row-column interactions, and the e's with all other miscellaneous effects which influence the observations. The variances of these variates will be denoted by σ_a^2, σ_b^2, σ_{ab}^2, and σ_e^2; the σ_{ab}^2 is used in preference to σ_c^2 to indicate more clearly that it refers to the population of row-column interactions. We shall leave the details as an exercise, since they are very similar to those of Sec. 9, and merely present the results. The final column of the accompanying table shows at a glance the appropriate ratios of mean squares for testing the various null hypotheses: for $\sigma_{ab}^2 = 0$, one compares the interaction mean square with the deviation mean square (this is sometimes called the test of additivity); the main effects are tested against interaction (not against deviations as was the case in Sec. 6).

Source	Sum of squares	Degrees of freedom	Expected mean square
Rows	$mr_2\Sigma(\bar{x}_{i..} - \bar{x})^2$	$r_1 - 1$	$\sigma_e^2 + m\sigma_{ab}^2 + mr_2\sigma_a^2$
Columns	$mr_1\Sigma(\bar{x}_{.j.} - \bar{x})^2$	$r_2 - 1$	$\sigma_e^2 + m\sigma_{ab}^2 + mr_1\sigma_b^2$
Interactions	$m\Sigma(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$	$(r_1 - 1)(r_2 - 1)$	$\sigma_e^2 + m\sigma_{ab}^2$
Deviations	$\Sigma(x_{ijk} - \bar{x}_{ij.})^2$	$r_1r_2(m - 1)$	σ_e^2
Total	$\Sigma(x_{ijk} - \bar{x})^2$	$r_1r_2m - 1$	

For the three-factor experiment, the expected mean squares for the table of Sec. 7 are:

Source	Expected mean square
A effect	$\sigma_e^2 + m\sigma_{abc}^2 + mr_3\sigma_{ab}^2 + mr_2\sigma_{ac}^2 + mr_2r_3\sigma_a^2$
B effect	$\sigma_e^2 + m\sigma_{abc}^2 + mr_3\sigma_{ab}^2 + mr_1\sigma_{bc}^2 + mr_1r_3\sigma_b^2$
C effect	$\sigma_e^2 + m\sigma_{abc}^2 + mr_2\sigma_{ac}^2 + mr_1\sigma_{bc}^2 + mr_1r_2\sigma_c^2$
AB interaction	$\sigma_e^2 + m\sigma_{abc}^2 + mr_3\sigma_{ab}^2$
AC interaction	$\sigma_e^2 + m\sigma_{abc}^2 + mr_2\sigma_{ac}^2$
BC interaction	$\sigma_e^2 + m\sigma_{abc}^2 + mr_1\sigma_{bc}^2$
ABC interaction	$\sigma_e^2 + m\sigma_{abc}^2$
Deviations	σ_e^2

where σ_a^2 is the variance of the population of A main effects, σ_{ab}^2 is the variance of the population of the two-factor (AB) interaction effects, σ_{abc}^2 is that of the three-factor interaction effects, and so forth. The expected mean squares for experiments with more than three factors may be readily written down as follows: Every expected mean square involves the deviation variance with coefficient one and all other variances which have subscripts containing all the letters corresponding to the mean square in question. The coefficients of these variances are the products of the ranges of all indices on the x's except those associated with subscripts on the variances.

A very troublesome difficulty is encountered in three-factor and higher order components-of-variance models. In the present instance one obviously tests the three-factor interaction against deviations, and he tests the two-factor interactions against the three-factor interaction, but what is to be done with the main effects? On putting $\sigma_a^2 = 0$ to test the main effect of A, there is still no pair of chi squares with common divisors. If it happens that one of the two-factor interactions is zero, there is no trouble. Thus, if the hypothesis $\sigma_{ab}^2 = 0$ is not rejected, then the main effect of A may be tested against the AC interaction.

If neither of the two-factor interactions is zero, then a theoretically satisfactory test for the main effect in question can become a troublesome matter. In practice, the following simple approximation device is ordinarily employed: Suppose it is desired to test $\sigma_a^2 = 0$. Let y_1, y_2, y_3 be the sums of squares for AB interaction, AC interaction, ABC interaction; let n_1, n_2, n_3 be their respective degrees of freedom; let k_1, k_2, k_3 be their respective expected mean squares. Since y_i/k_i is a chi-square variate, the mean and variance of y_i are n_ik_i and $2n_ik_i^2$. It is evident from the above table of expected values that the variate

$$v = \frac{y_1}{n_1} + \frac{y_2}{n_2} - \frac{y_3}{n_3} \tag{2}$$

has the right mean value for an F test of $\sigma_a^2 = 0$, but v does not have the distribution of a mean square. However, if the n_i are large, the shape of the distribution of v does not differ much from the shape of the distribution of a mean square, and the approximate test treats v as if it did have such a distribution. The only question remaining is how many degrees of freedom shall be associated with v. Letting N be this number of degrees of freedom, one determines N so that the variance of the approximating distribution is the same as the variance of the actual distribution. The true variance of v is

$$\sigma_v^2 = 2\left(\frac{k_1^2}{n_1} + \frac{k_2^2}{n_2} + \frac{k_3^2}{n_3}\right) \tag{3}$$

while the variance of a mean square with N degrees of freedom and with expected value $k_1 + k_2 - k_3$ is

$$\frac{2}{N}(k_1 + k_2 - k_3)^2 \tag{4}$$

On equating (3) and (4), N is found to be

$$N = \frac{(k_1 + k_2 - k_3)^2}{(k_1^2/n_1) + (k_2^2/n_2) + (k_3^2/n_3)} \tag{5}$$

In practice, of course, the k_t are unknown, but they can be estimated by the y_t; i.e., $\hat{k}_t = y_t/n_t$. Thus the approximate test for $\sigma_a^2 = 0$ is to treat

$$\frac{mr_2r_3\Sigma(\bar{x}_{k..} - \bar{x})^2}{(r_1 - 1)\left(\dfrac{y_1}{n_1} + \dfrac{y_2}{n_2} + \dfrac{y_3}{n_3}\right)} \tag{6}$$

as an F variate with $r_1 - 1$ and N degrees of freedom, where N is determined by (5) with the k_t replaced by y_t/n_t.

14.11. Mixed Models. We are now able to investigate the mathematical model ordinarily needed to analyze data from factorial experiments. In most experiments, the levels of some factors are to be regarded as fixed constants whereas the levels of other factors must be regarded as random variables; hence the required model must be a combination of the two models already discussed. As an illustration of such a model, we shall return to the experiment described in Sec. 6. The effects of the machines will be regarded as fixed constants, while the effects of the workmen will be regarded as a sample of observations from some population of workmen.

Using the notation of Sec. 6, the observations are now regarded as being of the form

$$x_{ijk} = \xi + a_i + \beta_j + c_{ij} + e_{ijk} \tag{1}$$

where now $i = 1, 2, \cdots, r_1$; $j = 1, 2, \cdots, r_2$; $k = 1, 2, \cdots, m$. The a_i are observations from a normal population with zero mean and variance σ_a^2; the β_j are constants whose sum is zero (the average machine effect, for example, is included in ξ); the c's and e's are random observations from normal populations with zero means and variances $\sigma_{a\beta}^2$ and σ_e^2.

The sum of squares is partitioned as before into parts associated with the various factors:

$$\begin{aligned}\Sigma(x_{ijk} - \bar{x})^2 = {} & \Sigma(x_{ijk} - \bar{x}_{ij.})^2 + m\Sigma(x_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2 \\ & + mr_2\Sigma(\bar{x}_{i..} - \bar{x})^2 + mr_1\Sigma(\bar{x}_{.j.} - \bar{x})^2\end{aligned} \tag{2}$$

On substituting for the x's in the first sum, it becomes $\Sigma(e_{ijk} - \bar{e}_{ij.})^2$; hence on division by σ_e^2 this sum has the chi-square distribution with

$r_1r_2(m-1)$ degrees of freedom. The second sum becomes

$$m\sum_{ij}[c_{ij}+\bar{e}_{ij.}-(\bar{c}_{i.}+\bar{e}_{i..})-(\bar{c}_{.j}+\bar{e}_{.j.})+(\bar{c}+\bar{e})]^2$$

and replacing the $c_{ij}+\bar{e}_{ij.}$ by y_{ij}, which has variance $\sigma^2_{\alpha\beta}+(\sigma^2_e/m)$, it is evident that this term, when divided by $m\sigma^2_{\alpha\beta}+\sigma^2_e$, has a chi-square distribution with $(r_1-1)(r_2-1)$ degrees of freedom and is distributed independently of the first sum, since the deviations $e_{ijk}-\bar{e}_{ij.}$ are independent of the $\bar{e}_{ij.}$. Similarly the third sum is independent of the first two and has the chi-square distribution with r_1-1 degrees of freedom on division by $\sigma^2_e+m\sigma^2_{\alpha\beta}+mr_2\sigma^2_\alpha$.

The final sum on the right of (2) becomes

$$mr_1\sum_j(\bar{c}_{.j}+\bar{e}_{.j.}-\bar{c}-\bar{e}+\beta_j)^2$$

which is independently distributed of the other sums but does not have the chi-square distribution. However the quantity

$$mr_1\Sigma(\bar{x}_{.j.}-\bar{x}-\beta_j)^2$$

does have the chi-square distribution (on division by $\sigma^2_e+m\sigma^2_{\alpha\beta}$); hence under the null hypothesis, $\beta_j=0$, the final sum on the right of (2) does have the chi-square distribution.

The analysis-of-variance table presented here may be compared with that of Sec. 10. The final column shows at a glance what ratios

Source	Sum of squares	Degrees of freedom	Expected mean square
A effect	$mr_2\Sigma(\bar{x}_{i..}-\bar{x})^2$	r_1-1	$\sigma^2_e+m\sigma^2_{\alpha\beta}+mr_2\sigma^2_\alpha$
B effect	$mr_1\Sigma(\bar{x}_{.j.}-\bar{x})^2$	r_2-1	$\sigma^2_e+m\sigma^2_{\alpha\beta}+\dfrac{mr_1\Sigma\beta_j^2}{r_2-1}$
AB interaction	$m\Sigma(\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x})^2$	$(r_1-1)(r_2-1)$	$\sigma^2_e+m\sigma^2_{\alpha\beta}$
Deviations	$\Sigma(x_{ijk}-\bar{x}_{ij.})^2$	$r_1r_2(m-1)$	σ^2_e

of mean squares are appropriate for testing the various null hypotheses. The main effects in both cases are to be tested against interaction, not against error.

14.12. Analysis of Covariance. The analysis of covariance is a technique employed in analyzing factorial experiments when the subject of the experiment is related via a regression function to certain observable parameters. As an example of an experiment in which the method would be used, let us suppose that penetration of different kinds of steel plates by 50-caliber projectiles is being studied. Suppose there are k plates, one of each kind, and that m projectiles are to be fired at each plate. The depth to which the jth projectile penetrates the ith plate will be denoted by x_{ij}. Thus far we have a one-factor experiment with k levels and m observations per cell. But the velocity of the projectiles will be a critical factor in the depth of penetration. We shall suppose that this factor is not of interest for purposes of the present experiment; we merely wish to observe for a fixed velocity whether the resistances of the plates differ significantly. However it is impossible to fire each bullet with exactly the same velocity; and in performing the experiment, the velocity of each one will be measured photographically, and then the effects of the variations in velocity will be taken account of in the analysis of the data. Let the velocities be denoted by z_{ij}. The observations x_{ij} are now assumed to be normally distributed with variance σ^2 about the linear regression functions

$$\alpha + \beta_i z_{ij} \tag{1}$$

In the experiment just described, the observable parameter z is associated with an extraneous factor (velocity) which cannot be entirely controlled and must be dealt with in the analysis of the data. In other experiments, the observable parameter may be associated with a factor of interest. Thus in the above experiment we may desire to study the two factors—type of plate and velocity—and might vary the velocities over a considerable range. But in this latter experiment the simple linear regression function might not be adequate, and we shall restrict our illustration to the simpler situation. In more elaborate experiments, there may be several observable parameters corresponding to each of several factors for which it is impossible or inconvenient to assign specific levels. Ordinarily, when it is possible, factors are studied in experiments by assigning to them a specific set of levels rather than an observable parameter, because the analysis of the resulting data is simpler.

Returning to the illustrative example, we have a two-factor experiment in a one-way classification. One factor (type of plate) is assigned specific levels which form the one-way classification, and the

other factor (velocity) is represented by an observable parameter z. The data consist of mk pairs of observations (x_{ij}, z_{ij}) with $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, m$. We wish to test whether either of the factors affect the subject (depth of penetration), and in particular whether the plates differ when effects due to differing velocities are removed.

The sum of squares of deviations from the regression function for the observations in a single cell may be partitioned just as was done in (2.5) to obtain

$$\sum_j (x_{ij} - \alpha_i - \beta_i z_{ij})^2 = \sum_j (x_{ij} - \hat{\alpha}_i - \hat{\beta}_i z_{ij})^2 + (\hat{\beta}_i - \beta_i)^2 \sum_j (z_{ij} - \bar{z}_{i.})^2 + m(\bar{x}_{i.} - \alpha_i - \beta_i \bar{z}_{i.})^2 \quad (2)$$

where the sums on the right are independently distributed by chi-square laws (on division by σ^2) with $m - 2$, one, and one degrees of freedom, respectively. If now (2) is summed on i, the total sum of squares will be partitioned into three parts independently distributed with $k(m - 2)$, k, and k degrees of freedom, respectively. The result is

$$\sum_{ij} (x_{ij} - \alpha_i - \beta_i z_{ij})^2 = \sum_{ij} (x_{ij} - \hat{\alpha}_i - \hat{\beta}_i z_{ij})^2 + \sum_{ij} (\hat{\beta}_i - \beta_i)^2 (z_{ij} - \bar{z}_{i.})^2 + m \sum_i (\bar{x}_{i.} - \alpha_i - \beta_i \bar{z}_{i.})^2 \quad (3)$$

We shall first investigate the hypothesis that the slopes of the regression lines are the same for all cells. To this end we write

$$\beta_i = \beta + \gamma_i \quad (4)$$

and the null hypothesis then may be put in the form $\gamma_i = 0$. To test this hypothesis, the middle sum on the right of (3) is to be partitioned into two parts: one with $k - 1$ degrees of freedom involving the γ_i and the other with one degree of freedom involving β.

If we let

$$w_i = \sum_j (z_{ij} - \bar{z}_{i.})^2 \quad (5)$$

then it is apparent from the middle term on the right of (2) that $\hat{\beta}_i$ is normally distributed with mean β_i and variance σ^2/w_i. Furthermore, the $\hat{\beta}_i$ are independently distributed. If their variances were equal,

one could partition $\Sigma(\hat{\beta}_i - \beta_i)^2$ directly into $\Sigma(\hat{\beta}_i - \gamma_i - \bar{\beta})^2$ and $k(\bar{\beta} - \beta)^2$ with $k - 1$ and one degrees of freedom, but this is not the proper procedure here (see Prob. 23 at the end of Chap. 12). The deviations of the $\hat{\beta}_i$ must be taken not from their simple average but from their weighted average, say

$$\hat{\beta} = \frac{\Sigma w_i \hat{\beta}_i}{\Sigma w_i} \tag{6}$$

Furthermore, β in equation (4) must be similarly defined so that the γ_i represent deviations of the β_i from

$$\beta = \frac{\Sigma w_i \beta_i}{\Sigma w_i} \tag{7}$$

Now the middle term on the right of (3) may be partitioned thus:

$$\begin{aligned} \sum_i w_i(\hat{\beta}_i - \beta_i)^2 &= \sum_i w_i[(\hat{\beta}_i - \gamma_i - \hat{\beta}) + (\hat{\beta} - \beta)]^2 \\ &= \sum_i w_i(\hat{\beta}_i - \gamma_i - \hat{\beta})^2 + (\hat{\beta} - \beta)^2 \sum w_i \end{aligned} \tag{8}$$

since the sum of cross-product terms vanishes in view of (6) and (7). It follows from the result of Prob. 31 of Chap. 10 that the two terms on the right of (8) are independently distributed by chi-square laws with $k - 1$ and one degrees of freedom, respectively. Under the null hypothesis, $\gamma_i = 0$, the first sum on the right of (8) with the first sum on the right of (3) determines an F variate with $k - 1$ and $k(m - 2)$ degrees of freedom. The other degree of freedom on the right of (8) provides an orthogonal test of the null hypothesis, $\beta = 0$.

Turning to the third sum on the right of (3), we should like to partition it so as to get an appropriate test of the hypothesis that the α_i are all equal. Unfortunately this is not possible unless the β_i are all zero. However, it is possible to partition the sum to get some useful information about the α_i, particularly when the β_i are equal. One sets up the null hypothesis,

$$E(\bar{x}_{i.}) = \alpha + \beta' \bar{z}_{i.} \tag{9}$$

which states that the cell means $(\bar{x}_{i.}, \bar{z}_{i.})$ fall, within experimental error, on a straight line; the nature of this hypothesis will be discussed further below, but now we proceed with the partition. The third sum

on the right of (3) may be written

$$m \sum_i [(\bar{x}_{i.} - \alpha - \beta'\bar{z}_{i.}) - (\alpha_i - \alpha) - (\beta_i - \beta')\bar{z}_{i.}]^2$$
$$= m \sum_i (\bar{x}_{i.} - \delta_i - \alpha - \beta'\bar{z}_{i.})^2 \quad (10)$$

where

$$\delta_i = (\alpha_i - \alpha) + (\beta_i - \beta')\bar{z}_{i.}$$

Regarding the $\bar{x}_{i.} - \delta_i$ as a new random variable, say u_i, the sum of squares on the right of (10) may be formally partitioned just as was done in equation (2.5) to get

$$m \sum_i (\bar{x}_{i.} - \delta_i - \alpha - \beta'\bar{z}_{i.})^2 = m \sum (\bar{x}_{i.} - \delta_i - \hat{\alpha}_\delta - \hat{\beta}_\delta'\bar{z}_{i.})^2$$
$$+ m(\hat{\beta}_\delta' - \beta')^2 \sum (\bar{z}_{i.} - \bar{z})^2 + mk(\bar{x} - \bar{\delta} - \alpha - \beta'\bar{z})^2 \quad (11)$$

in which, referring to equations (13.2.8) and (13.2.9), we have

$$\hat{\alpha}_\delta = \bar{x} - \bar{\delta} - \hat{\beta}_\delta'\bar{z} \quad (12)$$

$$\hat{\beta}_\delta' = \frac{\Sigma(\bar{x}_{i.} - \delta_i - \bar{x} + \bar{\delta})(\bar{z}_{i.} - \bar{z})}{\Sigma(\bar{z}_{i.} - \bar{z})^2} \quad (13)$$

and subscripts δ have been put on these two estimators to indicate that they are functions of the unknown parameters δ_i. Under the null hypothesis that the $E(\bar{x}_{i.})$ are linear functions of the $\bar{z}_{i.}$ (i.e., that the $\delta_i = 0$), these two estimators become

$$\hat{\alpha}_0 = \bar{x} - \beta_0'\bar{z} \quad (14)$$

$$\hat{\beta}_0' = \frac{\Sigma(\bar{x}_{i.} - \bar{x})(\bar{z}_{i.} - \bar{z})}{\Sigma(\bar{z}_{i.} - \bar{z})^2} \quad (15)$$

the ordinary regression coefficients fitted to the points $(\bar{x}_{i.}, \bar{z}_{i.})$; they are therefore called the *regression coefficients for the cell means.*

The three terms on the right of (11) are independently distributed by chi-square laws on division by σ^2, the first with $k - 2$ degrees of freedom and the other two with one degree of freedom each. The null hypothesis, $\delta_i = 0$, would be tested by putting $\delta_i = 0$ in the first term on the right of (11) and comparing it with the first term on the right of (3) in an F test. The nature of this null hypothesis is illustrated on the left of Fig. 67, where the solid lines represent within-cell regressions with equations $x = \hat{\alpha}_i + \hat{\beta}_i z$, and the dashed line represents the regression of cell means $x = \hat{\alpha}_0 + \hat{\beta}_0' z$. The points on the solid lines are $(\bar{x}_{i.}, \bar{z}_{i.})$, and the null hypothesis states that the expected values

of the vertical deviations of these points from the dashed line are zero. Rejection of the null hypothesis is good evidence that the cell parameters differ. However, as the right-hand graph shows, the cell means can be linear, and even though the within-cell slopes are the same, the α_i are different. That is, one can accept $\beta_i = \beta$ and $\delta_i = 0$, yet it does not follow that the α_i are equal. However if $\beta' = \beta$, then it would follow that the α_i were all equal.

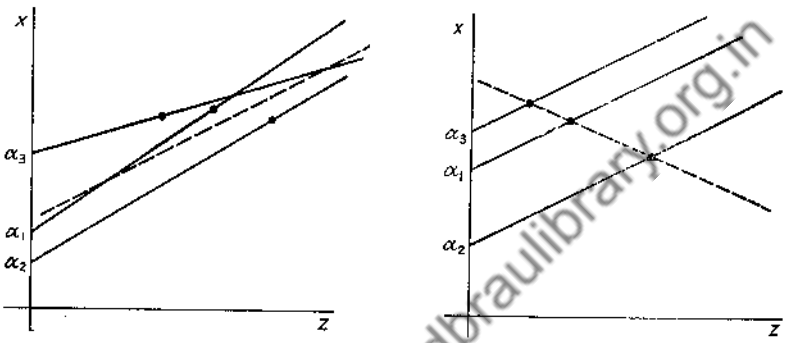


FIG. 67.

Assuming now that $\delta_i = 0$ and $\beta_i = \beta$ are acceptable hypotheses, let us construct a test for the null hypothesis $\beta' = \beta$. The random variables $\hat{\beta}$ of equation (8) and $\hat{\beta}_0'$ of equation (11) (putting $\delta_i = 0$) are independently normally distributed with means β and β', and variances $\sigma^2/\Sigma w_i$ and $\sigma^2/m\Sigma(\bar{z}_{i.} - \bar{z})^2$. Their difference is therefore distributed normally with mean $\beta - \beta'$ and variance equal to the sum of the individual variances. Hence

$$\frac{[\hat{\beta} - \hat{\beta}_0 - (\beta - \beta')]^2}{\sigma^2\left[\dfrac{1}{\Sigma w_i} + \dfrac{1}{m\Sigma(\bar{z}_{i.} - \bar{z})^2}\right]}$$

$$= \frac{[\hat{\beta} - \hat{\beta}_0' - (\beta - \beta')]^2}{\sigma^2 \sum\limits_{ij} (z_{ij} - \bar{z})^2}\, m \sum w_i \sum (\bar{z}_{i.} - \bar{z})^2 \qquad (16)$$

has the chi-square distribution with one degree of freedom. The weighted sum of $\hat{\beta}$ and $\hat{\beta}_0$

$$\Sigma w_i\hat{\beta} + m\Sigma(\bar{z}_{i.} - \bar{z})^2\hat{\beta}_0' \qquad (17)$$

is normally distributed independently of $\hat{\beta} - \hat{\beta}_0'$ [it is necessary only to show that the covariance between (17) and $\hat{\beta} - \hat{\beta}_0'$ is zero] with

mean $\Sigma w_i\beta + m\Sigma(z_{i.} - \bar{z})^2\beta'$ and variance $\sigma^2\Sigma(z_{ij} - \bar{z})^2$. Thus

$$\frac{[\Sigma w_i(\hat{\beta} - \beta) + m(\hat{\beta}_0' - \beta')\Sigma(\bar{z}_{i.} - \bar{z})^2]^2}{\sigma^2\Sigma(z_{ij} - \bar{z})^2} \tag{18}$$

has the chi-square distribution with one degree of freedom and is independent of (16). If the hypothesis $\beta = \beta'$ is accepted, then (18) provides a test of whether their common value is zero. The two independent degrees of freedom corresponding to $\hat{\beta}$ and $\hat{\beta}_0'$ in (8) and (11) have been transformed to two other independent degrees of freedom (16) and (18).

The complete partition of the sum of squares is exhibited in the accompanying table, in which all parameters have been put equal to zero. We shall review briefly the various tests:

Source	Sum of squares	Degrees of freedom
Deviations	$\sum_{ij} (x_{ij} - \hat{\alpha}_i - \hat{\beta}_i z_{ij})^2$	$k(m - 2)$
$\beta_i - \beta$	$\sum_{ij} (z_{ij} - \bar{z}_{i.})^2(\hat{\beta}_i - \hat{\beta})^2$	$k - 1$
β'	$m\sum_i (\bar{x}_{i.} - \hat{\alpha}_0 - \hat{\beta}_0'\bar{z}_{i.})^2$	$k - 2$
$\beta - \beta'$	$\dfrac{m(\hat{\beta} - \hat{\beta}_0')^2 \sum_i w_i \sum_i (\bar{z}_{i.} - \bar{z})^2}{\sum_{ij} (z_{ij} - \bar{z})^2}$	1
$\beta = \beta'$	$\dfrac{\left[\hat{\beta}\sum_i w_i + m\hat{\beta}_0'\sum_i (\bar{z}_{i.} - \bar{z})^2\right]^2}{\sum_{ij} (z_{ij} - \bar{z})^2}$	1
Total	$\sum_{ij} (x_{ij} - \bar{x})^2$	$km - 1$

1. $\beta_i - \beta = 0$. If the regression lines for the individual cells all have the same slope, then the second mean square (sum of squares divided by degrees of freedom) divided by the first mean square has the F distribution with $k - 1$ and $k(m - 2)$ degrees of freedom. If this hypothesis is rejected, then it is concluded at once that *both*

factors affect the subject of the experiment, for if the regression coefficients differ, at least one of them must be different from zero.

2. $\delta_i = 0$. The third mean square divided by the first mean square has the F distribution, *whether or not* the β_i are equal, when the cell means are linear.

3. $\beta - \beta' = 0$. The fourth mean square divided by the first has the F distribution if $\beta = \beta'$, only if it is true that $\beta_i = \beta$ and $\delta_i = 0$. One would not make this test if either of the first two null hypotheses were rejected. If all three of these null hypotheses are accepted, then it is inferred that the factor corresponding to the discrete classification does not affect the subject of the experiment (the α_i as well as the β_i are the same for all cells).

4. $\beta = \beta' = 0$. This test would be made only if all three of the other null hypotheses were accepted. If this fourth null hypothesis is accepted also, then one infers that neither of the two factors affects the subject of the experiment.

In many experiments there would be no thought of making all these tests; the primary object of the experiment might be to estimate the regression coefficients, it being well known in advance that both factors influence the subject. In such cases one would ordinarily make only the first test, in order to decide whether the same slope would suffice for all cells or whether a separate slope should be computed for each cell.

14.13. Analysis of Adjusted Means. There is one other aspect of the analysis of covariance that needs to be discussed. We may refer to the illustration at the beginning of the previous section. Suppose it is found that both factors affect the penetration; the α_i and β_i are different for the different plates, but this was to be expected anyway, and these results are of minor interest. The real question may be, Do the plates differ in their resistance for velocities $z = z_0$? (Thus z_0 may be the ordinary short-range velocity of 50-caliber bullets.) Admitted that some plates may be particularly good for very high velocities while others may be better for low velocities, how do they rank at the velocity of real interest?

Using the notation of Sec. 12, the cell means $\bar{x}_{i.}$ correspond to velocities $\bar{z}_{i.}$; in fact

$$\bar{x}_{i.} = \hat{\alpha}_i + \hat{\beta}_i \bar{z}_{i.} \tag{1}$$

With these regression coefficients we estimate that the cell means would have been

$$y_i = \hat{\alpha}_i + \hat{\beta}_i z_0 = \bar{x}_{i.} - \hat{\beta}_i(\bar{z}_{i.} - z_0) \tag{2}$$

if all the $\bar{z}_{i.}$ had been equal to z_0. The y_i are called *adjusted cell means*, and we are interested in testing the null hypothesis that the expected values of the adjusted means are the same for all cells. The y_i are independently normally distributed, as follows from equation (12.2), with variances

$$\sigma_i^2 = \sigma^2 \left[\frac{1}{m} + \frac{1}{w_i} (\bar{z}_{i.} - z_0)^2 \right] \tag{3}$$

and with means which may be denoted by η_i. Since the variances of the y_i are different, we test the null hypothesis $\eta_i = \eta$ by using the weighted sum of squares of deviations from their weighted mean, say,

$$\hat{y} = \frac{\Sigma(y_i/\sigma_i^2)}{\Sigma(1/\sigma_i^2)} \tag{4}$$

just as was done in the preceding section in testing the β_i. The sum of squares is

$$\sum \frac{1}{\sigma_i^2} (y_i - \hat{y})^2 = \frac{1}{\sigma^2} \sum \frac{m w_i (y_i - \hat{y})^2}{w_i + m(\bar{z}_{i.} - z_0)^2} \tag{5}$$

which has the chi-square distribution with $k - 1$ degrees of freedom when the η_i are equal and which is distributed independently of the first sum on the right of (12.2). Thus we have an F test for $\eta_i = \eta$.

If the first null hypothesis, $\beta_i = \beta$, of the preceding section is accepted, the $\bar{x}_{i.}$ are adjusted by the single regression coefficient $\hat{\beta}$, and the adjusted means are

$$y_i = \bar{x}_{i.} - \hat{\beta}(\bar{z}_{i.} - z_0) \tag{6}$$

The variances of the y_i then become

$$\sigma_i^2 = \sigma^2 \left[\frac{1}{m} + \frac{1}{\Sigma w_i} (\bar{z}_{i.} - z_0)^2 \right] \tag{7}$$

and equations (4) and (5) are altered accordingly. In this case the sum of squares for the denominator of the F test is often taken to be the sum of the first two sums in the table of the preceding section. Thus the deviation sum of squares would be

$$\sum_{ij} (x_{ij} - \hat{\alpha}_i - \hat{\beta}_i z_{ij})^2 + \sum_i w_i (\hat{\beta}_i - \hat{\beta})^2 = \sum_{ij} (x_{ij} - \hat{\alpha}_i - \hat{\beta} z_{ij})^2 \tag{8}$$

with $km - k - 1$ degrees of freedom.

In testing adjusted means, one would ordinarily choose $z_0 = \bar{z}$ unless there was good reason for not doing so.

14.14. Notes and References. The general field of experimental design was first thoroughly explored by Fisher, whose book [1] remains today the most important treatment of the subject. It was originally published in 1935. Yates [2] has introduced many valuable new designs. The tables of Fisher and Yates [3] describe most of the known designs and give instructions for using them.

The analysis-of-variance technique is also due to Fisher. Fisher used the test criterion $\frac{1}{2} \log F$ rather than F in his development. The latter version of the criterion is due to Snedecor, who named it F after Fisher. An excellent presentation of the practical aspects of experimental design and analysis of variance may be found in Snedecor's book [4], a large part of which is devoted to these subjects.

We have given in this chapter merely the barest introduction to the subject. Only the simplest designs have been considered, and they have not been fully analyzed. The total sum of squares may be further partitioned to study individual effects of factors and to study the linear, quadratic, cubic (and so forth) components of factors whose levels are chosen values of a continuous variate. Also the analysis was much simplified by assuming equal numbers of observations in the cells. When the cell frequencies are not equal, the analysis becomes much more tedious (except in the case of one-way classifications), primarily because the tests become nonorthogonal so that simple successive partition of the total sum of squares is no longer possible. The analysis of covariance can become quite difficult for more elaborate designs and more complicated regression functions; we have dealt only with the simplest case.

Most experimental work today is based on the rule: "Keep all variables constant but one," an ancient and erroneous dictum which guarantees a high degree of inefficiency. One well-designed experiment, taking account of all relevant factors, is worth dozens or even hundreds of experiments which study one factor at a time keeping the others constant.

1. R. A. Fisher: "Design of Experiments," 4th ed., Oliver & Boyd, Ltd., Edinburgh and London, 1945.
2. F. Yates: "Design and Analysis of Factorial Experiments," Imperial Bureau of Soil Science, Harpenden, 1937.
3. R. A. Fisher and F. Yates: "Statistical Tables," 3d ed., Hafner Publishing Co., Inc., New York, 1948.

4. G. W. Snedecor: "Statistical Methods," 4th ed., Iowa State College Press, Ames, 1947.

14.15. Problems

1. Test for differences between machines using the data of Sec. 3. The computations are usually easier to do if the sums of squares are put in forms which do not employ deviations from means. Thus, when the n_i are equal, say $n_i = m$,

$$\sum n_i(\bar{x}_{i.} - \bar{x})^2 = \frac{1}{m}\sum X_{i.}^2 - \frac{1}{n}X^2 \qquad \text{and}$$

$$\sum (x_{ij} - \bar{x}_{i.})^2 = \sum x_{ij}^2 - \frac{1}{m}\sum X_{i.}^2$$

where $X = \sum_{ij} x_{ij}$ and $X_{i.} = \sum_j x_{ij}$.

2. Use the data of Sec. 5 to test whether machine effects differ. Note that

$$c\sum (\bar{x}_{i.} - \bar{x})^2 = \frac{1}{c}\sum X_{i.}^2 - \frac{1}{rc}X^2 \qquad r\sum (\bar{x}_{.j} - \bar{x})^2 = \frac{1}{r}\sum X_{.j}^2 - \frac{1}{rc}X^2$$

and that the deviation sum of squares may be obtained by subtracting these two sums from $\Sigma x_{ij}^2 - (1/rc)X^2$.

3. Referring to Prob. 2, find a 95 per cent confidence interval for the difference between the effects of the first and third machines.

4. Four varieties of oats were compared on a block of land by dividing the block into 16 plots and using a 4×4 Latin square (chosen at random) in order to take account of possible fertility gradients in the soil. The resulting yields in pounds were found to be as follows, where the integers 1, 2, 3, 4 refer to varieties. Test for differences between variety effects. Was it worth while to use the Latin square?

3 47	4 40	2 50	1 57
2 49	1 53	3 37	4 29
4 28	3 34	1 46	2 37
1 48	2 44	4 25	3 30

5. Analyze the following data taken from a much larger table:

RETAIL PRICES OF BREAD

	New York	Chicago	Los Angeles
Chain stores and supermarkets	14, 15.5, 15, 13	14, 13, 11.5, 13	15, 15, 14, 13.5
Supermarkets (not chain)	14.5, 13, 12.5, 13	13, 13, 12, 13	13, 15, 14, 13.5
Neighborhood stores	18, 15, 15, 17	15, 15, 16, 15	16, 20, 15, 18

6. Analyze the following data:

AVERAGE NUMBER OF CHILDREN PER FAMILY

Family income	Cities		Towns		Rural Areas	
	White	Negro	White	Negro	White	Negro
Under $4,000	2.1	2.4	2.2	2.7	3.0	3.2
Over $4,000	1.5	1.8	1.8	2.1	2.5	2.9

7. A paint-manufacturing company tests new formulas for outside paint by painting 12 panels of each of three kinds of wood (36 panels in all) and exposing them for 2 years in four climates (warm dry, cold dry, warm humid, cold humid), putting three panels for each type of wood in each climate. A group of paint technologists then score the panels on a scale from 0 to 100. Analyze the following data for four formulas:

Type of wood	Climate	Formula 1	Formula 2	Formula 3	Formula 4
1	1	21, 15, 17	56, 59, 53	41, 38, 42	51, 47, 43
	2	20, 18, 19	61, 62, 62	46, 47, 45	55, 51, 54
	3	26, 30, 31	72, 67, 70	50, 48, 54	64, 63, 66
	4	31, 34, 32	66, 64, 67	54, 52, 55	64, 65, 64
2	1	24, 20, 23	54, 54, 56	39, 38, 39	50, 49, 50
	2	21, 25, 25	58, 64, 61	45, 44, 45	54, 53, 52
	3	30, 31, 31	71, 71, 71	49, 48, 53	59, 61, 60
	4	33, 34, 30	74, 71, 72	48, 56, 53	59, 62, 62
3	1	14, 17, 18	56, 55, 52	42, 40, 40	48, 49, 47
	2	21, 23, 22	61, 60, 58	46, 48, 50	53, 54, 55
	3	30, 30, 32	69, 71, 70	50, 47, 48	59, 62, 63
	4	36, 33, 35	68, 73, 77	55, 54, 51	62, 66, 64

8. A nutrition experiment studied the effects of five diets for fattening pigs for the market. Fifteen pigs, three for each diet, were put on the diets for 1 month. The following table gives the final and initial weights in pounds. Analyze the results.

Diet				
1	2	3	4	5
118, 72	102, 70	91, 63	104, 65	93, 68
108, 64	83, 55	97, 64	110, 60	79, 65
109, 63	99, 61	92, 62	95, 57	96, 69

9. A first-grade teacher with 20 pupils decided to test for herself the merits of two methods of teaching reading. The class was divided into two groups of ten and the pupils given an intelligence test (I). At the end of the year they were given a comprehensive achievement test (A) in reading. Compare the two methods.

Method 1	I	112	121	96	87	97	107	96	101	106	104
	A	81	98	71	66	65	79	63	70	71	79
Method 2	I	95	98	81	108	114	111	107	99	126	106
	A	59	60	59	72	77	79	71	63	96	68

10. Manufacturers of mass-production items often use statistical methods to control variations in the quality of their product. One technique is to take periodic samples of items from the production line and measure some critical dimension or other property (hardness, breaking strength, electrical resistance, etc.). Thus one might examine samples of size five every half hour over two 8-hour shifts, obtaining 32 samples in all. How would you use these data to test homogeneity of the production process over time, and what assumptions do you require? The null hypothesis is that no factors have crept in to alter the process—factors such as variations in incoming raw material; slipping of machine adjustments; failure of governors, thermostatic controls, etc.; differences in techniques of assembly-line workers; wear and tear on the equipment; and the like. If the null hypothesis is acceptable the process is said to be *in control.*

11. Samples of three fuses were taken every hour for 2 days from a process making 10-ampere fuses. The fuses were blown and the current measured with the following results. Is the process in control?

1	10.2, 10.1, 10.3	9	10.0, 9.8, 9.8
2	9.7, 9.9, 10.4	10	9.8, 9.7, 10.0
3	10.6, 10.1, 9.9	11	10.1, 10.1, 10.1
4	10.1, 9.8, 10.3	12	10.3, 10.2, 10.3
5	9.8, 10.0, 10.2	13	10.0, 10.2, 10.0
6	10.2, 10.1, 10.0	14	10.0, 10.1, 10.2
7	9.5, 10.1, 9.7	15	10.1, 10.1, 10.1
8	9.9, 9.9, 9.7	16	10.5, 10.2, 10.1

12. Referring to Prob. 11, let $\bar{x}$ be the mean of all observations and let s be the estimate of the standard deviation based on the within-sample deviations. Suppose now that another sample is drawn with measurements y_1, y_2, y_3. How would you test (assuming normality and common variances) the null hypothesis that $E(\bar{y}) = E(\bar{x})$?

13. In quality-control work, after a collection of samples has been analyzed, a *control chart* is constructed. The chart is simply a set of three horizontal lines drawn on graph paper at $\bar{x}$, $\bar{x} + 3s/\sqrt{m}$, $\bar{x} - 3s/\sqrt{m}$ on the vertical scale. Here s is the within-sample estimate of the standard deviation, and m is the sample size. The central line is called the *process average*, and the other two lines are called *control limits*. One continues to sample the process periodically and plots the successive sample means as points on the chart (the abscissa of the ith sample mean is i). When a point falls outside the control limits, the production process is halted and carefully examined for presence of disturbing factors. About how many times per thousand samples will the process be futilely examined if the process remains in control?

14. In the above problem, the plotting of each point constitutes a simplified test of the null hypothesis described in Prob. 12. Criticize this test. Under what circumstances would you regard the lack of independence between successive tests as not serious?

15. Verify equation (5.7) of the text.

16. Show that the expressions (5.21), (5.22), and (5.23) reduce to terms of (5.7).

17. Work through the details of the derivation of the analysis-of-variance table of Sec. 7.

18. Verify equation (8.2).

19. Referring to the components-of-variance model of Sec. 9, suppose one wished merely to estimate the variance components σ_a^2, σ_b^2, σ_c^2 and had no intention of testing hypotheses about them. Would it be necessary to assume normality? Would the obvious estimates determined from the analysis-of-variance table (by equating mean squares to expected values) necessarily be good estimates?

20. What are the maximum-likelihood estimators of σ_a^2, σ_b^2, σ_c^2 of Sec. 9?

21. Show that the four sums of squares in the first analysis-of-variance table of Sec. 10 are independently distributed by chi-square laws.

22. Derive the expected mean squares in the first analysis-of-variance table of Sec. 10.

23. Verify equations (10.3) and (10.4).

24. Derive the expected mean squares for the table of Sec. 11.

25. Show that (12.18) has the chi-square distribution and is independent of (12.16).

26. Verify equation (13.8).

27. Verify the total in the analysis-of-covariance table of Sec. 12.

28. In a two-factor experiment with each factor at two levels, it was possible to obtain only one observation for three of the cells and two for the fourth. Test for significance of the interaction.

	B_1	B_2
A_1	68	54
A_2	50	55.1, 54.9

29. Show that the analysis-of-covariance table would have been as follows had the cell frequencies been different, say m_i:

Source	Sum of squares	Degrees of freedom
Deviations	$\sum_{ij} (x_{ij} - \hat{\alpha}_i - \hat{\beta}_i z_{ij})^2$	$\Sigma m_i - 2k$
$\beta_i - \beta$	$\sum_{ij} (z_{ij} - \bar{z}_{i.})^2(\hat{\beta}_i - \hat{\beta})^2$	$k - 1$
δ_i	$\sum_i m_i(\bar{x}_{i.} - \hat{\alpha}_0 - \hat{\beta}_0'\bar{z}_{i.})^2$	$k - 2$
$\beta - \beta'$	$\dfrac{(\hat{\beta} - \hat{\beta}_0')^2 \sum_i w_i \sum m_i(\bar{z}_{i.} - \bar{z})^2}{\Sigma(z_{ij} - \bar{z})^2}$	1
$\beta = \beta'$	$\dfrac{(\hat{\beta}\Sigma m_i + \hat{\beta}_0'\Sigma m_i(\bar{z}_{i.} - \bar{z})^2}{\Sigma(z_{ij} - \bar{z})^2}$	1
Total	$\sum_{ij} (x_{ij} - \bar{x})^2$	$\Sigma m_i - 1$

30. Express all the α's and β's of the preceding table in terms of the x_{ij} and z_{ij}.

31. Test whether the regression function is of the form $\alpha + \beta z + \gamma z^2$ given the following observations (x, z) on a random variate x and an observable parameter z: (2.1, 0), (6, −1), (6, 4), (1.9, 0), (0, 2), (6.1, 4), (0.1, 1). Do not work through the arithmetic; merely specify all the steps in detail.

32. Using the data of Prob. 31, test whether the regression function is of the form $2 - 3z + z^2$.

33. Discuss the problem of testing whether the means of two samples from normal populations with the same variance are equal. Use the analysis of variance for one factor at two levels, and compare the resulting test with the one given in Sec. 12.7.

34. Consider a one-way classification with observations x_{ij}

$$(i = 1, 2, \cdots, k \quad \text{and} \quad j = 1, 2, \cdots, n_i)$$

there being unequal subclass numbers n_i. Show that the analysis-of-variance table for the components-of-variance model is:

Source	Sum of squares	Degrees of freedom	Expected mean square
Effects	$\sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2$	$k - 1$	$\sigma_e^2 + n_0\sigma_a^2$
Deviations	$\sum_{ij} (x_{ij} - \bar{x}_{i.})^2$	$N - k$	σ_e^2
Total	$\sum_{ij} (x_{ij} - \bar{x}_{..})^2$	$N - 1$	

where $N = \Sigma n_i$, σ_e^2 is the error variance, σ_a^2 is the effect variance, and

$$n_0 = \frac{1}{k - 1}\left(N - \frac{\Sigma n_i^2}{N}\right)$$

Observe also that n_0 reduces to m if all $n_i = m$.

CHAPTER 15

SEQUENTIAL TESTS OF HYPOTHESES

15.1. Sequential Analysis. *Sequential analysis* refers to techniques for testing hypotheses or estimating parameters when the sample size is not fixed in advance but is determined during the course of the experiment by criteria which depend on the observations as they occur.

In Sec. 12.2 we considered the test of a null hypothesis against a single alternative. It was shown that for samples of size n, $(x_1, x_2, \cdots, x_n)$, the test which minimizes the Type II error for fixed Type I

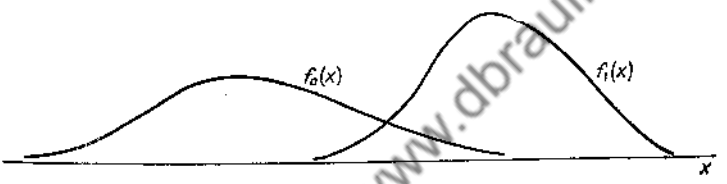


FIG. 68.

error is the likelihood-ratio test. Thus if the Type I error is chosen to be α, then α determines a number A by virtue of the equation

$$\int\int \underset{\lambda_n > A}{\cdots} \int f_0(x_1)f_0(x_2) \cdots f_0(x_n)dx_1 \cdots dx_n = \alpha \tag{1}$$

where

$$\lambda_n = \prod_1^n \frac{f_1(x_i)}{f_0(x_i)} \tag{2}$$

and the critical region for rejection of H_0 is the region

$$\lambda_n > A \tag{3}$$

This critical region minimizes the probability β (Type II error) of accepting H_0 when H_1 is true.

Suppose it is desired to fix both α and β in advance. One could do so as follows if the sample size were at his disposal: first determine A_n

as a function of n by means of (1), then determine β as a function of n,

$$\beta_n = \int\int \underset{\lambda_n < A_n}{\cdots} \int f_1(x_1)f_1(x_2) \cdots f_1(x_n)dx_1 \cdots dx_n \tag{4}$$

and finally select n so that β_n has the desired value.

Suppose further that for, say $\alpha = .01$ and $\beta = .01$, and for particular functions $f_0(x)$ and $f_1(x)$, we had worked through the computation and found n to be 100. The following considerations make sequential analysis interesting both from the theoretical and practical viewpoint: In drawing the 100 observations to test H_0, it is possible that among the first few observations there may be one or more so far to the left that eventual rejection of H_0 is out of the question and it would be a waste of time to make the remaining observations. In other instances the first 20 or first 30 or first 40 observations may provide quite sufficient evidence, relative to α and β, for accepting or rejecting H_0. In short, the possibility is raised that, by constructing the test in a fashion which permits termination of the sampling at any observation, one can test H_0 with fixed errors α and β and yet do so with fewer than 100 observations on the average. This is in fact the case, though it may at first appear surprising in view of the fact that the best test for fixed sample size does require 100 observations. The saving in observations is often quite large, sometimes more than 50 per cent. That is, in repeated tests of H_0 against H_1 for fixed control of both errors, 100 observations per test may be required for fixed sample sizes, but for sequential sampling and the same control of the errors, only 50 observations per test may be required on the average.

15.2. Construction of Sequential Tests. The theory of sequential testing has been developed only for the case of testing a null hypothesis H_0 against a single alternative H_1. It will become apparent in the later sections of the chapter that this restriction is not serious in application of the methods to practical problems. We shall let H_0 refer to a density function $f_0(x)$ and H_1 to $f_1(x)$. Observations will be denoted by $x_1, x_2, \cdots$, where the subscripts give the order in which the observations are taken.

The sequential test employs the likelihood ratio

$$\lambda_m = \prod_{i=1}^{m} \frac{f_1(x_i)}{f_0(x_i)} \tag{1}$$

and two positive numbers A and B, with $A > 1$ and $B < 1$. As observations are made, one computes the ratios $\lambda_1, \lambda_2, \lambda_3, \cdots$, and

continues taking observations as long as

$$B < \lambda_m < A \tag{2}$$

If for some m λ_m is less than or equal to B, H_0 is accepted and the test is completed. If λ_m becomes greater than or equal to A at some stage, H_0 is rejected and the test is completed. The procedure then is to continue sampling until λ_m falls outside the interval specified by (2), at which time the sampling ceases.

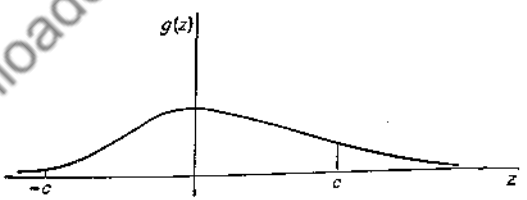
The first question that naturally arises is, What is to prevent the sampling from going on forever? It is easy to show that this cannot happen—that the probability is one that the process will terminate whatever the distribution of x. Let

$$z = \log \left[\frac{f_1(x)}{f_0(x)}\right] \tag{3}$$

then z will have some density function, say $g(z)$, determined by the density function of x [which need not be $f_0(x)$ or $f_1(x)$]. The sequence of observations $x_1, x_2, \cdots$ determines a sequence of z observations $z_1, z_2, \cdots$. The sequence of inequalities (2) becomes

$$\log B < \sum_1^m z_i < \log A \tag{4}$$

where $\log B$ is negative and $\log A$ is positive. Let $c = \log A - \log B$ and let p be the area under $g(z)$ between $-c$ and c. Now if any one of the z_i falls outside the interval $-c$ to c, one of the inequalities in (4) will necessarily be violated either at that stage or, if not then, at some



previous stage. Hence if (4) is to hold for all m, at the very least every z_i must fall between $-c$ and c. (Of course the inequalities may be violated though all the z's do fall in that interval.) The probability that every z_i falls in the interval is p^m for the first m observations (since they are independent), and this probability approaches zero as m increases, since p is less than one. Thus (4) cannot remain true indefinitely. [In case $g(z)$ is zero outside $-c$ to c, one would define

new variables y_i, letting y_1 be the sum of the first r, z's, y_2 the sum of the next r, z's, and so forth, taking r to be large enough that the nonzero range of the density function of y does not fall within $-c$ to c.]

We turn now to the determination of A and B. The probability α that H_0 will be rejected when it is true is found by computing the probability that λ_m will exceed A before it becomes less than B. It is clear that

$$\alpha = P(\lambda_1 \geq A) + P(B < \lambda_1 < A, \lambda_2 \geq A) + P(B < \lambda_1 < A, B < \lambda_2 < A, \lambda_3 \geq A) + \cdots \tag{5}$$

Similarly the probability β that H_0 will be accepted when H_1 is true is

$$\beta = P(\lambda_1 \leq B) + P(B < \lambda_1 < A, \lambda_2 \leq B) + P(B < \lambda_1 < A, B < \lambda_2 < A, \lambda_3 \leq B) + \cdots \tag{6}$$

For two specified density functions $f_0(x)$ and $f_1(x)$ one could compute all these probabilities, using $f_0(x)$ in (5) and $f_1(x)$ in (6). It follows then that α and β are known functions of A and B; hence if α and β are specified in advance, A and B are determined by (5) and (6).

As might be anticipated, the actual determination of A and B from (5) and (6) can be a major computational project. In practice, they are never determined that way because a very simple and accurate approximation is available. The approximate formulas are

$$A \cong \frac{1-\beta}{\alpha} \tag{7}$$

$$B \cong \frac{\beta}{1-\alpha} \tag{8}$$

and they arise from the following considerations. Suppose λ_m were a continuous function of a continuous variate m so that λ_m could be plotted as a curve against m, and suppose the test were performed by moving out along the m axis until λ_m first equaled A or B. That is, the test is continued as long as (2) is true and ceases when either $\lambda_m = B$ (H_0 accepted) or $\lambda_m = A$ (H_1 accepted). At all points of the $(x_1, x_2, \cdots)$ space where H_0 is accepted, the likelihood of H_1, say L_1, is exactly B times the likelihood L_0 of H_0, since $\lambda = L_1/L_0 = B$ at those points. Hence the integral of L_1 over those points is exactly equal to B times the integral of L_0 over those points. But the first integral is β, and the second is $1 - \alpha$ (the probability of accepting H_0 when it is true). So we would have β exactly equal to $B(1 - \alpha)$ if continuous

sampling were possible, and (8) would hold exactly. By a similar argument at $\lambda_m = A$, (7) would be an exact equality if m were a continuous variate. Since the error of using (7) and (8) is merely a consequence of the discreteness of m, one would expect it to be small, and analytical investigation shows that it is quite small when both α and β are less than one-half. We shall not, however, look into this matter.

Equations (7) and (8) make the actual performance of a sequential test astonishingly simple. It is not necessary to develop any sampling distribution theory at all; one merely selects α and β arbitrarily, computes A and B, and proceeds at once with the test.

15.3. Power Functions. Let a density function $f(x; \theta)$ have one parameter θ and let us test the null hypothesis, $\theta = \theta_0$, against the alternative hypothesis, $\theta = \theta_1$. We are interested in the behavior of the test for all possible values of θ. In particular, we shall examine the power function of the test, $P(\theta)$, which is the probability that θ_0 will be rejected when θ is the true parameter value. Of course

$$P(\theta_0) = \alpha \tag{1}$$

$$P(\theta_1) = 1 - \beta \tag{2}$$

and (supposing for definiteness that $\theta_0 < \theta_1$) we should expect the power function to have somewhat the shape of the curve of Fig. 69.

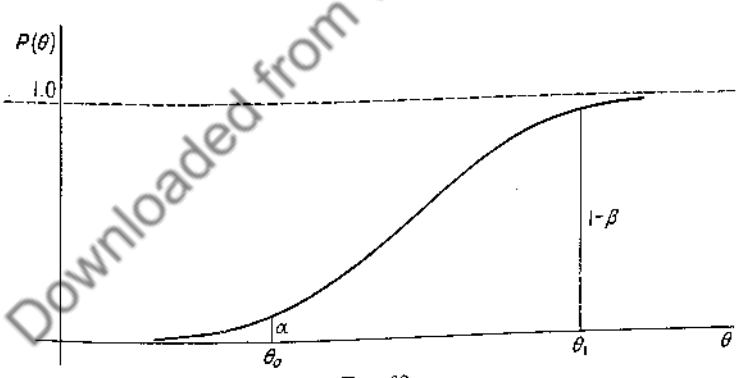


FIG. 69.

The straightforward way to compute $P(\theta)$ is simply to add the probabilities that H_0 will be rejected at each observation. Thus

$$P(\theta) = P(\lambda_1 > A) + P(B < \lambda_1 < A, \lambda_2 > A) + P(B < \lambda_1 < A, B < \lambda_2 < A, \lambda_3 > A) + \cdots \tag{3}$$

where, for example,

$$P(B < \lambda_1 < A, \lambda_2 > A) = \iint_R f(x_1, \theta) f(x_2, \theta) dx_1\, dx_2 \tag{4}$$

and the double integral is taken over the region R in the x_1, x_2 plane defined by the inequalities

$$B < \frac{f(x_1, \theta_1)}{f(x_1, \theta_0)} < A \qquad \frac{f(x_1, \theta_1) f(x_2, \theta_1)}{f(x_1, \theta_0) f(x_2, \theta_0)} > A \tag{5}$$

This procedure for determining the power function is tedious to say the least and is usually so troublesome as to be completely out of the question in practice.

To avoid the use of (3), a very ingenious device has been developed. We shall present it without a formal proof of its correctness, merely giving the general pattern of the proof. The argument requires first the existence of a nonzero number h such that

$$g(x; \theta) = \left[\frac{f(x; \theta_1)}{f(x; \theta_0)}\right]^h f(x; \theta) \tag{6}$$

is a density function; i.e., a number h such that

$$\int_{-\infty}^{\infty} g(x; \theta) dx = 1 \tag{7}$$

Of course $h = 0$ will make $g(x; \theta)$ a density function, because $f(x; \theta)$ is a density function. To show that such a nonzero value of h exists, we consider the expected value of $[f(x; \theta_1)/f(x; \theta_0)]^u$ as a function of u, say $\phi(u)$,

$$\phi(u) = \int_{-\infty}^{\infty} \left[\frac{f(x; \theta_1)}{f(x; \theta_0)}\right]^u f(x; \theta) dx \tag{8}$$

Obviously $\phi(u)$ is always positive, and furthermore $\phi(0) = 1$. We can also argue that $\phi(u)$ becomes infinite when u approaches infinity in either the positive or negative direction. Since $f(x, \theta_1)$ and $f(x, \theta_0)$ differ, there will be an interval or set of intervals where their ratio is greater than one. Over such intervals the integrand becomes large with increasing u, and $\phi(u) \to \infty$ as $u \to \infty$. Similarly there will be intervals where the inverse ratio is greater than one and the integrand becomes large for large negative value of u. This is enough to show the existence of h. (Of course, $\phi(u)$ may have a minimum at $u = 0$, in which case h would not exist, but this can happen only for particular

values of θ, not in general.) So far as our argument goes, there may be several values of u for which $\phi(u) = 1$. Actually there is only one, for the shape of $\phi(u)$ is as illustrated in Fig. 70; the minimum, though, may be to the left of the origin so that h may be negative. Thus there exists a nonzero h in general such that $\phi(h) = 1$, and (6) is therefore a density function.

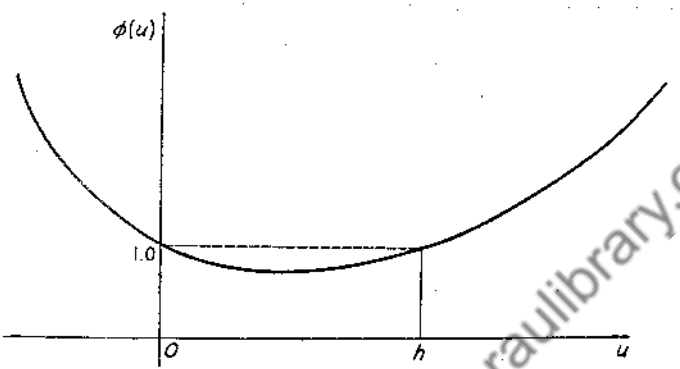


FIG. 70.

One now sets up a sequential test of the null hypothesis H_0' that the density function is $f(x;\ \theta)$ against the alternative hypothesis H_1' that the density function is $g(x;\ \theta)$. Of course the null hypothesis here is true by assumption. The limits for the likelihood ratio are taken to be A^h and B^h. Thus the test continues as long as

$$B^h < \frac{g(x_1;\ \theta)g(x_2;\ \theta)\ \cdots\ g(x_m;\ \theta)}{f(x_1;\ \theta)f(x_2;\ \theta)\ \cdots\ f(x_m;\ \theta)} < A^h \tag{9}$$

and ceases when the ratio equals or falls outside these limits. We are assuming here that h is positive; if it is negative, A and B are interchanged. In view of (6) the test defined by (9) is exactly equivalent to the original sequential test under consideration; i.e., (9) is equivalent to

$$B < \frac{f(x_1;\ \theta_1)f(x_2;\ \theta_1)\ \cdots\ f(x_m;\ \theta_1)}{f(x_1;\ \theta_0)f(x_2;\ \theta_0)\ \cdots\ f(x_m;\ \theta_0)} < A \tag{10}$$

Thus the rejection of H_0 implies the rejection of H_0'. But we can compute at once the probability that H_0' will be rejected when H_0' is true [$f(x;\ \theta)$ is the true density function]; hence we have $P(\theta)$ for $f(x;\ \theta)$. H_0' will be rejected when it is true with probability α' and accepted when H_1' is true with probability β' where, in accordance with

(2.7) and (2.8),

$$A^h \cong \frac{1 - \beta'}{\alpha'} \tag{11}$$

$$B^h \cong \frac{\beta'}{1 - \alpha'} \tag{12}$$

On solving this pair of equations for α', we find

$$\alpha' = P(\theta) \cong \frac{1 - B^h}{A^h - B^h} \tag{13}$$

Thus to find the ordinate of the power function at a point θ, one first finds the function $\phi(u)$ defined by (8) for that value of θ; then puts $\phi(u) = 1$ and solves for u; the nonzero root is the number h of (13), which then determines $P(\theta)$.

As an illustration, let us consider the null hypothesis that the mean of a normal distribution is μ_0 against the alternative that the mean is μ_1 (with $\mu_0 < \mu_1$), assuming that the variance σ^2 is known. We wish to find the probability $P(\mu)$ that μ_0 will be rejected when the true mean is μ. The function $\phi(u)$ is

$$\phi(u) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-[(x-\mu)^2/2\sigma^2]} \left(\frac{e^{-[u(x-\mu_1)^2/2\sigma^2]}}{e^{-[u(x-\mu_0)^2/2\sigma^2]}} \right) dx \tag{14}$$

The integral is easily evaluated, and on putting $\phi(u) = 1$ and solving for u, we find that one root is $u = 0$ while the other is

$$h = \frac{\mu_1 + \mu_0 - 2\mu}{\mu_1 - \mu_0} \tag{15}$$

On substituting this expression for h in (13), we have an explicit formula for $P(\mu)$ in terms of μ.

15.4. Average Sample Size. The sample size n in sequential testing is a random variable with a density function, say $p(n)$, which may be determined in terms of the true density function $f(x; \theta)$. Thus

$$p(1) = P(\lambda_1 < B) + P(\lambda_1 > A) \tag{1}$$

$$p(2) = P(B < \lambda_1 < A, \lambda_2 < B) + P(B < \lambda_1 < A, \lambda_2 > A) \tag{2}$$

and so forth, where the probabilities on the right are determined by integrals like that of equation (3.4). In this section we shall find an approximate expression for the expected sample size $E(n)$ and then illustrate the extent to which sequential methods may save observations.

Let

$$z = \log \frac{f(x; \theta_1)}{f(x; \theta_0)} \tag{3}$$

and let n be the smallest integer for which $z_1 + z_2 + \cdots + z_n = Z_n$ does not satisfy

$$\log B < Z_n < \log A \tag{4}$$

We shall show that the expected value of the variate Z_n, which depends on the random z's and the random variate n, is simply

$$E(Z_n) = E(n)E(z) \tag{5}$$

To do this, we let N be some very large but fixed value of n and disregard that part of the distribution of n to the right of N. The resulting error can be made arbitrarily small by taking N sufficiently large. Since N is fixed, it follows that

$$E(Z_N) = NE(z) \tag{6}$$

The variate Z_N may be put in the form

$$Z_N = Z_n + W_n \tag{7}$$

defining another variate W_n, and by virtue of (6)

$$E(Z_n + W_n) = NE(z) \tag{8}$$

The trouble with trying to get (5) directly is that the range of z_i depends on whether $i \leq n$ or $i > n$. In the latter case $E(z_i) = E(z)$, but when $i < n$, the range of z_i is restricted by (4). Now in (8) the variate W_n consists of z's with $i > n$, so that the expected value of each z in W_n is $E(z)$. Thus

$$E(W_n) = E(z)E(N - n) \tag{9}$$

where the second factor on the right depends only on the distribution of n. Combining (8) and (9),

$$NE(z) = E(Z_n) + E(W_n) \tag{10}$$

$$= E(Z_n) + E(z)[N - E(n)] \tag{11}$$

which is the same as (5); solving for $E(n)$,

$$E(n) = \frac{E(Z_n)}{E(z)} \tag{12}$$

This last expression enables one to get a simple approximate formula for the expected sample size. The variate Z_n takes on only values

beyond $\log A$ and smaller than $\log B$. If one ignores the amounts by which Z_n exceeds $\log A$ or falls short of $\log B$, he may say that Z_n takes essentially only two values, $\log A$ and $\log B$. When the true distribution is $f(x;\theta)$, the probability that Z_n takes the value $\log A$ is $P(\theta)$, while the probability it takes the value $\log B$ is $1 - P(\theta)$. Hence

$$E(Z_n) \cong P(\theta) \log A + [1 - P(\theta)] \log B \tag{13}$$

which together with (12) gives

$$E(n) \cong \frac{P(\theta) \log A + [1 - P(\theta)] \log B}{E(z)} \tag{14}$$

This result enables one to compare sequential tests with fixed-sample-size tests.

As an illustration, we shall consider the test that $\mu = 0$ against $\mu = 1$ for a normal population with unit variance. We shall choose $\alpha = .01$ and $\beta = .01$; then (2.7) and (2.8) give $A = 99$ and $B = 1/99$. Let us further assume that the true parameter value is zero so that $P(\theta)$ in (14) is just .01. Also we need to compute the expected value of

$$z = \log \frac{e^{-[(x-1)^2/2]}}{e^{-(x^2/2)}} = x - \frac{1}{2} \tag{15}$$

which is $-1/2$ under the true distribution. Thus

$$\begin{aligned} E(n) &\cong \frac{.01 \log 99 + .99 \log 1/99}{-1/2} \\ &\cong 1.96 \log 99 \cong 9 \end{aligned} \tag{16}$$

To get the same control of the two errors with a sample of fixed size, we recall that the best test is made by choosing a number c and accepting or rejecting $\mu = 0$ according as $\bar{x}$ is less than or greater than c. The probability α that H_0 will be rejected (under $\mu = 0$) is

$$\alpha = \sqrt{\frac{n}{2\pi}} \int_c^\infty e^{-(n/2)\bar{x}^2}\, d\bar{x} = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{n}c}^\infty e^{-t^2/2}\, dt$$

so that for $\alpha = .01$,

$$\sqrt{n}\, c = 2.326 \tag{17}$$

The probability β that H_0 would be accepted under H_1 ($\mu = 1$) is

$$\beta = \sqrt{\frac{n}{2\pi}} \int_{-\infty}^c e^{-(n/2)(\bar{x}-1)^2}\, d\bar{x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{n}(c-1)} e^{-(t^2/2)}\, dt$$

so that for $\beta = .01$,

$$\sqrt{n}\,(c - 1) = -2.326 \tag{18}$$

On solving (17) and (18) for n, we find it to be 22. Thus in repeated tests of the hypothesis in question, the sequential procedure would require on the average only $9/22$ or 41 per cent as many observations as the fixed-sample-size procedure.

15.5. Sampling Inspection. A particularly important application of sequential testing is in inspection of manufactured items. Large consumers such as retail chains, assembly plants, government agencies, and the like usually contract for periodic deliveries of items in large groups called *lots*. Certain specifications for the items in question are stipulated in the contract, and it is further stipulated that the items shall be inspected or partially inspected to ensure that only a small proportion of the delivered items fail to meet the specifications. Ordinarily, defective items are not so crucial as to warrant the expense of complete inspection of all items, and sampling inspection is used. That is, the supplier will inspect a sample of the items of a lot and estimate the proportion of the lot defective. If the quality of the lot appears satisfactory, it is delivered; otherwise it may be sold to a less exacting consumer, or to the original consumer at a lower price, or it may be completely inspected (if the inspection is not destructive) and the defective items removed. When sampling inspection is to be used, the actual sampling procedure is often a part of the contract. The supplier does not guarantee that the proportion of defective items in submitted lots will be smaller than a given amount; he merely guarantees to submit only lots which have passed a specified sampling inspection test.

The simplest sort of sampling inspection plan is the so-called *single-sampling plan*. One inspects a sample of size n and accepts the lot as satisfactory if the number of defective items is less than or equal to a given number c; otherwise the lot is rejected. The probability of accepting a lot under such a plan depends, of course, on the proportion of defectives in the lot. The density function for the number of defectives x is

$$g(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{1}$$

where N is the lot size and M is the number of defectives in the lot.

This distribution is somewhat troublesome to work with, and since n is usually quite small relative to N, it is customary to approximate the function by the binomial

$$f(x) = \binom{n}{x} p^x(1 - p)^{n-x} \tag{2}$$

where $p = M/N$ is the proportion of defectives in the lot.

The performance of a sampling inspection plan may be portrayed by the *operating-characteristic* curve, which is simply a graph of the probability of accepting the lot plotted over the range of p. This probability for the single-sampling plan is

$$L(p) = \sum_{x=0}^{c} g(x) \cong \sum_{x=0}^{c} f(x) \tag{3}$$

using the binomial approximation as we shall do in this and the next section. An operating characteristic is plotted in Fig. 71. If, for example, one wished to pass all lots with 6 per cent or less defective

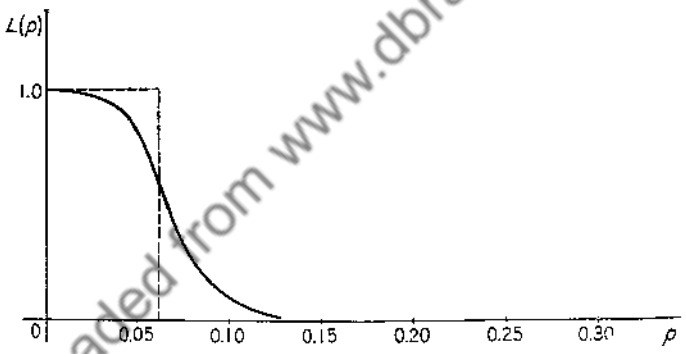


FIG. 71.

and reject all lots with more than 6 per cent defective, the ideal operating characteristic would be the dashed curve of Fig. 71. This could not be achieved without complete inspection. Sampling inspection will necessarily reject some of the acceptable lots and will accept some lots which should be rejected. The more sampling one is willing to do, the more nearly he can force the operating characteristic to approximate the ideal operating characteristic. The actual extent of the sampling in any instance depends, of course, on various economic factors associated with the particular problem at hand—factors such as production cost per item, inspection cost per item, difference in market value of accepted and rejected lots, etc.

Sampling inspection plans may be regarded as procedures for testing hypotheses. Thus the single-sampling plan is just the procedure one would use to test the null hypothesis that the parameter p of a binomial distribution has the value p_0 against alternatives $p > p_0$. Given a sample size n and a specified size α for the Type I error, the Type II error would be minimized for any $p > p_0$ by choosing the integer c for which

$$\sum_0^c \binom{n}{x} p_0^x (1 - p_0)^{n-x} = 1 - \alpha \tag{4}$$

and rejecting the null hypothesis when $x > c$. It is to be observed that the operating-characteristic function is simply one minus the power function of the test.

Somewhat more sophisticated inspection plans use *double sampling*. A small sample of size n_1 is examined, and the lot may be accepted or rejected on the basis of this sample. But in borderline cases a second sample of size n_2 is examined before the lot is finally classified one way or the other. Formally the procedure is:

1. Examine a sample of size n_1.
2. If x_1 (number of defectives in n_1) $\leq c_1$, accept the lot.
3. If $x_1 \geq c_2$, reject the lot.
4. If $c_1 < x_1 < c_2$, examine a second sample of size n_2.
5. If $x_1 + x_2 \leq c_3$, accept the lot.
6. If $x_1 + x_2 > c_3$, reject the lot.

This procedure contains the germ of the sequential idea. It is better than single sampling in the following sense: Given a single-sampling plan with sample size n and a double-sampling plan with average sample size n, one can more nearly approximate the ideal operating characteristic with the latter. Or in other words, for a given operating characteristic double sampling will require on the average fewer observations than single sampling.

15.6. Sequential Sampling Inspection. We shall suppose that large lots are being dealt with, so that the error of using the binomial distribution is of no practical importance. Let us further suppose that the supplier's production process, when all is well, produces about 2 per cent defectives and that the sampling inspection plan is supposed to accept most lots with less than 3 per cent defective and reject most lots with more than 3 per cent defective. This is the usual situation; a supplier who contracted to provide better quality than his production process was capable of would have little use for sampling inspection.

In setting up a sequential plan, one must first put the test in terms of a null hypothesis and a single alternative. Thus in the present instance one might test the null hypothesis $p_0 = .025$ against the alternative $p_1 = .04$, accepting the lot whenever the null hypothesis is accepted. In general, two values p_0 and p_1 are chosen and two probabilities α and β for the Type I and Type II errors. Thus one has at his disposal two points on the operating characteristic: $(p_0, 1 - \alpha)$ and (p_1, β). One could make the inspection plan very critical at $p = .03$ by choosing, for example, the two points (.029, .999) and (.031, .001), but in doing so he would ensure that considerable sampling would be done. The actual choice of these two points depends on economic considerations.

The individual observations y_i have the density function

$$p^y(1 - p)^{1-y} \tag{1}$$

and if $\sum_1^n y_i$ is denoted by x_n, the likelihood ratio is

$$\lambda_n = \frac{p_1^{x_n}(1 - p_1)^{n-x_n}}{p_0^{x_n}(1 - p_0)^{n-x_n}} \tag{2}$$

Observations are taken until either $\lambda_n \leq B$, in which case the lot is accepted, or $\lambda_n \geq A$, in which case the lot is rejected. A and B are computed from (2.7) and (2.8).

To get the operating characteristic, one first finds $\phi(u)$, which is simply

$$\phi(u) = E\left[\frac{p_1^y(1 - p_1)^{1-y}}{p_0^y(1 - p_0)^{1-y}}\right]^u \tag{3}$$

$$= \sum_{y=0}^{1} p^y(1 - p)^{1-y}\left(\frac{p_1}{p_0}\right)^{uy}\left(\frac{1 - p_1}{1 - p_0}\right)^{u(1-y)}$$

$$= p\left(\frac{p_1}{p_0}\right)^u + (1 - p)\left(\frac{1 - p_1}{1 - p_0}\right)^u \tag{4}$$

and the number h of Sec. 3 is the nonzero root of $\phi(u) = 1$, so that h is defined by

$$p\left(\frac{p_1}{p_0}\right)^h + (1 - p)\left(\frac{1 - p_1}{1 - p_0}\right)^h = 1 \tag{5}$$

This equation together with

$$L(p) = \frac{A^h - 1}{A^h - B^h} \tag{6}$$

[obtained by subtracting both sides of (3.13) from one] determine the operating-characteristic function. Since the solution of (5) for h is

a troublesome computation, one computes points on the curve by choosing values for h arbitrarily and calculating the corresponding values of p and $L(p)$ from (5) and (6).

Often a sufficient appraisal of the operating characteristic can be obtained from five easily computed points on the curve:

$$L(0) = 1 \tag{7}$$

$$L(1) = 0 \tag{8}$$

$$L(p_0) = 1 - \alpha \tag{9}$$

$$L(p_1) = \beta \tag{10}$$

$$L(p') = \frac{\log A}{\log A - \log B} \tag{11}$$

where

$$p' = \frac{\log [(1 - p_0)/(1 - p_1)]}{\log (p_1/p_0) - \log [(1 - p_1)/(1 - p_0)]} \tag{12}$$

The fifth point $[p', L(p')]$ is between p_0 and p_1 and corresponds to $h = 0$; the formulas (11) and (12) are obtained by letting h approach zero in (5) and (6), which become indeterminate at $h = 0$.

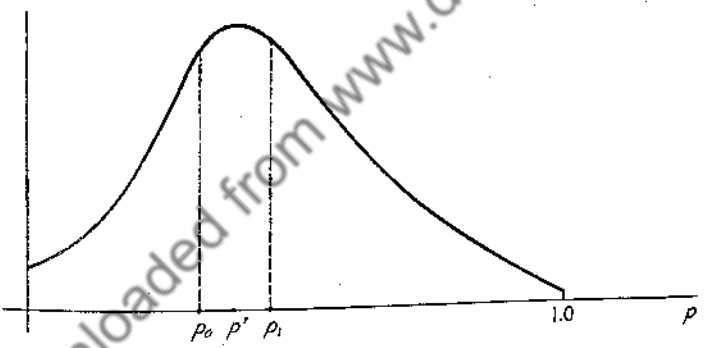


FIG. 72.

The average-sample-size curve may be plotted easily after $L(p)$ has been plotted. Referring to equation (4.15), the ordinate of this curve (Fig. 72) is given by

$$E(n) \cong \frac{[1 - L(p)] \log A + L(p) \log B}{p \log (p_1/p_0) + (1 - p) \log [(1 - p_1)/(1 - p_0)]} \tag{13}$$

where we have substituted $1 - L(p)$ for $P(p)$ and

$$E(z) = E\left[\frac{\log p_1^y(1 - p_1)^{1-y}}{p_0^y(1 - p_0)^{1-y}}\right] \tag{14}$$

$$= p \log \frac{p_1}{p_0} + (1 - p) \frac{1 - p_1}{1 - p_0} \tag{15}$$

The maximum value of $E(n)$ occurs very nearly at the point p' given by (12). At that point, (13) becomes an indeterminate form whose limiting value is

$$\frac{\log A \log B}{\log (p_1/p_0) \log [(1 - p_1)/(1 - p_0)]} \tag{16}$$

This is approximately the maximum average sample size and occurs when the true proportion defective has the value given by (12).

15.7. Sequential Test for the Mean of a Normal Population. As a final example of sequential testing, we shall consider the two-sided test of the null hypothesis H_0 that the mean of a normal population has the value μ_0. It is assumed that the variance σ^2 is known. It is necessary to frame the test in terms of a single alternative H_1. If we were interested in a one-sided test, say against alternatives $\mu > \mu_0$, we should simply choose some arbitrary value μ_1 (greater than μ_0) for the alternative. But that alternative will not serve for the two-sided test, because the power function approaches zero as μ moves to the left.

The trick here is to phrase the hypotheses in terms of another parameter δ which measures the distance of μ from μ_0. The new parameter δ takes only positive values and is defined by

$$\delta = \mu - \mu_0 \qquad \text{if } \mu > \mu_0 \tag{1}$$

$$\delta = \mu_0 - \mu \qquad \text{if } \mu < \mu_0 \tag{2}$$

The null hypothesis is now $\delta = 0$, and the alternative is $\delta = \delta_1$, where δ_1 is an arbitrarily chosen number. Now one must set up a somewhat artificial alternative distribution function, because the number δ_1 actually refers to two distributions—one with mean $\mu_0 - \delta_1$ and one with mean $\mu_0 + \delta_1$. The alternative density function is defined to be

$$f_1(x) = \frac{1}{2} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-[(x-\mu_0+\delta_1)^2/2\sigma^2]} + \frac{1}{2} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-[(x-\mu_0-\delta_1)^2/2\sigma^2]} \tag{3}$$

which is clearly a density function. Under H_0 the density function is, of course,

$$f_0(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-[(x-\mu_0)^2/2\sigma^2]} \tag{4}$$

It is apparent that the likelihood ratio will behave as we wish. If μ is to the left of $\mu_0 - \delta_1$, the ratio f_1/f_0 will usually be large because of the first term on the right of (3), while if $\mu > \mu_0 + \delta_1$, it will be large because of the second term.

The test is now performed in accordance with the usual procedure. One chooses probabilities α and β for the two types of error and computes A and B from (2.7) and (2.8). For a very sensitive test one

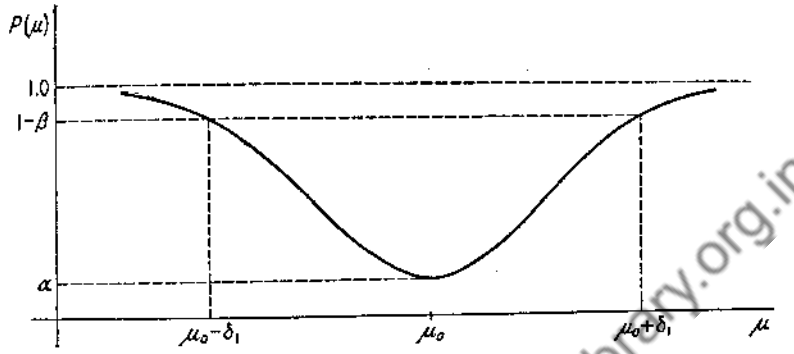


FIG. 73.

would choose δ_1 as well as α and β to be small. Observations are made until

$$\lambda_n = \Pi \frac{f_1(x_i)}{f_0(x_i)} \tag{5}$$

exceeds A or becomes less than B.

The test given here for H_0 is merely one of many possibilities. We have been quite arbitrary in setting up the alternative density function, and it is entirely conceivable that some other form might improve the test, might reduce the average sample size under the null hypothesis, for example, or might have other desirable properties.

When the variance is unknown, several tests are available; most of them use weight functions of one kind or another. Perhaps the simplest test is that based on the t distribution. If we denote by $g_n(t; \mu)$ the density function for t with n degrees of freedom and with μ the mean of the normal population, then one may define

$$\lambda_n = \frac{g_n(t; \mu_1)}{g_n(t; \mu_0)} \tag{6}$$

with $n = 2, 3, 4, \cdots$. Although this function is not of the same type as the others we have considered (because the numerator and denominator are not products of density functions of independent variates), it can be shown that the test terminates and that (2.7) and (2.8) determine A and B as before.

The criterion (6) refers, of course, to the one-sided test of μ_0 against an alternate μ_1 greater than μ_0. For a two-sided test of μ_0, one may use

$$\lambda_n = \frac{\frac{1}{2}g_n(t;\, \mu_0 + \delta_1) + \frac{1}{2}g_n(t;\, \mu_0 - \delta_1)}{g_n(t;\, \mu_0)} \tag{7}$$

where δ has the same meaning as in (1) and (2).

15.8. Notes and References. Sequential analysis is a quite recent development in the theory of statistics, having been started in 1943. The theory is due primarily to Wald [1], whose excellent and quite readable book on the subject contains most of the developments made up to the date of its publication. Wald's work has stimulated much research, and the techniques of sequential analysis will doubtless be extended considerably during the next few years.

Thus far, most attention has been given to the matter of testing hypotheses, but sequential methods also promise to increase the efficiency of estimation procedures. The problem here is to choose in advance a $1 - \alpha$ confidence interval of specified length and make observations until the confidence interval can be said to cover the true parameter value with the desired probability.

The matter of testing composite hypotheses requires further development. Wald has shown that this problem may be dealt with by means of certain weight functions chosen in an optimum fashion. But a detailed general theory is not yet available.

A good exposition of sampling inspection from the practical point of view is given by the second reference.

1 A. Wald: "Sequential Analysis," John Wiley & Sons, Inc., New York, 1947.

2. H. A. Freeman, M. Friedman, F. Mosteller, and W. A. Wallis: "Sampling Inspection," McGraw-Hill Book Company, Inc., New York, 1948.

15.9. Problems

1. Perform a sequential test of the null hypothesis that $p = .45$ against the alternative that $p = .30$. Let p refer to the probability of a head in tossing a coin, and carry through the test by tossing a coin using $\alpha = .10$ and $\beta = .10$. The arithmetic is simplified by solving $\log \lambda_n = B$ and $\log \lambda_n = A$ for x_n (the number of heads in n tosses), thus obtaining acceptance and rejection numbers as linear functions of n.

2. Show that equation (3.13) is correct when h is negative.

3. Assuming a lot has size N with M defectives, what is the exact expression for the operating-characteristic function?

4. Show that the ratio γ_1/γ_2 obtained at the end of Sec. 4 depends only on the values of α and β and not on the sizes of σ^2, μ_0, and μ_1.

5. Compare the average sequential sample size with the fixed sample size for the one-sided test of the mean of a normal population when $\alpha = .01$, $\beta = .05$, and the alternative hypothesis is true.

6. Show that the one-sided test for the mean of a normal population with known variance may be performed by plotting the two lines

$$y = \frac{\sigma^2}{\mu_1 - \mu_0} \log B + \frac{\mu_0 + \mu_1}{2} n$$

$$y = \frac{\sigma^2}{\mu_1 - \mu_0} \log A + \frac{\mu_0 + \mu_1}{2} n$$

in the n, y plane; then plotting $\sum_1^n x_i$ against n as the observations are made. The test ends when one of the lines is crossed.

7. Referring to Prob. 6, let $c = (\mu_0 + \mu_1)/2$ and let the two constants in the equations be denoted by b and a; i.e.,

$$a = \frac{\sigma^2 \log A}{\mu_1 - \mu_0}$$

Show that the power function for the test may be put in the form

$$P(\mu) \cong \frac{1 - e^{2(c-\mu)b/\sigma^2}}{e^{2(c-\mu)a/\sigma^2} - e^{2(c-\mu)b/\sigma^2}}$$

8. Referring to Probs. 6 and 7, show that the expression for the average sample size may be written

$$E(n) \cong \frac{b + P(\mu)(b - a)}{\mu - c}$$

9. Verify equations (6.11) and (6.12).

10. Plot the power function and average-sample-size function for the test of Prob. 1.

11. Plot the power function and the average-sample-size function for the test that the mean of a normal population is zero against the alternative that it is one. Let $\sigma^2 = 1$, $\alpha = .01$, $\beta = .05$.

12. Find formulas for the power function and average sample size for sequential tests on the mean of a Poisson distribution.

13. Suppose a production process produces lots of size N with M defectives in such a way that M has a binomial distribution. Show that a sample of size n (with x defectives) can provide no information about the proportion of defectives in the remaining $N - n$ items of a lot.

14. Suppose lots which are rejected under a sequential sampling inspection procedure are completely inspected and the defective items replaced by good items; this is a common practice. Let p be the proportion of defectives in the original lots. What will be the average proportion of defectives over all delivered lots counting both those completely inspected and those passed by the sampling plan? This function of p is called the *average outgoing quality* function; the maximum of the function is called the *average outgoing quality limit.* Make a rough sketch showing the general shape of the function.

15. Referring to the situation described in Prob. 14, find the average percentage of items inspected as a function of p, counting both passed and completely inspected lots. Make a rough sketch showing the general shape of the function.

16. Suppose a uniform distribution has the range $0 < x < \theta$. Discuss the sequential test of $\theta = \theta_0$ against $\theta = \theta_1$ with $\theta_0 < \theta_1$. Be careful here; some of the general formulas may not be applicable.

17. By an argument similar to that used to obtain (4.5), Wald has shown that

$$E\{e^{Z_n t}[\phi(t)]^{-n}\} = 1$$

where $\phi(t)$ is the moment generating function of z, i.e., $\phi(t) = E(e^{zt})$, and where the expectation E is over the joint distribution of the z's and the random variable n. This is called the fundamental identity of sequential analysis. Use it to obtain (4.5).

18. Use the identity of Prob. 17 to show that

$$E(n) = \frac{E(Z_n^2)}{E(z^2)}$$

when $E(z) = 0$.

19. Use the result of Prob. 18 to obtain (6.16).

20. Use the result of Prob. 18 to show that the maximum average sample size for one-sided tests of the mean of normal population is approximately $-ab/\sigma^2$, where a and b are defined in Prob. 7. Assume, do not try to prove, that the maximum occurs at $h = 0$.

CHAPTER 16

DISTRIBUTION-FREE METHODS

16.1. Introduction. In Sec. 7.8 the important place ascribed to the normal distribution in statistical theory was justified on the basis that any known continuous distribution could be transformed to the normal distribution. But, of course, experimenters quite frequently have no knowledge of the form of the distribution with which they are dealing, or at least so little information that they cannot prescribe a normalizing transformation. Until recently there was not much to be done in this situation, and experimenters were more or less forced to make wholesale assumptions of normality. During the past few years, however, techniques have been developed for estimating parameters and testing hypotheses which require no assumption about the form of the distribution function. These techniques are called *nonparametric methods*, or better, *distribution-free methods*. While the collection of distribution-free methods is not nearly so comprehensive as that based in normal theory, a good beginning has been made, and this chapter will present some of the results.

Heretofore in denoting a sample by $x_1, x_2, \cdots, x_n$, the symbol x_1 referred to the first observation made, x_2 to the second, and so on. Throughout this chapter the notation will be interpreted quite differently. The symbol x_1 will refer to the smallest of the n observations, x_2 will represent the second smallest of the observations, and so on, with x_n the largest. Thus, for the sample of four observations, 2, -4, -1, 1, x_1 refers to the second observation, x_2 to the third, x_3 to the fourth, and x_4 to the first. The phrase *ordered sample* is often used to indicate this interpretation of the notation. Distribution-free methods are based entirely on these ordered observations, or *order statistics*.

The methods to be presented are applicable to both continuous and discrete variates, but we shall direct our attention almost entirely to the continuous case, merely pointing out occasionally the modifications that would be required in the case of discrete variates.

16.2. A Basic Distribution. The whole structure of distribution-free methods rests on a simple property of order statistics: the distribution of the area under the density function between any two ordered

observations is independent of the form of the density function. To show this, we merely make the probability transformation described in Sec. 6.1. The density function for the ordered sample $x_1, x_2, \cdots, x_n$ is

$$n!f(x_1)f(x_2) \cdots f(x_n) \tag{1}$$

if $f(x)$ is the population density function. The factor $n!$ arises from the fact that there are $n!$ permutations of the observations and every permutation gives rise to the same ordered sample. The density for any given permutation is just $\Pi f(x_i)$, so the density for the ordered sample is obtained by summing this expression over all permutations of the x_i. The variates in (1) are restricted by the inequalities

$$-\infty < x_1 < x_2 < x_3 \cdots < x_n < \infty \tag{2}$$

If we let

$$u_i = \int_{-\infty}^{x_i} f(x) = F(x_i) \tag{3}$$

then in accordance with Sec. 6.1, the density function for the u_i is simply

$$g(u_1, u_2, \cdots, u_n) = n! \qquad 0 < u_1 < u_2 < \cdots < u_n < 1 \tag{4}$$

which does not depend on $f(x)$.

The density function $g(u_1, \cdots, u_n)$ enables one to find the distribution of any set of areas under $f(x)$ between pairs of ordered observations. For example, suppose we desire the density function for the area under $f(x)$ between x_1 and x_n. This area, say v, is

$$v = F(x_n) - F(x_1) = u_n - u_1 \tag{5}$$

We first integrate out $u_2, u_3, \cdots, u_{n-1}$ in (4), then make the substitution $u_n = u_1 + v$ and integrate out u_1. Thus

$$h(u_1, u_n) = \int_{u_1}^{u_n} \cdots \int_{u_1}^{u_4}\int_{u_1}^{u_3} n!\, du_2\, du_3 \cdots du_{n-1} \tag{6}$$

$$= n!\frac{(u_n - u_1)^{n-2}}{(n-2)!} \qquad 0 < u_1 < u_n < 1 \tag{7}$$

and the density of u_1 and v is

$$k(u_1, v) = n(n-1)v^{n-2} \qquad 0 < u_1 < (1 - v) < 1 \tag{8}$$

On integrating out u_1, we obtain the required density

$$m(v) = n(n-1)v^{n-2}(1 - v) \qquad 0 < v < 1 \tag{9}$$

which is a beta density function.

More generally, the area w between x_r and x_s $(s > r)$ under $f(x)$ can readily be shown to have the density function

$$\frac{n!}{(s-r-1)!(n-s-r)!} w^{s-r-1}(1-w)^{n-s+r} \qquad 0 < w < 1 \tag{10}$$

Also one can obtain a joint density function for several such areas.

The expected value of u_i is

$$\begin{aligned} E(u_i) &= \int_0^1 \cdots \int_0^{u_3}\int_0^{u_2} n!u_i\, du_1\, du_2 \cdots du_n \\ &= \frac{i}{n+1} \end{aligned} \tag{11}$$

hence the expected area under $f(x)$ between two successive observations is

$$E(u_i) - E(u_{i-1}) = \frac{1}{n+1} \tag{12}$$

Thus, on the average, the n ordered observations divide the area under $f(x)$ into $n + 1$ equal parts of area $1/(n + 1)$ each.

16.3. Location and Dispersion. In the parametric case we have used the population mean and standard deviation as measures of location and dispersion. The distribution-free methods use other measures. The center of the population is defined to be the median, say ν, which is the point that divides the area under the frequency function in half. Thus ν is defined by

$$\tfrac{1}{2} = \int_{-\infty}^{\nu} f(x)dx = F(\nu) \tag{1}$$

where $f(x)$ is the density function and $F(x)$ is the cumulative distribution. The median is often denoted by $\xi_{.50}$, and a similar notation is used for other percentage points; thus

$$F(\xi_{.15}) = .15 \tag{2}$$

defines the 15 per cent point, $\xi_{.15}$, of the population.

As a measure of dispersion one uses the distance between two percentage points. Thus, one frequently used measure of dispersion is

$$\tau_{.50} = \xi_{.75} - \xi_{.25} \tag{3}$$

which is called the 50 *per cent range*, or the *interquartile range*. But many other ranges are often used, for example, the 90 per cent range $\tau_{.90} = \xi_{.95} - \xi_{.05}$, or the 33⅓ per cent range $\tau_{1/3} = \xi_{2/3} - \xi_{1/3}$.

Point Estimation. The population median ν is estimated by the sample median $\tilde{x}$, which is the middle observation if the sample size is odd or the average of the two middle observations if the sample size is even. Thus

$$\tilde{x} = x_{k+1} \qquad \text{if } n = 2k + 1 \tag{4}$$

$$= \tfrac{1}{2}(x_k + x_{k+1}) \qquad \text{if } n = 2k \tag{5}$$

The sample median $\tilde{x}$ is not ordinarily an unbiased estimate of ν even when n is odd, for the fact that $E[F(\tilde{x})] = F(\nu)$ does not imply that $E(\tilde{x}) = \nu$. However the bias is not serious and must approach zero as the sample size increases.

To estimate percentage points, the x_i themselves furnish estimates of the $100i/(n + 1)$ per cent points. For other values one may use linear interpolation. Thus to estimate $\xi_{.25}$ from a sample of size $n = 10$, we observe that x_2 estimates the $\frac{2}{11}$ point and x_3 the $\frac{3}{11}$ point; hence we use as the estimate

$$\tilde{x}_{.25} = x_2 + \frac{.25 - \frac{2}{11}}{\frac{1}{11}}(x_3 - x_2) \tag{6}$$

Given estimates of percentage points, one can obviously estimate the various ranges.

Confidence Intervals. A confidence interval for ν is easily constructed by means of the binomial distribution. The probability that an observation falls to the left or right of ν is one-half in either case. The probability that exactly i observations fall to the left of ν is just

$$\binom{n}{i}\left(\frac{1}{2}\right)^n \tag{7}$$

The probability that x_r, the rth-order statistic, exceeds ν is then

$$P(x_r > \nu) = \sum_{i=0}^{r-1} \binom{n}{i}\left(\frac{1}{2}\right)^n \tag{8}$$

and similarly

$$P(x_s < \nu) = \sum_{i=s}^{n} \binom{n}{i}\left(\frac{1}{2}\right)^n \tag{9}$$

If we now suppose $s > r$, add (8) and (9), and subtract both sides from unity, we have

$$P(x_r < \nu < x_s) = \sum_{i=r}^{s-1} \binom{n}{i}\left(\frac{1}{2}\right)^n \tag{10}$$

which provides a confidence interval for ν. Ordinarily s is taken to be $n - r + 1$ so that the rth observations in order of magnitude from the top and from the bottom are used. Thus for a sample of size 6, two possible confidence intervals for the median are

$$P(x_1 < \nu < x_6) = 1 - (\tfrac{1}{2})^6 - (\tfrac{1}{2})^6 = 31/32 \cong .97 \tag{11}$$

and

$$P(x_2 < \nu < x_5) = 1 - \frac{14}{2^6} = \frac{50}{64} \cong .78 \tag{12}$$

If one wished to do so, he could approximate a 95 or 90 per cent confidence interval by using linear interpolation between (11) and (12), but this is rarely done in practice. One ordinarily restricts himself to the confidence levels available with the simple order statistics.

If the sample size is small, one has only a few confidence levels available; in particular, when $n = 2$, there is only the 50 per cent confidence interval given by

$$P(x_1 < \nu < x_2) = .50 \tag{13}$$

For moderate sample sizes the binomial sum in (10) may be computed directly or found in tables of the incomplete beta function. For large n one would use the normal approximation to the binomial. Since the index i in (7) is approximately normal with mean $n/2$ and standard deviation $\sqrt{n}/2$ for large n, a 95 per cent confidence interval, for example, is obtained by counting $1.96\sqrt{n}/2$ observations to the left and right of the sample median.

A similar technique is employed to obtain confidence intervals for percentage points. If ξ_p is the $100p$ per cent point of the distribution, then the same argument used to obtain (10) shows that

$$P(x_r < \xi_p < x_s) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \tag{14}$$

Thus for a sample of size 6, a possible confidence interval for the 25 per cent point is given by

$$P(x_1 < \xi_{.25} < x_4) = \sum_{1}^{3} \binom{6}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{6-i} \cong .78 \tag{15}$$

A 96 per cent upper bound for $\xi_{.25}$ is given by

$$P(\xi_{.25} < x_4) = \sum_{0}^{3} \binom{6}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{6-i} \cong .96 \tag{16}$$

Tests of Hypotheses. To test the null hypothesis $\nu = \nu_0$ against alternatives $\nu > \nu_0$, one uses the relation (8), choosing in advance an integer r so that the probability of a Type I error will have as nearly as possible the desired value. Thus for a sample of size 6 one can make the probability of a Type I error $\frac{7}{64} \cong .11$ by choosing $r = 2$. If after drawing the sample one finds $x_2 < \nu_0$, the null hypothesis is accepted; if $x_2 > \nu_0$, it is rejected. In the same fashion two-sided tests of $\nu = \nu_0$ may be constructed; the two-sided test is obviously equivalent to constructing a confidence interval for ν and accepting or rejecting $\nu = \nu_0$ according as the confidence interval does or does not cover ν_0. Tests on a percentage point ξ_p would be carried out by the same technique, using probabilities p and $(1 - p)$ instead of $\frac{1}{2}$ and $\frac{1}{2}$.

It is now apparent that the distribution-free methods, besides being extremely general in that they require no assumption about the form of the distribution function, are also extraordinarily simple. No complex analysis or distribution theory is needed; the simple binomial provides all the necessary equipment for estimation and testing hypotheses when one is dealing with a single population. The only inconvenience is in the paucity of significance levels or confidence levels when the sample size is quite small.

A word about the discrete case is in order here. We have assumed the frequency function was continuous. If it is discrete, then the equalities obtained in this section for confidence intervals and tests need to be replaced by inequalities. Thus (10), for example, becomes

$$P(x_r < \nu < x_s) \geq \sum_{r}^{s-1} \binom{n}{i} \left(\frac{1}{2}\right)^n \tag{17}$$

The reason for the inequality is in the fact that certain observations may be duplicated. Thus suppose one wished to estimate ν for a discrete distribution using a sample of size 6 and a 78 per cent confidence interval given by x_2 and x_5. Now and then the two smallest observations x_1 and x_2 will be equal so that the (x_2, x_5) interval is equivalent to the (x_1, x_5) interval and hence corresponds to a probability larger than .78. The same thing may happen at the upper limit; x_5 and x_6 may be equal so that sometimes the (x_2, x_5) interval is equivalent to the (x_2, x_6) interval; occasionally it can even be the same as the (x_1, x_6) interval and thus correspond to the 97 per cent rather than the 78 per cent level.

16.4. Comparison of Two Populations. A great many distribution-free methods have been developed for testing whether two populations

have the same distribution. We shall consider only two of them, and at the end of this section we shall derive a confidence interval for the difference between two population medians. First, we shall obtain a simple result on the distribution of arrangements of two sets of observations from the same population.

Let $x_1, x_2, \cdots, x_{n_1}$ be an ordered sample from a population with a density function $f(x)$, and let $y_1, y_2, \cdots, y_{n_2}$ be a second ordered sample from the same population. Let the two samples be combined and arranged in order of magnitude; thus, for example, one might have

$$y_1, x_1, x_2, y_2, x_3, y_3, y_4, y_5, x_4, \cdots \tag{1}$$

We wish to find the probability of obtaining a specific arrangement of this kind.

If the x's are transformed to u's by the relation (2.3), and the y's transformed to v's by the same relation, the joint frequency function of the u's and v's is

$$g(u_1, u_2, \cdots, u_{n_1}, v_1, v_2, \cdots, v_{n_2}) = n_1!n_2! \tag{2}$$

The probability of a given arrangement such as (1) is found by integrating (2) over the region defined by

$$0 < v_1 < u_1 < u_2 < v_2 < u_3 < \cdots < 1 \tag{3}$$

i.e., v_1 is integrated from zero to u_1, then u_1 from zero to u_2, etc. It is readily seen that the value of the integral is $n_1!n_2!/(n_1 + n_2)!$, or simply $1\Big/\binom{n_1 + n_2}{n_1}$. Since there are exactly $\binom{n_1 + n_2}{n_1}$ arrangements of n_1 x's and n_2 y's, it follows that all arrangements of the x's and y's are equally likely.

Run Test. We now turn to the question of testing the null hypothesis that two samples have come from the same population. The observations in the two samples will be denoted by x's and y's as above. The two sets of observations are combined as in (1) and the number d of runs counted. A run is a sequence of letters of the same kind bounded by letters of the other kind. Thus (1) starts with a run of one y; then follows a run of two x's, then a run of one y, and so on; six runs are exhibited in (1). It is apparent that if the two samples are from the same population, the x's and y's will ordinarily be well mixed and d will be large. If the two populations are widely sepa-

rated so that their ranges do not overlap, d will be only two, and, in general, differences between the two populations will tend to reduce d. Thus the two populations may have the same mean or median, but if the x population is concentrated while the y population is dispersed, there will be a long y run on each end of the combined sample and there will thus be a tendency to reduce d. The test then is performed by observing the total number of runs in the combined sample, accepting the null hypothesis if d is greater than some specified number d_0, or rejecting the null hypothesis if $d \leq d_0$. Our task now is to determine the distribution of d under the null hypothesis in order that we may specify d_0 for a given level of significance.

We have seen that all of the $\binom{n_1 + n_2}{n_1}$ arrangements of n_1 x's and n_2 y's are equally likely under the null hypothesis. It is necessary now to count all arrangements with exactly d runs. Suppose d is even, say $2k$; then there must be k runs of x's and k runs of y's. To get k runs of x's, the n_1 x's must be divided into k groups, and we wish to count all permutations of the k numbers in each group. In short, we wish to count all the ordered k-part partitions of n_1 with zero parts excluded. This is readily done with the aid of the generating function described in Sec. 2.6 for enumerating the ways of getting a given total with a set of dice. The required number is the coefficient of t^{n_1} in

$$(t + t^2 + t^3 + \cdots)^k = \left(\frac{t}{1 - t}\right)^k \tag{4}$$

$$= t^k \sum_{i=0}^{\infty} \binom{k - 1 + i}{k - 1} t^i \tag{5}$$

which is $\binom{n_1 - 1}{k - 1}$. Similarly there are $\binom{n_2 - 1}{k - 1}$ k-part partitions of n_2, excluding zero parts. Any partition of the x's may be combined in any partition of the y's in two ways to form a sequence like (1); the first x partition or the first y partition may be put at the beginning of the sequence. Thus we have found the density for even values of d:

$$h(d) = 2\frac{\binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_1}} \qquad d = 2k \tag{6}$$

and by similar reasoning one finds for odd values of d:

$$h(2k+1) = \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n_1+n_2}{n_1}} \tag{7}$$

To test the null hypothesis in question with a probability p for the Type I error, one finds the integer d_0 so that (as nearly as possible)

$$\sum_{d=0}^{d_0} h(d) = p \tag{8}$$

and rejects the null hypothesis if the observed d does not exceed d_0.

The computation involved in (8) can become quite tedious unless both n_1 and n_2 are small. The distribution of d becomes approximately normal for large samples, and in fact the approximation is usually good enough for practical purposes when both n_1 and n_2 exceed 10. The mean and variance of $h(d)$ are

$$E(d) = \frac{2n_1n_2}{n_1+n_2} + 1 \tag{9}$$

$$\sigma_d^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \tag{10}$$

and if we let

$$n_1 + n_2 = n \qquad n_1 = n\alpha \qquad n_2 = n\beta \tag{11}$$

these moments become, for large n, approximately

$$E(d) \cong 2n\alpha\beta \tag{12}$$

$$\sigma_d^2 \cong 4n\alpha^2\beta^2 \tag{13}$$

The large-sample normality of $h(d)$ is demonstrated by using Stirling's formula to evaluate the factorials in (6), substituting for d in terms of t defined by

$$t = \frac{d - 2n\alpha\beta}{2\alpha\beta\sqrt{n}} \tag{14}$$

and showing that the logarithm of the resulting expression approaches

$$-\log\sqrt{2\pi} - \tfrac{1}{2}t^2$$

as n becomes infinite. We shall omit the details. Using this result,

one would determine d_0 for testing the null hypothesis at the .05 level, for example, by putting the right-hand side of (14) equal to -1.645 and solving for d.

The run test is sensitive to both differences in shape and differences in location between two distributions. Often, however, in practice, one does not care about differences in shape; he is concerned only with location. That is to say, he would like to test merely the null hypothesis that the population medians are equal: $\nu_1 = \nu_2$. It is not possible to make such a test, but the following test of $f_1(x) = f_2(y)$ is sensitive primarily to differences in location and very little to differences in shape.

Median Test. As before, let there be an ordered sample $x_1, x_2, \cdots, x_{n_1}$ from $f_1(x)$ and a sample $y_1, y_2, \cdots, y_{n_2}$ from $f_2(y)$. Let $z_1, z_2, \cdots, z_{n_1+n_2}$ be the ordered combined sample. The test of the null hypothesis $f_1(x) = f_2(x) = f(x)$ will consist in finding the median $\tilde{z}$ of the combined sample, then counting the number of x's, say m_1, which exceed $\tilde{z}$ and the number of y's, say m_2, which exceed $\tilde{z}$. If the null hypothesis is true, we should expect m_1 to be approximately $n_1/2$ and m_2 approximately $n_2/2$. It is clear that this test will be sensitive to differences in location between $f_1(x)$ and $f_2(x)$ but not to differences in their shape. Thus if $f_1(x)$ and $f_2(x)$ have the same median, we should expect the null hypothesis to be accepted ordinarily even though their shapes were quite different.

To make this test, the distribution of m_1 and m_2 under the null hypothesis is required. Let z_a be the ath observation in order of magnitude, let m_1 be the number of x's which exceed z_a, and let m_2 be the number of y's which exceed z_a. The joint density function of m_1, m_2, z_a under the null hypothesis is

$$\left\{\frac{n_1!}{m_1!(n_1 - m_1 - 1)!}[F(z_a)]^{n_1 - m_1 - 1}[1 - F(z_a)]^{m_1}\, dF(z_a)\right\}$$
$$\left\{\binom{n_2}{m_2}[F(z_a)]^{n_2 - m_2}[1 - F(z_a)]^{m_2}\right\} + \left\{\binom{n_1}{m_1}[F(z_a)]^{m_1 - m_1}[1 - F(z_a)]^{m_1}\right\}$$
$$\left\{\frac{n_2!}{m_2!(n_2 - m_2 - 1)!}[F(z_a)]^{n_2 - m_2 - 1}[1 - F(z_a)]^{m_2}\, dF(z_a)\right\} \quad (15)$$

where the first term takes account of the case in which z_a is an x observation and the second term of that in which z_a is a y observation; $F(x)$ is the cumulative form of $f(x)$, and $dF(z_a)$ represents $f(z_a)\Delta z_a$. On integrating out z_a and combining the two resulting terms, one finds

the frequency function for m_1 and m_2 to be, say,

$$g(m_1, m_2) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1 + n_2}{a}} \tag{16}$$

We observe, on comparing this expression with equation (12.10.17), that it is just the distribution of the cell frequencies in a 2×2 contingency table with all marginal totals fixed when there is independence. The contingency table is

m_1	m_2	$n_1 + n_2 - a$
$n_1 - m_1$	$n_2 - m_2$	a
n_1	n_2	

where the marginal totals are shown to the right of and below the closed part of the table. If $n_1 + n_2$ were odd, one would choose $a = (n_1 + n_2 + 1)/2$, whereas if the sum were even, one would choose $a = (n_1 + n_2)/2$. Thus the null hypothesis may be tested by using either the λ criterion given by (12.10.8) or the chi-square criterion given by (12.10.20). If $n_1 + n_2$ were small, one would use (16) to compute the exact probabilities instead of using the approximate probability given by the chi-square distribution with one degree of freedom. The approximation is fairly good if both n_1 and n_2 exceed 10.

Confidence Intervals. In order to obtain exact confidence intervals for the difference between the medians of two populations, it is necessary to assume that the distributions differ only in location. Letting $x_1, x_2, \cdots, x_{n_1}$ be a sample from a population with median ν_1, and $y_1, y_2, \cdots, y_{n_2}$ a sample from one with median ν_2, we assume that the variates

$$u_i = x_i - \nu_1 \qquad \text{and} \qquad v_i = y_i - \nu_2$$

have the same density function, say $f(u)$, with median zero. The sample of u's and the sample of v's are then two samples from the same population. If one chooses two integers r and s, he may compute the probability that u_r exceeds v_s as follows:

$$P(u_r > v_s) = \int_{-\infty}^{\infty} s\binom{n_2}{s}[F(v_s)]^{s-1}[1 - F(v_s)^{n_2-s}\, dF(v_s) \sum_{i=0}^{r-1}\binom{n_1}{i} [F(v_s)]^i[1 - F(v_s)]^{n-i} \tag{17}$$

$$= \sum_{i=0}^{r-1} \frac{s\binom{n_2}{s}\binom{n_1}{i}}{(s+i)\binom{n_1+n_2}{s+i}} \tag{18}$$

$$= \sum_{i=0}^{r-1} \frac{\binom{s+i-1}{s-1}\binom{n_1+n_2-s-i}{n_2-s}}{\binom{n_1+n_2}{n_1}} \tag{19}$$

Similarly

$$P(u_{r'} < v_{s'}) = \sum_{i=r'}^{n_1} \frac{\binom{s'+i-1}{s'-1}\binom{n_1+n_2-s'-i}{n_2-s'}}{\binom{n_1+n_2}{n_1}} \tag{20}$$

If we now suppose $r < s$, $r' > s'$, and $\nu_2 > \nu_1$, then

$$P(y_{s'} - x_{r'} < \nu_2 - \nu_1 < y_s - x_r)$$

$$= P(u_r < v_s \text{ and } u_{r'} > v_{s'}) \tag{21}$$

$$= 1 - P(u_r > v_s) - P(u_{r'} < v_{s'}) \tag{22}$$

and the left-hand side of this relation provides a confidence interval for $\nu_2 - \nu_1$ with a confidence level which is calculable by means of (19) and (20). The confidence interval provides a third test of the null hypothesis that the two distributions are the same; the hypothesis would be rejected if the interval did not include zero.

We shall outline a large-sample approximation which may be used when n_1 and n_2 both exceed 10. Since the sum expressed in (19) is one when taken over the whole range of i, we may regard the summand as a density function for a variate i and find the normal approximation to that function. The sum may then be approximated by integrating the approximating function. The mean and variance of i are

$$E(i) = \frac{sn_1}{n_2+1} \tag{23}$$

$$\sigma_i^2 = \frac{sn_1}{n_2+1}\left[\frac{(s+1)(n_2+3)}{n_2+2} + \frac{(s+1)n_1}{n_2+2} - (2s+1) - \frac{sn_1}{n_2+1}\right] \tag{24}$$

and their approximate values when n_1 and n_2 are large may be found by letting

$$n_1 + n_2 = n \qquad n_1 = n\alpha \qquad n_2 = n\beta \qquad s = \gamma n_2 = \beta\gamma n \tag{25}$$

and keeping only terms involving the highest power of n. The results are

$$E(i) \cong n\alpha\gamma \tag{26}$$

$$\sigma_i^2 \cong n\alpha\gamma \frac{1-\gamma}{\beta} \tag{27}$$

The large-sample normality of i may be proved in the same manner as outlined for d in (6). The sum in (19) may then be approximated by

$$\int_{-\infty}^{A} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \tag{28}$$

where

$$A = \frac{(r - 1 + \frac{1}{2}) - n\alpha\gamma}{\sqrt{n\alpha\gamma(1-\gamma)/\beta}} \tag{29}$$

Given s, one would choose A to give the desired probability level (-1.96, for example, to make the probability .025) and solve for r.

The question arises as to how s should be chosen. Clearly s should be greater than $n_2/2$ and r should be less than $n_1/2$. One might, for example, make the two differences equal, but a shorter confidence interval may be expected by making the differences equal on "standard" scale. The number of x observations less than v_1 is approximately normally distributed with mean $n_1/2$ and standard deviation $\sqrt{n_1}/2$; similarly the number of y observations exceeding v_2 is approximately normally distributed with mean $n_2/2$ and standard deviation $\sqrt{n_2}/2$. We shall then determine s so that

$$\frac{(n_1/2) - r}{\sqrt{n_1}} = \frac{s - (n_2/2)}{\sqrt{n_2}} \tag{30}$$

If one substitutes for n_1, n_2, and s in this relation in terms of (25) and solves for r, then equates the result to the solution of (29) for r, he finds

$$\gamma \cong \frac{1}{2} + \frac{A}{2\sqrt{n}\,(\sqrt{\alpha\beta} + \beta)} \tag{31}$$

neglecting terms with higher powers of $1/\sqrt{n}$; in terms of the original symbols this becomes

$$s \cong \frac{n_2}{2} + \frac{A\sqrt{n_2}\sqrt{n_1 + n_2}}{2(\sqrt{n_1} + \sqrt{n_2})} \tag{32}$$

and from (30)

$$r \cong \frac{n_1}{2} - \frac{\Lambda \sqrt{n_1}\sqrt{n_1 + n_2}}{2(\sqrt{n_1} + \sqrt{n_2})} \tag{33}$$

In similar fashion one would argue that good choices for r' and s' in (22) are given by changing the signs in (32) and (33).

16.5. A Distribution-free Test for One-factor Experiments.* A factor is tested at k levels with n_i observations at the ith level; the observations may be denoted by x_{ij} with $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, n_i$. The null hypothesis that the factor has no effect will be tested by testing in fact whether all the $n = \Sigma n_i$ observations may be regarded as coming from the same population. Ordinarily in practice one is not much concerned with whether the cell distributions differ in shape; he is primarily concerned with whether they differ in location. Hence the test we shall consider will be a generalization of the second test given in the preceding section.

Let m_i be the number of observations in the ith cell which exceed the median of the whole set of n observations and construct the contingency table:

m_1	m_2	$\cdots$	m_k	a
$n_1 - m_1$	$n_2 - m_2$	$\cdots$	$n_k - m_k$	$n - a$
n_1	n_2		n_k	

where $a = n/2$ if n is even or $(n - 1)/2$ if n is odd. It is easily shown by the argument used in the preceding section that the density function for the m_i is

$$g(m_1, m_2, \cdots, m_k) = \frac{\prod_{i=1}^{k} \binom{n_i}{m_i}}{\binom{n}{a}} \tag{1}$$

This is just the ordinary distribution for a $2 \times k$ contingency table with all marginal totals fixed when there is independence. Hence the null hypothesis may be tested by means of the λ criterion or the chi-square criterion of Sec. 12.10. The chi-square criterion is ordinarily easier to use, and using the present notation, it may be put in the form

$$\chi^2 = \frac{n(n-1)}{a(n-a)} \sum_{i=1}^{r} \frac{1}{n_i}\left(m_i - \frac{n_i a}{n}\right)^2 \tag{2}$$

*Sections 16.5 through 16.9 are based in part on unpublished work of George W. Brown (see Preface).

where we have retained a factor $n - 1$ in the numerator of the coefficient instead of replacing it by n, as is usually done, because n is assumed large. The expression (1) has a distribution very accurately approximated by the chi-square distribution with $r - 1$ degrees of freedom even if n is only of the order of twenty provided all the n_i are at least five. For smaller values of the n's one should compute exact probabilities from (1).

To estimate the difference between main effects of the factor at two levels, one would use the difference of the cell medians, and a confidence interval for the difference would be constructed by the method described in the preceding section.

16.6. Two-factor Experiments, One Observation per Cell. The observations are denoted by x_{ij} with $i = 1, 2, \cdots, r$ and

$$j = 1, 2, \cdots, c$$

The row factor is thus being tested at r levels and the column factor at c levels. The distributions of the x_{ij} have medians

$$\nu_{ij} = \nu + \alpha_i + \beta_j \tag{1}$$

where the median of the numbers α_i is zero as is the median of the β_j. The α_i and β_j are identified with row and column effects. The distributions of the x_{ij} are assumed to be identical except for location; thus the variates $x_{ij} - \nu_{ij}$ are all supposed to have the same density, say $f(u)$. Also the x's are assumed to be continuous variates. If one or both the factors have randomly chosen levels, we may suppose that the density takes account of random interaction effects as well as error effects. Otherwise it is necessary to assume the interactions are zero.

We shall examine the null hypothesis that the row effects, α_i, are zero. Under this hypothesis all the observations in a given column have the same distribution. Let $\tilde{x}_j$ be the median of the observations in the jth column, and in the two-way table let the observation x_{ij} be replaced by a plus sign if it exceeds $\tilde{x}_j$ or by a minus sign if it does not. The $r \times c$ table then consists of plus and minus signs in equal number if r is even, or with c more minus signs than plus signs if r is odd. Let m_i be the number of plus signs in the ith row. If there are in fact no row effects, then we should expect the m_i to differ from $c/2$ only by random sampling deviations, but if there are row effects, then the rows with positive effects would have an excess of plus signs while those rows with a negative effect would have a deficiency of plus signs. The null hypothesis is therefore tested by testing whether the signs are

divided evenly in rows. In fact, we may construct a $2 \times r$ contingency table:

m_1	m_2	$\cdots$	m_r	ca
$c - m_1$	$c - m_2$	$\cdots$	$c - m_r$	$c(r - a)$

where $a = r/2$ if r is even or $(r - 1)/2$ if r is odd. It turns out that the m_i do not have the ordinary contingency-table distribution as was the case in the preceding section. However the distribution of the m_i is such that the large-sample distribution of an analogous chi-square criterion has, in fact, the chi-square distribution with $r - 1$ degrees of freedom. So this table may be tested like an ordinary contingency table with all marginal totals fixed.

The distribution of the m_i is best exhibited in the form of a generating function; the distribution itself does not have a simple closed form. Suppose we let t_1 be associated with a plus sign in the first row, t_2 with a plus sign in the second row, and so forth. Let $\phi_a(t_1, t_2, \cdots, t_r)$ consist of the sum of all terms that can be formed by multiplying the t's together a at a time. Thus, for example,

$$\phi_2(t_1, t_2, t_3, t_4) = t_1t_2 + t_1t_3 + t_1t_4 + t_2t_3 + t_2t_4 + t_3t_4 \tag{2}$$

Each term of $\phi_a(t_1, \cdots, t_r)$ describes a possible arrangement of signs in a given column. Furthermore it is easily argued that each arrangement of signs is equally likely; hence the probability of a particular arrangement is $1 \Big/ \binom{r}{a}$. Now we consider the function

$$\phi = \left[\frac{\phi_a(t_1, t_2, \cdots, t_r)}{\binom{r}{a}}\right]^c \tag{3}$$

A little reflection will convince one that there is a one-to-one correspondence between ways of getting terms $t_1^{m_1}t_2^{m_2} \cdots t_r^{m_r}$ in the numerator of ϕ and arrangements of signs in the $r \times c$ table which give rise to $m_1, m_2, \cdots, m_r$ plus signs in the respective rows. Hence

$$\phi = \sum_{m_1}\sum_{m_2} \cdots \sum_{m_r} g(m_1, m_2, \cdots, m_r)t_1^{m_1}t_2^{m_2} \cdots t_r^{m_r} \tag{4}$$

where g is the density function for the m_i. On putting all the $t_i = 1$, ϕ becomes one, since the sum in (4) is then just the sum of the density over its whole space; this is evident from (3) also, since

$$\phi_a(1, 1, \cdots, 1) = \binom{r}{a},$$

there being $\binom{r}{a}$ terms in $\phi_a(t_1, t_2, \cdots, t_r)$.

It is evident from (4) that ϕ is a factorial-moment generating function for the m_i. Thus,

$$E(m_1) = \frac{d\phi}{dt_1} \quad \text{with all } t_i = 1 \tag{5}$$

$$= c\phi_{a-1}(t_2, t_3, \cdots, t_r) \frac{[\phi_a(t_1, t_2, \cdots, t_r)]^{c-1}}{\binom{r}{a}^c} \quad \text{at } t_i = 1 \tag{6}$$

$$= c\frac{\binom{r-1}{a-1}}{\binom{r}{a}} \tag{7}$$

$$= \frac{ca}{r} \tag{8}$$

which is the same for all m_i, and similarly the variances and covariances of the m_i are found to be

$$\sigma_{ii} = \frac{ca(r-a)}{r^2} \tag{9}$$

$$\sigma_{ij} = -\frac{ca(r-a)}{r^2(r-1)} \qquad i \neq j \tag{10}$$

Taking m_r to be the dependent variate (they are related by $\Sigma m_i = ca$), the matrix of variances and covariances for $m_1, m_2, \cdots, m_{r-1}$ may be inverted to get

$$\sigma^{ii} = \frac{2r(r-1)}{ca(r-a)} \tag{11}$$

$$\sigma^{ij} = \frac{r(r-1)}{ca(r-a)} \qquad i \neq j \tag{12}$$

We shall not demonstrate that the m_i are asymptotically normally distributed. The simplest proof appeals to a generalization of the central-limit theorem. If variates $y_1, y_2, \cdots, y_{r-1}$ are distributed with finite variances and covariances, γ_{ij}, then it can be shown that

the averages $\bar{y}_i$ for a large sample of size c are approximately normally distributed with variances and covariances γ_{ij}/c. In the present instance, y_1 would be defined to be one or zero according as there was or was not a plus sign in the first cell of a column, and similarly for the other y's. The c columns are then regarded as c observations on the y_i, and the $\bar{y}_i$ are then m_i/c, which by the general theorem must be asymptotically normal as c becomes large.

A more direct proof of normality for large c could be constructed by replacing the t_i in ϕ by $e^{s_i/\sqrt{c}}$ (except $t_r = 1$), then showing that $\log \phi$ approaches, as c becomes large, the expression

$$\sum_{i=1}^{r-1} s_i \frac{\sqrt{c}\, a}{r} + \frac{1}{2} \sum_{ij} \frac{\sigma_{ij}}{c} s_i s_j \tag{13}$$

the exponent of e in the moment generating function of a normal distribution, as shown by equation (9.5.4). The quadratic form of the large-sample normal distribution will have the chi-square distribution with $r - 1$ degrees of freedom; the quadratic form is

$$\chi^2 = \sum_{1}^{r-1} \sum_{1}^{r-1} \sigma^{ij} \left(m_i - \frac{ca}{r}\right)\left(m_j - \frac{ca}{r}\right) \tag{14}$$

and it may be reduced to the expression

$$\chi^2 = \frac{r(r-1)}{ca(r-a)} \sum_{1}^{r} \left(m_i - \frac{ca}{r}\right)^2 \tag{15}$$

The ordinary chi-square criterion given by equation (12.10.20), if applied to the $2 \times r$ table at the beginning of this section, would differ from (15) only in that the numerator of the coefficient of the sum would be r^2 instead of $r(r-1)$. Here r is not assumed to be large so the difference may be appreciable.

The null hypothesis that the row effects are zero may therefore be tested by the criterion (15), using the ordinary chi-square distribution unless c is small. For practical purposes the large-sample distribution is satisfactory if c is as large as 10, or even if c is only 5 provided rc is 20 or more; for smaller values the exact probability should be computed by means of (3). To test column effects, one would, of course, simply reverse the roles of rows and columns in the above test.

16.7. Two-factor Experiments, Several Observations per Cell. We shall suppose that there are r rows, c columns, and h observations

per cell. The observations are denoted by x_{ijk} with $i = 1, 2, \cdots, r$; $j = 1, 2, \cdots, c$; and $k = 1, 2, \cdots, h$. It is assumed that the variates are continuous and have the same distribution except for location; if the population cell medians are ν_{ij}, then the variates $x_{ijk} - \nu_{ij}$ all have the same distribution. The ν_{ij} may be put in the form

$$\nu_{ij} = \nu + \alpha_i + \beta_j + \gamma_{ij} \tag{1}$$

where the α_i have zero median, the β_j have zero median, and the γ_{ij} have zero medians in every row and column. If the levels of a factor are randomly chosen, then the effects are random variables and are regarded as having a zero population median rather than a zero median themselves.

Main Effects against Interaction. To make the test analogous to the test of main effects against interaction in the ordinary analysis of variance, one simply finds the cell medians $\tilde{x}_{ij}$ (median of the h observations in the i, j cell) and uses the tests presented in the preceding section on these cell medians.

Joint Tests of Main Effects and Interaction. By using a procedure similar to that of the preceding section it is possible to construct a simple test of the hypothesis that a factor has no effect whatever, either in main effects or in interaction effects. Thus we shall consider the null hypothesis: $\alpha_i = 0$ and $\gamma_{ij} = 0$. Let $\tilde{x}_j$ represent the median of all rh observations in the jth column, and let m_{ij} be the number of observations in the i, j cell which exceed $\tilde{x}_j$. Considering a specific column, we have just the one-factor situation discussed in Sec. 5, and the m_{ij} have the density

$$\frac{\prod_{i=1}^{r} \binom{h}{m_{ij}}}{\binom{rh}{a}} \tag{2}$$

where $a = rh/2$ or $(rh - 1)/2$ whichever is an integer. The density for all the m_{ij} is therefore obtained by taking the product of (2) over j from one to c. We need not, however, deal with this distribution expect in the case of small numbers. To test the null hypothesis, one would compute the chi square of equation (5.2) for each column and add the results to obtain

$$\chi^2 = \frac{r(rh - 1)}{a(rh - a)} \sum_{i,j} \left(m_{ij} - \frac{a}{r}\right)^2 \tag{3}$$

which is approximately distributed as chi square with $c(r-1)$ degrees of freedom, and the approximation is satisfactory enough for most practical purposes if h is 5 or more, or if rch is 20 or more.

Main Effects against Deviations. The distribution-free test analogous to the analysis-of-variance test of main effects against deviations is fairly simple if interactions are assumed to be zero. Referring to the numbers m_{ij} of the paragraph above, let

$$\sum_j m_{ij} = n_i \tag{4}$$

i.e., n_i is the number of observations in the ith row which exceed their column medians. There being ch observations in a row, we should expect the n_i to be roughly $ch/2$ under the null hypothesis. The hypothesis is tested by means of a chi-square criterion much like that of the preceding section. We shall merely outline its derivation. The $2 \times r$ contingency table here is:

n_1	n_2	$\cdots$	n_r	ca
$ch - n_1$	$ch - n_2$	$\cdots$	$ch - n_r$	$c(ch - a)$

but it does not have the ordinary contingency-table distribution.

A factorial-moment generating function for the n_i is

$$\phi(t_1, t_2, \cdots, t_r) = \text{coefficient of } \prod_j x_j^a \text{ in } \prod_{i,j} \frac{(1 + x_i t_i)^h}{\binom{rh}{a}^c} \tag{5}$$

Using this, one finds the means, variances, and covariances of the n_i to be

$$E(n_i) = \frac{ca}{r} \tag{6}$$

$$\sigma_{ii} = \frac{ca(r-1)(rh-a)}{r^2(rh-1)} \tag{7}$$

$$\sigma_{ij} = -\frac{ca(rh-a)}{r^2(rh-1)} \qquad i \neq j \tag{8}$$

The inverse of the variance-covariance matrix for $i = 1, 2, \cdots, r-1$ is found to be

$$\sigma^{ii} = \frac{2r(rh-1)}{ca(rh-a)} = 2\sigma^{ij} \tag{9}$$

and the chi-square criterion is

$$\chi^2 = \frac{r(rh-1)}{ca(rh-a)} \sum_i \left(n_i - \frac{ca}{r}\right)^2 \tag{10}$$

with $r - 1$ degrees of freedom.

Interaction. All the tests described thus far are computationally quite simple—merely a matter of counting observations and computing a chi square. To test the null hypothesis that the γ_{ij} of (1) are zero, it is necessary first to remove both row and column main effects by an iterative reduction; then one proceeds with a test similar to those already described.

Letting $\tilde{x}_j$ be the column medians as before, one removes the column effects to a first approximation by subtracting the $\tilde{x}_j$ from the observations of the jth column to get a reduced set of observations:

$$x'_{ijk} = x_{ijk} - \tilde{x}_j \tag{11}$$

One then finds the row medians $\tilde{x}'_i$ and subtracts these out to get

$$x''_{ijk} = x'_{ijk} - \tilde{x}'_i \tag{12}$$

If the plus and minus signs are balanced in the columns (they will obviously be balanced in the rows), the reduction is complete. But ordinarily the subtraction of the row medians will upset the balance of signs in the columns, and it is necessary to find the column medians $\tilde{x}''_j$ of the x''_{ijk} and subtract these out to get

$$x'''_{ijk} = x''_{ijk} - \tilde{x}''_j \tag{13}$$

This process is continued until both rows and columns have zero medians. One could, of course, start the reduction with the row medians of the original observations rather than the column medians.

After the reduction is completed, one counts the number of plus signs, m_{ij}, in each cell, counting the zeros as one-half plus and one-half minus. The numbers m_{ij} and $h - m_{ij}$ form a $2 \times r \times c$ contingency table with all marginal totals fixed, and the null hypothesis may be tested by the ordinary chi-square criterion for testing independence in such a table. This interaction test is very nearly but not completely distribution free. The approximate chi-square criterion is

$$\chi^2 = m^2 h \sum_{ij} \frac{[m_{ij} - (m_{i.}m_{.j}/m)]^2}{m_{i.}m_{.j}(h - m_{i.}m_{.j})} \tag{14}$$

with $(r - 1)(c - 1)$ degrees of freedom where

$$m_{i.} = \sum_j m_{ij} \qquad m_{.j} = \sum_i m_{ij} \qquad m = \sum_{i,j} m_{ij} \tag{15}$$

The expression simplifies somewhat if $m_{i.} = ch/2$ and $m_{.j} = rh/2$, but this will not always be the case owing to the presence of zeros among the reduced observations.

All the elements for testing in factorial experiments have been presented in this discussion of the two-factor experiment. The methods carry over directly to more complicated situations. The general rule for testing one factor or set of factors assuming the presence of a second set of factors is to fit the second set of factors using medians, then classify the data according to the first set of factors and test for fifty-fifty splits between positive and negative deviations in the various classifications. All these tests are special cases of tests in the general linear regression problem which will be described briefly in Sec. 9.

16.8. Simple Linear Regression. A continuous variate x has a density $f(x)$ whose median is of the form

$$\nu = \alpha + \beta z \tag{1}$$

where α and β are unknown parameters and z is an observable parameter. On the basis of a sample of n observations, (x_1, z_1), (x_2, z_2), $\cdots$, (x_n, z_n), it is desired to estimate α and β or test hypotheses regarding α and β.

Point Estimation. Supposing the paired observations to be plotted as n points in the x, z plane, the problem here is to fit a regression line of the form

$$x = \alpha + \beta z \tag{2}$$

to the plotted points. If we denote the estimates of α and β by $\tilde{\alpha}$ and $\tilde{\beta}$, the two conditions which determine $\tilde{\alpha}$ and $\tilde{\beta}$ are

$$\text{Median of } (x_i - \tilde{\alpha} - \tilde{\beta} z_i) = 0 \qquad \text{for } z_i \leq \tilde{z} \tag{3}$$

$$\text{Median of } (x_i - \tilde{\alpha} - \tilde{\beta} z_i) = 0 \qquad \text{for } z_i > \tilde{z} \tag{4}$$

where $\tilde{z}$ is the median of the z_i. Thus one divides the observations into two groups, using the median of the z's, and chooses that line which makes the median of the deviations zero in each group. (If it happens that several z values fall at $\tilde{z}$, then the $\leq$ sign in (3) and $>$ sign in (4) would be replaced by $<$ and $\geq$ if such a replacement would more nearly divide the points into groups of equal size.)

In practice, unless the number of observations is quite large, the simplest method of determining the line is to plot the points and use a transparent ruler to locate the line by eye. For machine work, the following iterative procedure may be used: Find the medians of the x's and z's in each of the two groups. The slope of the line joining the two points determined by these four medians is a first approximation, say β', to $\tilde{\beta}$. Let the deviations of the x_i from the line, $x = \beta' z$, be

$$x_i' = x_i - \beta' z_i \tag{5}$$

A slope δ' is fitted to these deviations in the same manner as above to get a correction to β'. The second approximation to $\tilde{\beta}$ is

$$\beta'' = \beta' + \delta' \tag{6}$$

Now new deviations

$$x_i'' = x_i - \beta'' z_i = x_i' - \delta' z_i \tag{7}$$

are computed and a slope δ'' fitted to them. The third approximation to $\tilde{\beta}$ is

$$\beta''' = \beta'' + \delta'' \tag{8}$$

and the iteration continues until $\tilde{\beta}$ is determined to the desired degree of accuracy. Then $\tilde{\alpha}$ is the median of the final set of deviations.

Tests of Hypotheses. To test the null hypothesis, $\alpha = \alpha_0$ and $\beta = \beta_0$, one divides the points into two groups at $\bar{z}$ and tests whether the two groups are both evenly divided by the line. Let m_1 be the number of points above the line for $z_i \leq \bar{z}$, and let m_2 be the number of points above the line for $z_i > \bar{z}$. Both m_1 and m_2 have the binomial distribution with parameter one-half; hence

$$\chi^2 = \frac{8}{n}\left[\left(m_1 - \frac{n}{4}\right)^2 + \left(m_2 - \frac{n}{4}\right)^2\right] \tag{9}$$

will have approximately the chi-square distribution with two degrees of freedom, unless n is small, in which case one would use the exact distribution to compute the probability.

To test $\alpha = \alpha_0$ only, one would fit a line, $x = \alpha_0 + \tilde{\beta} z$, to the points, determining $\tilde{\beta}$ by the condition

$$\underset{z_i \leq \bar{z}}{\text{Median}}\,(x_i - \alpha_0 - \tilde{\beta} z_i) = \underset{z_i > \bar{z}}{\text{median}}\,(x_i - \alpha_0 - \tilde{\beta} z_i) \tag{10}$$

The number of points, m, above the fitted line (in both groups combined) has the binomial distribution with mean $n/2$ under the null hypothesis.

To test $\beta = \beta_0$, one would fit a line, $x = \tilde{\alpha} + \beta_0 z$, to the points determining $\tilde{\alpha}$ by

$$\tilde{\alpha} = \text{median } (x_i - \beta_0 z_i) \tag{11}$$

The points are again divided into two groups on $\tilde{z}$ and the numbers m_1 and m_2 of points above the line in each group counted. Those with $(n/2) - m_1$ and $(n/2) - m_2$ form a 2×2 contingency table with all margins fixed, and (unless n is small) the null hypothesis may be tested by (9), which in this case has only one degree of freedom, and, in fact, may be put in the form

$$\chi^2 = \frac{16}{n}\left(m_1 - \frac{n}{4}\right)^2 \tag{12}$$

Confidence Intervals. To obtain a confidence interval for α, one first fits a line $x = \tilde{\beta} z$ to the data by the condition

$$\underset{x_i \leq \tilde{z}}{\text{Median}} (x_i - \tilde{\beta} z_i) = \underset{x_i > \tilde{z}}{\text{median}} (x_i - \beta z_i) \tag{13}$$

If the deviations of the x_i from this line are denoted by x_i', i.e., if

$$x_i' = x_i - \tilde{\beta} z_i \tag{14}$$

then the estimate of α is the median of the x_i', and a confidence interval for α is obtained by applying the method described for ν in Sec. 3 to the x_i'.

The simplest description of a confidence interval for β is to say that a $1 - p$ confidence interval is the set of points β_0 which would not be rejected at the p level of significance by the test described above. Thus one might determine the confidence interval by trial and error. An approximate method, which may be ordinarily expected to be quite satisfactory, is to fit the line $x = \tilde{\alpha} + \tilde{\beta} z$ and rotate it about the point where it intersects the line $z = \tilde{z}$. Since the number of points, m_1, above the line and to the left of $\tilde{z}$ is approximately normally distributed with mean $n/4$ and variance $n/16$, the limits of the confidence interval would be obtained by rotating the line until m_1 reached its $p/2$ and its $1 - (p/2)$ levels. The slopes of the line in these two positions approximate the $1 - p$ confidence limits of β.

16.9. General Linear Regression. The treatment of the more general case is a straightforward extension of the methods already described. Let there be k observable parameters $z_1, z_2, \cdots, z_k$ and

let the regression equation be of the form

$$y = \alpha_0 + \sum_1^k \alpha_r z_r \tag{1}$$

On the basis of n observations $(y_i, z_{1i}, z_{2i}, \cdots, z_{ki})$ with

$$i = 1, 2, \cdots, n$$

it is desired to estimate the α's or test hypotheses about the α's.

Suppose first that the regression does not involve α_0, so that we merely wish to estimate $\alpha_1, \alpha_2, \cdots, \alpha_k$. The k conditions on the observations which determine these estimates are

$$\underset{z_{ri} \leq \bar{z}_r}{\text{Median}} \left(y_i - \sum_1^k \tilde{\alpha}_r z_{ri}\right) = \underset{z_{ri} > \bar{z}_r}{\text{median}} \left(y_i - \sum_1^k \tilde{\alpha}_r z_{ri}\right) \tag{2}$$

there being k such conditions, one for each value of r. Thus the observations are divided into two groups by the median of each of the k z's, and the medians of the deviations in each group of any pair of groups are required to be equal. Now turning to the case in which a constant α_0 is involved, the condition for determining α_0 is

$$\tilde{\alpha}_0 = \text{median } (y_i - \Sigma\tilde{\alpha}_r z_{ri}) \tag{3}$$

or, what is the same thing,

$$\text{Median } (y_i - \tilde{\alpha}_0 - \Sigma\tilde{\alpha}_r z_{ri}) = 0 \tag{4}$$

If we consider any one of the relations (2), it is clear that the median on each side of the equation must in fact be the median of the whole set of deviations, hence must be $\tilde{\alpha}_0$. Thus to fit a regression function of the form (1), one may specify the conditions (2) and (4), or he may combine them into

$$\underset{z_{ri} \leq \bar{z}_r}{\text{Median}} (y_i - \tilde{\alpha}_0 - \Sigma\tilde{\alpha}_r z_{ri}) = \underset{z_{ri} > \bar{z}_r}{\text{median}} (y_i - \tilde{\alpha}_0 - \Sigma\tilde{\alpha}_r z_{ri}) = 0 \tag{5}$$

It is worth noting that the estimation of $\alpha_1, \alpha_2, \cdots, \alpha_r$ is entirely independent of α_0; one could make any assumption he wished about α_0 without influencing the estimates of the other α's.

To test hypotheses about the α's or estimate them by confidence intervals, one would use the procedures described in the preceding section. Thus to test $\alpha_1 = \alpha_{10}$, one would fit the other constants by means of (5) with $\tilde{\alpha}_1$ replaced by α_{10} and the relation for $r = 1$ would of course be omitted. One would then test, using a 2×2 table,

whether the signs of the deviations were split fifty-fifty below and above $\tilde{z}_1$. A confidence interval for α_1 is the set of values α_{10} which would not be rejected by the test. To test a set of the α's, for example, to test whether the first c of the α's had the values $\alpha_t = \alpha_{t0}$

$$(t = 1, 2, \cdots, c)$$

one would fit the other α's using the last $k - c$ relations of (5), then construct c 2×2 tables like that used in the preceding section to test β, adding the individual chi squares to get a criterion with c degrees of freedom. A confidence region for the $\alpha_1, \alpha_2, \cdots, \alpha_c$ may be constructed from those points in the $(\alpha_{10}, \alpha_{20}, \cdots, \alpha_{c0})$ space which would not be rejected by the test.

The actual fitting of a regression function requires an iterative computation. The constant term α_0 is estimated last by equation (3). A first approximation α_r' to $\tilde{\alpha}_r$ is obtained by finding the slope of the line joining the median of (y_i, z_{ri}) for $z_{ri} < \tilde{z}_r$ and the median of (y_i, z_{ri}) for $z_{ri} > \tilde{z}_r$. The α_r' are then used to compute deviations

$$y_i' = y_i - \sum_r \alpha_r' z_{ri}$$

which are again fitted to the z_{ri} in the same fashion. The slopes obtained are added to the α_r' to obtain second approximations, α_r''. The process continues until the desired accuracy is achieved; then α_0 is estimated as the median of the final set of deviations.

16.10. Tests of Association. Given a sample of n observations from a bivariate population, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, the problem is to test whether the two variates are independently distributed. We assume both variates are continuous so that the probability is zero that two observations have the same value.

Contingency Test. The simplest test that comes to mind for this problem is to test whether a regression line fitted to the points has zero slope. The test amounts merely to dividing the n points into four groups by the two lines $y = \tilde{y}$ and $x = \tilde{x}$. The numbers of points in the four quadrants form a 2×2 contingency table and have the contingency-table distribution under the null hypothesis. The chi-square criterion (with one degree of freedom, since all marginal totals are fixed) may therefore be used unless n is small.

Corner Test. The so-called corner test appears to be the best test yet developed for the problem at hand. There is no proof that it is best, but in the event x and y are not independently distributed, this

test appears most likely to reject the null hypothesis. It is as simple to use as the contingency test.

The test is performed as follows: First the observations are divided into four groups by the medians as in the contingency test (the solid lines of Fig. 74). Now we shall arrange to deal always with an even number of points. If $n = 2m$, the two median lines will not intersect any points. (In practice they may, because of coarse measurement. If, for example, the horizontal line intersects two points, one may choose one of them arbitrarily and move it slightly up or down according as a tossed coin falls heads or tails; the other would be moved in the opposite direction. A similar procedure would be used for four, six, · · · points on a median line.) If $n = 2m + 1$, the two median lines will each pass through a point. In case it is the same point, that point is omitted. In case the two points are different, they are both omitted and a new point constructed from the two coordinates of the omitted points which are not medians. The circled point in Fig. 74 was added to the original data in this manner. Thus, in any case, we shall deal with an even number of points, say $2m$.

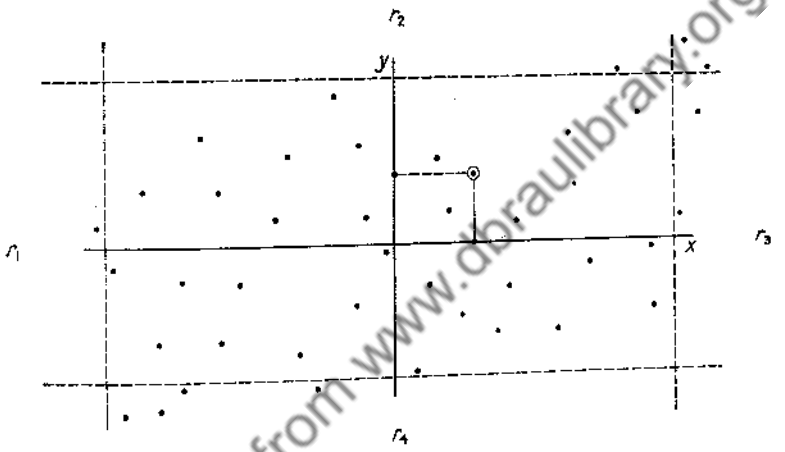


FIG. 74.

The dashed lines of Fig. 74 are now constructed. Starting at the left, a vertical line is moved to the right until one encounters a point on the opposite side of the horizontal line from the first point encountered. The upper horizontal dashed line is moved down from above until one encounters a point on the opposite side of $x = \bar{x}$ from the first point encountered. The other two dashed lines are located

similarly, the right one by moving to the left, the lower one by moving up. Let r_1 denote the number of points to the left of the left line; r_2 denote the number of points above the upper line, etc. Points in the upper right and lower left quadrants are counted positive while those in the other two quadrants are counted negative. Thus in Fig. 74, $r_1 = -1$, $r_2 = 3$, $r_3 = 4$, $r_4 = 4$.

The test criterion is

$$r = r_1 + r_2 + r_3 + r_4 \tag{1}$$

and it is intuitively clear that a large positive or negative value of r is evidence against independence while small values of r are expected if the null hypothesis is true. We must find the distribution of r under the null hypothesis in order to determine the critical level for a desired probability of a Type I error.

If x and y are independently distributed, then a random sample of n pairs (x, y) is nothing more than a sample of n x's and a sample of n y's paired at random. If the x's are ordered with x_1 the smallest, x_2 the second smallest, and so on, then the sample of n y's may be paired with the x's in $n!$ ways corresponding to the $n!$ permutations of the ordered y values, and under the null hypothesis all of the permutations are equally likely. Our distribution problem therefore is simply a matter of counting the number of permutations of the $2m$ y values which give a specified value of r; this number divided by $(2m)!$ is the probability of r.

Let us suppose for the moment that all four of r_1, r_2, r_3, r_4 are positive, and suppose all that the number of points in the upper right quadrant is j; then there will be $m - j$ points in the upper left and in the lower right quadrants and j points in the lower left quadrant. The numbers r_2 and r_3 depend only on the m x's greater than $\bar{x}$ and the m y's greater than $\bar{y}$. For r_2 to be positive, the j y values in the upper right quadrant must include the top r_2 y's but not the one just below them. The other $j - r_2$ y's in this quadrant must therefore be selected from the $m - r_2 - 1$ smallest of the m y's greater than $\bar{y}$; this selection can be made in $\binom{m - r_2 - 1}{j - r_2}$ ways. The j y's that have been selected must now be associated with j of the x's to right of $\bar{x}$ and among these must be the top r_3 x's, since r_3 is assumed positive, but not x_{2m-r_2}. The other $j - r_3$ values of x may therefore be selected in $\binom{m - r_3 - 1}{j - r_3}$ ways from the smallest $m - r_3 - 1$ values of x to the right of $\bar{x}$.

A similar argument shows that there are

$$\binom{m - r_4 - 1}{j - r_4}\binom{m - r_1 - 1}{j - r_1}$$

selections of y's and x's which give positive values of r_1 and r_4 in the lower left quadrant. After all these selections have been made, the remaining $m - j$ y values above $\tilde{y}$ are assigned to the remaining $m - j$ values of x to the left of $\tilde{x}$, and the remaining $m - j$ y values less than $\tilde{y}$ are assigned to the remaining x values to the right of $\tilde{x}$. The y values in any quadrant may be permuted at will. Thus there are $j!$ permutations of y values in the upper right quadrant, $(m - j)!$ permutations of y values in the upper left quadrant, and so on. The total number of y permutations which give j points in the upper right quadrant and the given values of the r's is therefore

$$\binom{m - r_1 - 1}{j - r_1}\binom{m - r_2 - 1}{j - r_2}\binom{m - r_3 - 1}{j - r_3}\binom{m - r_4 - 1}{j - r_4} j!j!(m - j)!(m - j)! \quad (2)$$

For any other assignment of signs to the r's, the argument is just the same, and the expression (2) would be changed only in that the lower index of the binomial coefficients would be different for negative r's. If we let s_a $(a = 1, 2, 3, 4)$ represent the numerical value of r_a, i.e., $s_a = r_a$ if r_a is positive and $s_a = -r_a$ if r_a is negative, then the binomial coefficient corresponding to r_a in (2) is

$$\binom{m - s_a - 1}{j - s_a}$$

if r_a is positive, and is

$$\binom{m - s_a - 1}{m - j - s_a}$$

if r_a is negative. We have given enough details now that it is fairly easy to show that the factorial-moment generating function for r is

$$\begin{aligned}\phi_n(t) &= E(t^r) = E(t^{r_1+r_2+r_3+r_4}) \\ &= \sum_{j=0}^{m} \frac{j!^2(m - j)!^2}{(2m)!} \\ &\qquad \left[\sum_{s=1}^{j}\binom{m - s - 1}{j - s} t^s + \sum_{s=1}^{m-j}\binom{m - s - 1}{m - j - s} t^{-s}\right]^4 \quad (3)\end{aligned}$$

and, of course, the probability for a given value of r is the coefficient of t^r in (3).

When n is small, (3) may be used to tabulate the distribution of r. The large-sample distribution of r is not normal. It can be shown, though we shall not do so here, that when n becomes infinite, the generating function (3) becomes simply

$$\phi(t) = \left(\sum_{s=1}^{\infty} 2^{-(s+1)}t^s + \sum_{s=1}^{\infty} 2^{-(s+1)}t^{-s}\right)^4 \tag{4}$$

The limiting distribution has been tabulated, and it is found that the 5 per cent limits on r are ± 11 and the 1 per cent limits are ± 14; i.e.,

$$P(-11 < r < 11) \cong .95 \tag{5}$$

so that if r equals or exceeds 11, the hypothesis of independence is rejected at the 5 per cent level of significance.

The small-sample distribution of r has been tabulated, and it is found that the limiting 5 and 1 per cent levels are quite satisfactory if the sample size is ten or more. Thus, though the distribution problem is rather troublesome, the application of the test is quite simple.

16.11. Power Functions. No generally accepted theory of power functions for distribution-free tests has yet been developed, and we shall therefore confine our discussion to a few brief remarks.

The great difficulty in obtaining a power function arises from the fact that the functional form of the distribution is not specified. Suppose, for example, that one wishes to test the null hypothesis that the median ν of a population has the value $\nu = 0$. What is the power of the test of Sec. 3 at $\nu = 1$? It is apparent that it depends entirely on the form of the distribution. If the distribution happens to be normal and $\sigma = 0.1$, the power will be very high at $\nu = 1$ even for small samples, but if $\sigma = 10$, the power will be quite low for small samples. It is thus apparent that a power function in the ordinary sense does not exist even for a specified family of distributions; the actual distribution is needed.

To circumvent this difficulty, it has been suggested that the power be computed as a function of $F(\nu)$ instead of as a function of ν; $F(x)$ is the cumulative distribution of the population. The null hypothesis mentioned above takes the form $F(0) = \frac{1}{2}$, and the alternatives are $0 \leq F(0) \leq 1$ excepting $F(0) = \frac{1}{2}$. Thus the null hypothesis states that $x = 0$ is the median of the population while the alternatives state

that $x = 0$ is some other percentage point of the population. This power function for the test of the median is then identical with the ordinary power function of the test that $p = \frac{1}{2}$ for a binomial population. It is clear that this device is applicable to many of the tests discussed in this chapter.

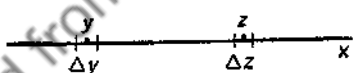
16.12. Notes and References. Many writers have contributed to the development of distribution-free techniques of analysis. A large share of the credit for these developments belongs to S. S. Wilks, who was among the first to realize the importance and potentialities of this field and who encouraged many of his students to work in it. A paper which gives a comprehensive survey of all the developments up to the date of its publication and a rather complete bibliography is S. S. Wilks, "Order statistics," *Bulletin of the American Mathematical Society*, Vol. 54 (1948), pp. 6–50.

16.13 Problems

1. Find the density function for $u = F(x_r)$, where x_r is the rth ordered observation of a sample of size n from a population with cumulative distribution $F(x)$.

2. Derive the density function given in equation (2.10) by integrating (2.4).

3. Derive (2.10) by a geometrical argument, considering the x axis divided into five intervals as illustrated. The sample is regarded as coming from a multinomial population with five categories having



probabilities $F(y - \Delta y/2)$, $f(y)\Delta y$, $F(z - \Delta z/2) - F(y + \Delta y/2$, $f(z)\Delta z$, $1 - F(z + \Delta z/2)$, and in such a way that $r - 1$ observations fall in the first category, one in the second, and so on. The density of x is $f(x)$ with cumulative $F(x)$.

4. Use the geometrical method of Prob. 3 to find the joint density function of u, the area between x_q and x_r, and v, the area between x_s and x_t, with $q < r < s < t$.

5. Show that the expected value of the larger of a sample of two observations from a normal population with zero mean and unit variance is $1/\sqrt{\pi}$, and hence that for the general normal population the expected value is $\mu + (\sigma/\sqrt{\pi})$.

6. If (x, y) is an observation from a bivariate normal population with zero means, unit variances, and correlation ρ, show that the expected value of the larger of x and y is $\sqrt{(1 - \rho)/\pi}$.

7. Derive equation (4.7).

8. Verify equations (4.9) and (4.10).

9. Show that t defined by equation (4.14) is approximately normally distributed for large n.

10. Verify equations (4.23) and (4.24).

11. Verify equation (4.31).

12. Derive the distribution given in (5.1).

13. Verify (6.9) and (6.10), and show that (6.11) and (6.12) do in fact define the inverse matrix.

14. Provide the details of the argument for normality which uses (6.13).

15. Verify (6.15).

16. Show that (7.5) is a generating function for the factorial moments of the n_i.

17. Verify equations (7.6) through (7.10).

18. Show that the distribution of r of Sec. 10 is symmetric about $r = 0$, hence that $E(r) = 0$.

19. Show that the limiting variance of r is 24.

20. Check the statement at the end of Sec. 10 by tabulating the cumulative distribution of the numerical value of r for $n = 10$. If s is the numerical value, it is found that $P(s \geq 10) = .0642$,

$$P(s \geq 11) = .0436$$

$P(s \geq 14) = .0127$, $P(s \geq 15) = .0095$. The corresponding values for n infinite are .0533, .0342, .0082, .0050.

21. Complete the derivation of (10.3).

22. If $x_1, x_2, \cdots, x_n$ is an ordered sample from a population with cumulative distribution $F(x)$, find the density for

$$u = \frac{[F(x_n) - F(x_2)]}{[F(x_n) - F(x_1)]}$$

23. The active life x, in hours, of radioactive atoms has the density $(1/\theta)e^{-x/\theta}$. To estimate θ for a particular kind of atom, a sample of n atoms is put under observation, but the experiment is to stop when the rth atom has expired; i.e., it is intended not to wait until all the atoms have ceased activity, but only until r of them (r chosen in advance) have. The data consist then of r measurements $x_1, x_2, \cdots, x_r$ and $n - r$ measurements known only to exceed x_r. Find the maximum-likelihood estimate of θ, and show that it has a chi-square distribution. Note that the likelihood contains the factor $[1 - F(x_r)]^{n-r}$ where $F(x)$ is the cumulative distribution.

24. Referring to Prob. 23, must one start with newly activated atoms, or is it all right to start with atoms that have already been active for various lengths of time (and are still active)?

25. If x is uniformly distributed between $\theta - \frac{1}{2}$ and $\theta + \frac{1}{2}$, find the density for the median $\tilde{x}$ for samples of size $2k + 1$.

26. Referring to Prob. 25, find the density for $z = (x_1 + x_{2k+1})/2$. Is z or $\tilde{x}$ the better estimator of θ?

27. Show that the sample median is a consistent estimator of the population median.

28. We have seen that the sample mean for a distribution with infinite variance (like the Cauchy distribution) does not necessarily converge in any sense toward the center of the distribution as the sample size increases. Does the sample median converge to the population median in such cases?

29. If a population has density function

$$\begin{aligned} f(x) &= \tfrac{1}{2}e^{-(x-\theta)} \qquad x \geq \theta \\ &= \tfrac{1}{2}e^{(x-\theta)} \qquad x \leq \theta \end{aligned}$$

find the maximum-likelihood estimate of θ for samples of size n.

30. A common measure of association for two variates x and y is the *rank correlation,* or *Spearman's correlation.* The x values are ranked and the observations replaced by their ranks; similarly the y observations are replaced by their ranks. Thus for samples of size n one might have:

x	1	2	3	$\cdots$	n
y	7	4	11	$\cdots$	6

Using these paired ranks, the ordinary correlation is computed

$$S = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2\Sigma(Y_i - \bar{Y})^2}} = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

where the capital letters represent the ranks, and $d_i = X_i - Y_i$. Verify that the given relation is true.

$$[\text{NOTE: } \sum_1^n i^2 = n(n+1)(2n+1)/6]$$

31. Show that the distribution of S of Prob. 30 is independent of the form of the distributions of x and y, provided that they are inde-

pendently distributed, hence that S is a distribution-free criterion for testing the null hypothesis of no association.

32. Show that the mean and variance of S under the hypothesis of independence are zero and $1/(n-1)$. To do this, show that S may be put in the form

$$S = \frac{12}{n^3 - n}\left[Q - \frac{n(n+1)^2}{4}\right]$$

where $Q = \Sigma i Y_i$ (replacing X_i by i), and observe that the coefficient of $\prod_1^n u_i$ in

$$\phi(t) = \frac{1}{n!}\prod_{j=1}^{n}\left(\sum_{i=1}^{n} u_i t^{ij}\right)$$

is a factorial-moment generating function for Q.

33. Apply some of the distribution-free methods to sets of data to be found in problems of Chaps. 13 and 14.

TABLES

DESCRIPTION OF TABLES

I. *Ordinates of the Normal Density Function.* This table gives values of

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for values of x between zero and four at intervals of 0.01. Of course one uses the fact that $f(-x) = f(x)$ for negative values of x.

II. *Cumulative Normal Distribution.* This tabulates

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

for values of x between zero and 3.5 at intervals of 0.01. For negative values of x, one uses the relation $F(-x) = 1 - F(x)$. Values of x corresponding to a few round values of F are given separately beneath the main table.

III. *Cumulative Chi-square Distribution.* This table gives values of u corresponding to a few selected values of $F(u)$ where

$$F(u) = \int_0^u \frac{x^{(n-2)/2} e^{-x/2}\, dx}{2^{n/2}[(n-2)/2]!}$$

for n, the number of degrees of freedom, equal to 1, 2, $\cdots$, 30. For larger values of n, a normal approximation is quite accurate. The quantity $\sqrt{2u} - \sqrt{2n-1}$ is nearly normally distributed with zero mean and unit variance. Thus u_α, the α point of the distribution, may be computed by

$$u_\alpha = \tfrac{1}{2}(x_\alpha + \sqrt{2n-1})^2$$

where x_α is the α point of the cumulative normal distribution. As an illustration, we may compute the .95 value of u for $n = 30$ degrees of freedom:

$$\begin{aligned} u_{.95} &= \tfrac{1}{2}(1.645 + \sqrt{59})^2 \\ &= 43.5 \end{aligned}$$

which is in error by less than 1 per cent.

IV. *Cumulative "Student's" Distribution.* This table gives values of t corresponding to a few selected values of

$$F(t) = \int_{-\infty}^{t} \frac{\left(\frac{n-1}{2}\right)!}{\left(\frac{n-2}{2}\right)!\sqrt{\pi n}\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}\,dx$$

with $n = 1, 2, \cdots, 30, 40, 60, 120, \infty$. Since the density is symmetric in t, it follows that $F(-t) = 1 - F(t)$. One should not interpolate linearly between degrees of freedom but on the reciprocal of the degrees of freedom, if good accuracy in the last digit is desired. As an illustration, we shall compute the .975 value for 40 degrees of freedom. The values for 30 and 60 are 2.042 and 2.000. Using the reciprocals of n, the interpolated value is

$$2.042 - \frac{1/30 - 1/40}{1/30 - 1/60}(2.042 - 2.000) = 2.021$$

which is the correct value. Interpolating linearly, one would have obtained 2.028.

V. *Cumulative F Distribution.* This table gives values of F corresponding to five values of

$$G(F) = \int_0^F \frac{\left(\frac{m+n-2}{2}\right)!\, m^{m/2}n^{n/2}x^{m/2}(n+mx)^{-(m+n)/2}\,dx}{\left(\frac{m-2}{2}\right)!\left(\frac{n-2}{2}\right)!}$$

for selected values of m and n; m is the number of degrees of freedom in the numerator of F, and n is the number of degrees of freedom in the denominator of F. The table also provides values corresponding to $G = .10, .05, .025, .01$, and .005 because $F_{1-\alpha}$ for m and n degrees of freedom is the reciprocal of F_α for n and m degrees of freedom. Thus for $G = .05$ with 3 and 6 degrees, one finds

$$F_{.05}(3, 6) = \frac{1}{F_{.95}(6, 3)} = \frac{1}{8.94} = .112$$

One should interpolate on the reciprocals of m and n as in Table IV for good accuracy.

TABLE I. ORDINATES OF THE NORMAL DENSITY FUNCTION

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Table II. Cumulative Normal Distribution

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

x	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$F(x)$	.90	.95	.975	.99	.995	.999	.9995	.99995	.999995
$2[1 - F(x)]$	.20	.10	.05	.02	.01	.002	.001	.0001	.00001

Table III. Cumulative Chi-square Distribution*

$$F(u) = \int_0^u \frac{x^{(n-2)/2} e^{-x/2}\, dx}{2^{n/2}(n-2)/2!}$$

n \ F	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	$.0^{4}393$	$.0^{3}157$	$.0^{3}982$	$.0^{2}393$	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

* This table is abridged from "Tables of percentage points of the incomplete beta function and of the chi-square distribution," *Biometrika*, Vol. 32 (1941). It is here published with the kind permission of the author, Catherine M. Thompson, and the editor of *Biometrika*.

Table IV. Cumulative "Student's" Distribution*

$$F(t) = \int_{-\infty}^{t} \frac{\left(\frac{n-1}{2}\right)!}{\left(\frac{n-2}{2}\right)!\sqrt{\pi n}\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}\,dx$$

n \ F	.75	.90	.95	.975	.99	.995	.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	1.282	1.645	1.960	2.326	2.576	3.291

* This table is abridged from the "Statistical Tables" of R. A. Fisher and Frank Yates published by Oliver & Boyd, Ltd., Edinburgh and London, 1938. It is here published with the kind permission of the authors and their publishers.

Table V. Cumulative F Distribution*

m degrees of freedom in numerator; n in denominator

$$G(F) = \int_0^F \frac{[(m+n-2)/2]!\, m^{m/2}\, n^{n/2}\, x^{m/2}\, (n+mx)^{-(m+n)/2}\, dx}{[(m-2)/2]!\, [(n-2)/2]!}$$

G	n \ m	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
.90		39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.3	62.8	63.1	63.3
.95		161	200	216	225	230	234	237	239	241	242	244	246	248	250	252	253	254
.975	1	648	800	864	900	922	937	948	957	963	969	977	985	993	1000	1010	1010	1020
.99		4,050	5,000	5,400	5,620	5,760	5,860	5,930	5,980	6,020	6,060	6,110	6,160	6,210	6,260	6,310	6,340	6,370
.995		16,200	20,000	21,600	22,500	23,100	23,400	23,700	23,900	24,100	24,200	24,400	24,600	24,800	25,000	25,200	25,400	25,500
.90		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	9.46	9.47	9.48	9.49
.95		18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
.975	2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5
.99		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
.995		199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
.90		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	5.17	5.15	5.14	5.13
.95		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	8.62	8.57	8.55	8.53
.975	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9	13.9
.99		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	26.5	26.3	26.2	26.1
.995		55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.4	43.1	42.8	42.5	42.1	42.0	41.8
.90		4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.93	3.92	3.90	3.87	3.84	3.82	3.79	3.78	3.76
.95		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	5.75	5.69	5.66	5.63
.975	4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.36	8.31	8.26
.99		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	13.8	13.7	13.6	13.5
.995		31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.7	20.4	20.2	19.9	19.6	19.5	19.3
.90		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	3.17	3.14	3.12	3.11
.95		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	4.50	4.43	4.40	4.37
.975	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.12	6.07	6.02
.99		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	9.38	9.20	9.11	9.02
.995		22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.4	13.1	12.9	12.7	12.4	12.3	12.1
.90		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	2.80	2.76	2.74	2.72
.95		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	3.81	3.74	3.70	3.67
.975	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.96	4.90	4.85
.99		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	7.23	7.06	6.97	6.88
.995		18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.2	10.0	9.81	9.59	9.36	9.12	9.00	8.88
.90		3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.51	2.49	2.47
.95		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	3.38	3.30	3.27	3.23
.975	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.25	4.20	4.14
.99		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	5.99	5.82	5.74	5.65
.995		16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.53	7.31	7.19	7.08
.90		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	2.38	2.34	2.31	2.29
.95		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	3.08	3.01	2.97	2.93
.975	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.78	3.73	3.67
.99		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	5.20	5.03	4.95	4.86
.995		14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.40	6.18	6.06	5.95

.90		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	2.25	2.21	2.18	2.16
.95		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	2.86	2.79	2.75	2.71
.975	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.45	3.39	3.33
.99		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	4.65	4.48	4.40	4.31
.995		13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.62	5.41	5.30	5.19
.90		3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.15	2.11	2.08	2.06
.95		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.62	2.58	2.54
.975	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.20	3.14	3.08
.99		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	4.25	4.08	4.00	3.91
.995		12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.07	4.86	4.75	4.64
.90		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	2.01	1.96	1.93	1.90
.95		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	2.47	2.38	2.34	2.30
.975	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.85	2.79	2.72
.99		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	3.70	3.54	3.45	3.36
.995		11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.33	4.12	4.01	3.90
.90		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	1.87	1.82	1.79	1.76
.95		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	2.25	2.16	2.11	2.07
.975	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.52	2.46	2.40
.99		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	3.21	3.05	2.96	2.87
.995		10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.69	3.48	3.37	3.26
.90		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	1.74	1.68	1.64	1.61
.95		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	2.04	1.95	1.90	1.84
.975	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.22	2.16	2.09
.99		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	2.78	2.61	2.52	2.42
.995		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.12	2.92	2.81	2.69
.90		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	1.61	1.54	1.50	1.46
.95		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	1.84	1.74	1.68	1.62
.975	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.07	1.94	1.87	1.79
.99		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	2.39	2.21	2.11	2.01
.995		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.63	2.42	2.30	2.18
.90		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	1.48	1.40	1.35	1.29
.95		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	1.65	1.53	1.47	1.39
.975	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.67	1.58	1.48
.99		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	2.03	1.84	1.73	1.60
.995		8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.19	1.96	1.83	1.69
.90		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	1.41	1.32	1.26	1.19
.95		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	1.55	1.43	1.35	1.25
.975	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.69	1.53	1.43	1.31
.99		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	1.86	1.66	1.53	1.38
.995		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.98	1.75	1.61	1.43
.90		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	1.34	1.24	1.17	1.00
.95		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	1.46	1.32	1.22	1.00
.975	∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.57	1.39	1.27	1.00
.99		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	1.70	1.47	1.32	1.00
.995		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.79	1.53	1.36	1.00

* This table is abridged from "Tables of percentage points of the inverted beta distribution," *Biometrika*, Vol. 33 (1943). It is here published with the kind permission of the authors, Maxine Merrington and Catherine M. Thompson, and the editor of *Biometrika*.

INDEX

G

H

I

J

L

M

N