WILEY PUBLICATIONS IN STATISTICS

Elements of Statistics

WILEY PUBLICATIONS IN STATISTICS

Walter A. Shewhart, Editor

Mathematical Statistics

HANSEN, HURWITZ, and MADOW • Sample Survey Methods and Theory, Volume II
DOOB • Stochastic Processes
RAO • Advanced Statistical Methods in Biometric Research
KEMPTHORNE • The Design and Analysis of Experiments
DWYER • Linear Computations
FISHER • Contributions to Mathematical Statistics
WALD • Statistical Decision Functions
FELLER • An Introduction to Probability Theory and Its Applications, Volume I
WALD • Sequential Analysis
HOEL • Introduction to Mathematical Statistics, Second Edition (in press)

Applied Statistics

BENNETT and FRANKLIN • Statistical Analysis in Chemistry and the Chemical Industry (in press)
FRYER • Elements of Statistics
COCHRAN • Sampling Techniques
WOLD and JURÉEN • Demand Analysis
HANSEN, HURWITZ, and MADOW • Sample Survey Methods and Theory, Volume I
CLARK • An Introduction to Statistics
TIPPETT • The Methods of Statistics, *Fourth Edition*
ROMIG • 50-100 Binomial Tables
GOULDEN • Methods of Statistical Analysis, *Second Edition*
HALD • Statistical Theory with Engineering Applications
HALD • Statistical Tables and Formulas
YOUDEN • Statistical Methods for Chemists
MUDGETT • Index Numbers
TIPPETT • Technological Applications of Statistics
DEMING • Some Theory of Sampling
COCHRAN and COX • Experimental Designs
RICE • Control Charts
DODGE and ROMIG • Sampling Inspection Tables

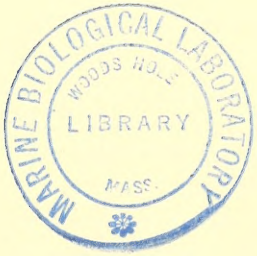
Related Books of Interest to Statisticians

ALLEN and ELY • International Trade Statistics
HAUSER and LEONARD • Government Statistics for Business Use

Elements of Statistics

H. C. FRYER

Professor of Mathematics
Statistician, Agricultural Experiment Station
Kansas State College



New York · John Wiley & Sons, Inc.
London · Chapman & Hall, Limited

Library of Congress Catalog Card Number: 54-5517

PRINTED IN THE UNITED STATES OF AMERICA

Preface

THIS BOOK HAS BEEN DEVELOPED TO PROVIDE THE BASIS OF AN introductory course in probability and statistics for the college and university student. It includes material which has been tried out in many classes and by several instructors for almost a decade.

The instructor using the book as a text or the student interested in the subject will find that college algebra is a necessary and sufficient prerequisite for this course, which aims to teach modern but elementary ideas, methods of reasoning, and methods of analysis fundamental but not peculiar to any particular specialized field. Once the student has acquired a background of elementary methods, probability, and frequency distributions, he can be taught some of the simpler sampling statistics in common use today. Thus, he may learn their importance as well as their application. The serious student will find included in this book problems to provoke thought and provide practice in statistical methods and reasoning.

Some colleges and universities offer statistics courses—often with graduate credit—in which the elementary concepts and methods are not assumed to be known and hence are taught during the first part of the course. It seems to me that one general course in probability and statistics, with emphasis on statistical reasoning and modern methods, helps to avoid useless duplication of instruction. It also leaves time in subsequent courses to do more advanced work in specialized fields. Such an introductory course also is rapidly becoming a necessary part of a student's education even if he does not use statistics directly in his specialized field.

It is helpful to the students during the studies of sampling to provide them with some mathematical models of populations so that they can obtain sampling experiences which—for a whole class—empirically verify for them the sampling distributions given in some of the tables which they will be using. It has been my experience that most students need this sort of empirical evidence before they really understand the nature and the use of sampling distributions. Numbers, and other symbols, written on plastic discs can be made to correspond closely to normal, non-normal, and binomial populations which are

met in actual practice. These populations, if properly employed, will enable the student to understand the more common sampling distributions rather well despite a lack of familiarity with mathematical statistics. Some of the problems in this book assume that such populations are available to the students.

It is desirable to have calculating machines available so that the students can learn what operations can be performed on them and can solve some of their problems more efficiently. However, I do feel that the acquisition of routine computational skills is not worthy of much college credit; hence, whenever there are heavy computations in a problem in this book, the necessary computations are usually given with the problem. For example, $\Sigma(X)$ and $\Sigma(X^2)$ are given for most problems with even a moderate amount of computation and asking for the mean and the standard deviation.

It has been my experience that it takes most of the equivalent of a three-semester-hour course to equip the student with the ideas and methods he needs before he can solve even the most elementary sampling problems in any particular field. For that reason we offer at Kansas State College a two-hour course in which the rest of the work on sampling contained herein can be given, and some applications to the students' fields of interest can be considered.

It is a pleasure to acknowledge the assistance given me by my colleague J. I. Northam, who pointed out errors in previous lithoprinted versions and made suggestions regarding the way the material should be presented. This book also has derived considerable benefit from the reviews made available to me during the last revisions of the manuscript. Obviously, the responsibility for all remaining shortcomings of the book is solely mine.

H. C. Fryer

Kansas State College
Manhattan, Kansas
December, 1953

Contents

CHAPTER 1

History and Introduction

1.1 HISTORY

The word "statistics," the associated mathematical analyses, and the general process of statistical reasoning appear to have begun their evolution around the time of Aristotle. This evolution can be described in terms of the following four phases, some of which occurred simultaneously among different groups of persons:

(1.11) An early, highly philosophical, study of "matters of state" which did little more for the statistical science used today than help to suggest its name.

(1.12) A semi-numerical and strongly sociological stage typified by the mathematical and philosophical study of large groups of numerical measurements bearing on health, insurance, foreign and domestic trade, and political matters.

(1.13) The development of the mathematical theory of probability starting in the sixteenth century with mathematical attacks on the various problems associated with games of chance.

(1.14) The current phase, starting late in the nineteenth century, during which phases (1.12) and (1.13) were combined, improved, and extended to produce a branch of mathematics which can handle a wide variety of problems pertaining to the drawing of valid and useful inferences from relatively small groups of numerical measurements.

During Aristotle's time interest developed in comparative descriptions of states. Aristotle is reported * to have written at least one hundred and fifty-eight descriptions of states, covering their histories, public administrations, arts, sciences, and religious practices. It was customary to refer to such compositions as treatises on "matters of state." That apparently is an important part of the origin of the

* Harald L. Westergaard (1942), *Contributions to the History of Statistics*, King.

term "statistics," although the name itself was coined many years after Aristotle's death.

For quite a long time after Aristotle a weak interest was maintained in descriptions of states partly by the intellectuals who enjoyed that pastime and partly by the rulers of the various states through their natural desire to know how many subjects they ruled, and to ascertain the wealth within their realms. Hence it is probable that some sort of crude census taking was attempted.

During the seventeenth and eighteenth centuries sufficient interest was generated in the study of the political, sociological, and economic features of states that societies developed for that purpose. In Germany this line of intellectual effort caused the development of the Staatenkunde, a name which appears to have led rather directly to the actual coining of the term "statistics." However, the Germans remained content to pursue the philosophical aspects of "matters of state"; hence the Staatenkunde never did become either very mathematical in character, or very useful. It merely typifies the last stages of the purely philosophical phase of the development of the science of statistical analysis, and points out its socio-political ancestry.

Another, and more fruitful, step in the evolution of the present-day type of statistical reasoning originated in England under the leadership of John Graunt. This was a semi-mathematical study of vital statistics, insurance, and economic statistics which came to be known as "Political Arithmetic." Epidemic diseases periodically decimated the populations of European nations; problems of agricultural production, foreign trade, and public administration became too complex to be handled without some form of numerical measurement and an objective means of interpretation of such measurements. Hence there was a natural interest in numbers of births and deaths, in estimates of the populations in various areas, in figures on agricultural production and foreign trade, and in methods for administering insurance against the economic situations created by death and disability.

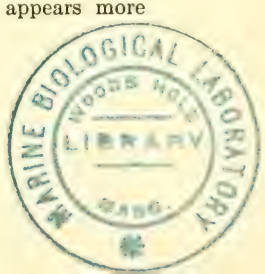
Public interest in specific measurements of populations and of resources was heightened by the constant danger of war with a neighboring state, and by the advent of an industrial revolution during the eighteenth century. It was the objective of the political arithmeticians to help with the collection and interpretation of data pertinent to the economic, sociological, and political problems which were becoming increasingly important and numerous. They devised methods for estimating the numbers of persons residing in certain political

units, and methods for summarizing groups of data. Their efforts to apply mathematical analysis to such problems helped to lay the foundation for the statistical methods now in use.

A third step in the evolution of statistical analysis and reasoning came in the development of the mathematical theory of probability, without which statistical reasoning could never have attained its present reliability and usefulness. Games of chance were especially popular among the well-to-do of the sixteenth and seventeenth centuries; and many problems involving probability were presented to the mathematicians of the day for solution. For example, an Italian nobleman asked Galileo to explain the following facts: If three dice are thrown, the numbers 9 and 10 can each be obtained from six different combinations of the numbers on the faces of the dice; but it has been found from experience that a sum of 10 appears more frequently than a sum of 9. Why so? By an enumeration of all the physically different ways that three dice can produce sums of 9 or 10, Galileo was able to answer this question clearly and convincingly. His answer appears to be the first published application of the theory of probability.* Other prominent mathematicians such as Pascal, Fermat, James and Daniel Bernoulli, de Moivre, Laplace, Gauss, Simpson, Lagrange, Hermite, and Legendre developed many important theorems and methods of attacking problems involving chance events, and they passed this information on for later use by mathematical statisticians.

During the last quarter of the nineteenth century, Sir Francis Galton took the lead in the development of the ideas of regression and correlation when two (or more) measurements are made simultaneously on each member of a group of objects. He appears to have built his ideas around problems in genetics. Karl Pearson and C. Spearman extended this theory and applied it to studies in the social sciences, especially psychology. Karl Pearson and others also had

* The nature of Galileo's solution is as follows. A sum of 9 can be obtained from any of the following combinations of numbers on three dice: 1, 2, 6; 1, 3, 5; 1, 4, 4; 2, 2, 5; 2, 3, 4; or 3, 3, 3. A sum of 10 is obtained from any of these: 1, 3, 6; 1, 4, 5; 2, 2, 6; 2, 3, 5; 2, 4, 4; or 3, 3, 4. There are six different combinations giving each of the sums 9 and 10; *but* the different combinations do not occur equally frequently. For example, the combination 3, 3, 3 can be thrown but one way whereas the combination 3, 3, 4 can occur on any of three different throws, and hence would tend to appear three times as often as 3, 3, 3. As a matter of fact, a 9 can be thrown in twenty-five different ways, a 10 in twenty-seven different ways, which is the reason that the 10 appears more frequently in games than the 9.



begun to study the effects of sampling errors on conclusions drawn from samples.

By the end of the nineteenth century the Staatenkunde had ceased to exist, and "Political Arithmetic" had died in name but had developed into a science of statistical analysis, with emphasis on sociological and economic applications. The theory of mathematical probability had grown extensively as a branch of pure mathematics, and also was beginning to be associated with applied statistics. Thus the groundwork was laid for the present phase in the evolution of statistical theory and methods.

In 1908 William Seely Gosset, who wrote under the pseudonym "Student," published an article in the journal *Biometrika* which was later to typify the opening of a new era in the statistical analysis and interpretation of sampling data. From 1899 until his death in 1937, Gosset worked for the brewing firm, Messrs. Guiness. His associations with this firm led him into a variety of experiences and suggested uses for statistical methods which are typical of several of the present-day applications of statistics.

Messrs. Guiness were interested in barley, not just any barley, but in those varieties, growing conditions, and practices which would produce the best barley for breweries to use. These circumstances brought Gosset into contact with agricultural experimentation aimed at the improvement of crops and of agricultural practices. Moreover, Messrs. Guiness did not wish to subsidize the raising of large crops purely for the sake of scientific experimentation; they were a commercial firm which wanted to show a profit from their enterprises. That fact, plus the shortage of tillable land in Ireland and England, made Gosset well aware of the importance of small samples and of methods for deriving reliable information from such samples.

Furthermore, a large brewery conducts many chemical analyses, and hence needs to take proper account of errors of measurement. And, finally, the firm with which Student was associated was confronted with problems concerning industrial statistics: production and marketing analyses, price analyses, and methods for controlling the quality of the products which it manufactures for market. Thus Student came into contact with a wide variety of agricultural, economic, and industrial problems which would require some form of statistical study. Moreover, those problems had to be solved for a commercial firm, a situation which demanded efficiency and reliability with a minimum cost consistent with these qualities. The twentieth century renaissance of statistical theory and methods appears to be

based on that attitude toward the purposes of statistical analysis.

Unfortunately, the statistical ideas and procedures which Student introduced in 1908 did not become familiar to persons outside his own firm for nearly a decade, at which time R. A. Fisher and his colleagues in England began to extend and to popularize the theory of small samples and its applications. The theory of statistics was developed extensively by Jerzy Neyman and Karl Pearson's son, Egon S. Pearson. They placed special emphasis on rigor in statistical reasoning and led the way by publishing many papers in this field. Many others have followed their lead since their papers began to appear. The results of this research are being applied in many fields, such as biology, the physical sciences, industry, economics, sociology, medicine, education, and psychology.

1.2 SOME OF THE PURPOSES OF STATISTICAL REASONING

Early in his history man displayed a desire to take numerical measurements of the various phenomena involving himself and his environment. At first, those measures probably consisted of simple counts, or of crude measures of weight, volume, length, and area. At present many instruments are available for the precise measurement of those features of man's self and environment which interest him. He constantly is taking groups of numerical measurements because such a procedure can furnish a relatively precise and standard means of obtaining the information desired, of using it efficiently, and of transmitting that information to others. The general purpose of statistical analysis is to assist in the collection and the interpretation of sets of numerical measurements which supposedly have been taken for some useful purpose.

Once it is decided that a particular phenomenon should be measured numerically, one of two general classes of data is then obtained. It may be that it was both possible and practicable to secure every measurement of that particular kind which exists or could be obtained under the particular circumstances. Such a complete record is one type of statistical *population* of numerical measurements. An example is the record of the ages of all the legal residents of the state of Kansas on April 1, 1950, as contained in the official United States Census for that date. Another example is a list of the I.Q.'s of all the students entering a particular university in a given year.

However, it is more commonly true that it is impossible, or unwise, to collect a whole population of numerical measurements. In that event we obtain but a portion of a population for actual analysis, and attempt to draw from it useful conclusions about the population which was merely sampled. If the sample is to be useful it must be adequately representative of the population; that is, it should faithfully reflect the important features of the population.

In the event that the whole population of data is available for analysis, the purpose of statistical analysis is to reduce what is a relatively large bulk of numbers to a comprehensible form by means of graphs and tables and/or by calculating a few figures which contain most of the important information theoretically available in the original mass of data. For example, the ACE scores at the beginning of Chapter 2 are numerical measurements which the college took in the belief that they would be of value to the student and to the school, perhaps by helping to determine what profession the student should prepare to enter. Obviously those data are so bulky that they demand some sort of condensation.

It is worth noting at this point that even though the necessity to analyze whole populations of data is a rare circumstance, it is not logical to study the statistical analysis of samples without some adequate knowledge of the statistical features of the populations from which the samples are taken. Fortunately a considerable amount of useful statistical analysis can be learned and appreciated without studying more than two general types of populations.

Whenever we attempt to base conclusions concerning a statistical population of numerical measurements upon relatively few observations (a sample) from that population, we face two important general questions. (*a*) How shall the sample be taken so as to maximize its chance of being representative of that population? (*b*) Having obtained some numerical observations from the population with question *a* in mind, how do we draw valid conclusions from the sample? As an illustration, consider the following sampling problem. Suppose that a highway commission is considering the purchase of some cement for highway construction, and that two companies are offering their products for purchase. The commission wishes to compare the seven-day tensile strengths of the two cements before letting the contract. Obviously they must resort to sampling because they can test only a tiny portion of each company's total output of a particular sort of cement. It will be supposed, for purposes of illustration, that it has been decided that ten of the stand-

ard laboratory specimens will be tested from each company's product. The test of each specimen will produce an "observation" from the population of all such tensile strengths possible from that company's cement. All told, there will be twenty samples taken, ten from each company.

Would it be satisfactory to inform each company of the plans for testing the cements and ask each company to provide ten specimens of concrete for testing? Or, would it be better to go into the open market and purchase a sack of each company's cement from each of ten stores and have a laboratory uniformly make up the ten testing specimens? Rather clearly the latter method would be much more likely to produce specimens which were representative of the respective strengths of those concretes at seven days of age.

One of the purposes of statistical theory is to devise methods for taking samples in such a way that they do yield essentially the same information as is contained in the population which was sampled. For the most part, that phase of statistics lies beyond the scope of this book; hence no attempt will be made to do more than to remind the reader of a few commonsense considerations from time to time.

Suppose now that the ten sample specimens of concrete have been tested for tensile strength at seven days of age, with the following results:

Cement	Seven-Day Tensile Strength (lb./sq. in.) of Concrete
No. 1	425, 410, 425, 460, 430, 445, 445, 415, 450, and 440
No. 2	420, 450, 405, 400, 400, 415, 435, 425, 400, and 430

How do we decide from such evidence whether one concrete will, as a rule, excel the other in tensile strength; or if either or both conform to pre-assigned standards for such building materials? Casual observation indicates that cement No. 1 tends to produce greater tensile strength in its concrete than No. 2; but there are several specimens from cement No. 2 that produced greater strength than certain of the specimens from No. 1 cement. For example, five of the No. 1 specimens had tensile strengths at or below 430 pounds per square inch, and two of the specimens of cement No. 2 had strengths greater than 430 pounds per square inch. Without doubt, then, some batches of cement No. 2 are better than *some* batches of cement No. 1. Such a situation is met quite frequently in sampling studies; only rarely do progressive improvements in methods or materials come on such a large scale that all previous methods or materials are excelled without exception. What is needed—and now available to a highly use-

ful degree—is a method of reasoning which enables us to induce from relatively few samples useful information regarding the population sampled.

Inductive reasoning based on evidence obtained from samples necessarily runs some risk of error; but as long as the extent of this risk can be measured, the process offers real hope for useful application. Much of the recent research in mathematical statistics has been devoted to the development of methods of reasoning based on sampling observations.

The reader should not feel from the preceding remarks that sampling is useful only in scientific research, because everyone is constantly being confronted with sampling studies of one sort or another. Radio advertising is quite full of alleged sampling investigations during which various products presumably have been tested and shown to be superior. Life insurance premiums are based on samples of mortality rates among insurable persons. Public opinion polls, economic polls, and the like, often reported in the newspapers, are attempts to reason from a sample to conclusions about a whole population of possible responses to one or more questions. Persons who have visited other parts of the world return and, upon the basis of relatively small samples, attempt to say how whole nations or societies are reacting to certain world events. The reader undoubtedly can think of many other examples of sampling followed by more or less valid applications of either inductive or deductive reasoning, or what might be better described in this instance as statistical inference.

In closing these introductory remarks, it seems fair to warn the student that, as in many other lines of thought, he cannot immediately jump into interesting applications of statistical methods and reasoning. He must first learn some fundamental principles and some statistical tools with which to work, a process which necessarily occupies most of the time in a first course in statistics. There is, however, nothing to prevent him from reading the rest of the book for himself, and from taking other courses in statistics.

CHAPTER 2

The Summarization of Sets of Data Involving One Type of Measurement

Whenever a statistical investigation is to be made, two initial steps must be taken: (*a*) A group of objects (persons, plants, bolts, or anything capable of being measured) is specified as the subject to be studied. (*b*) A decision is made regarding the feature of these objects that is to be measured numerically or by some qualitative designation. Such a set of measurements is called a *population* if it includes every member of the group to be defined in *a*. For example, suppose that an economist proposes to study the net cash incomes of beef-cattle ranchers in Kansas during the decade from January 1, 1944, to January 1, 1954. It would be necessary first to define the group of ranchers to be included in this study. How many beef animals must he raise? Must the raising of beef cattle be his major source of income according to some standard? Are absentee owners included? There are many other matters which would have to be considered. When a specific group of Kansas ranchers has been defined, part *a* above has been completed.

Next it is necessary to decide upon the specific meaning of the term, *net cash income*. Is the measurement to be on a per-animal basis, or the total for the ranch regardless of its size? Is any adjustment to be made for inflation, cost-of-living indexes, and the like? When net cash income has been defined specifically, part *b* listed above has been completed, and the population is defined.

In some situations it is feasible to obtain every possible one of the measurements in a population, as would be the case if every beef-cattle rancher in the group discussed above were to be interviewed and his net cash income determined according to the definition adopted by the investigator. Under these circumstances, the purpose of statistical analysis is to summarize the information in the

data as clearly and as concisely as is possible. A statistical description of a population will be found to be important also when the population is to be studied by means of samples rather than in its entirety. There are various widely used methods of accomplishing such a purpose. The choice of a method depends upon what is to be learned from the data, and upon the statistical characteristics of the population which is being summarized.

The need for statistical descriptions and summaries is pointed out rather specifically by means of the data in Table 2.01. It contains the 1290 ACE scores made by students entering Kansas State College for the first time in 1947. An ACE score is intended to measure certain features of a student's intellect and aptitudes which are thought to be related to his success in pursuing one of the various possible college curricula. If so, ACE scores should help the student and the staff to do a better job of fitting the students' abilities and interests to the facilities which the college has to offer.

The reader is already generally familiar with the term *average* as some kind of usefully typical number which partially replaces a whole group of numbers; but he may be less familiar with the fact that there are several averages in use. It should be intuitively obvious that no single number, like an average, can be expected to summarize adequately the set of data in Table 2.01. Some measure of the variability exhibited by these ACE scores is needed; that is, an adequate description of the way these scores are dispersed, or distributed, between the lowest and highest scores is needed in addition to a description of the general level of performance. More specifically, we need a standard method of describing any particular student's score relative to the whole group of scores. With such information at hand, a trained adviser may be able to give a student considerable assistance in the choice of a vocation or a profession, or in the solution of personal problems.

The statistical procedures described and illustrated in this chapter will make it possible to replace the 1290 ACE scores by relatively few statistical constants, graphs, and tables which still contain all the really pertinent information embodied in the original population of numerical measurements. Some of these possible procedures will be introduced by means of small sets of data for the sake of convenience. Thereafter, reference again will be made to Table 2.01.

TABLE 2.01

1290 ACE Test Scores Made by Students Entering Kansas State College during 1947

(Data furnished by the Counseling Bureau of Kansas State College.)

70	108	84	104	103	92	103	128	96	82	145	139	128	136	94	67	78
49	84	121	86	108	65	47	75	95	67	53	95	104	95	89	79	82
119	106	52	134	131	108	145	110	54	68	123	112	82	123	141	93	75
108	95	82	59	120	105	79	115	118	112	109	104	70	112	125	126	37
137	103	80	130	80	93	97	116	48	98	93	102	36	114	119	130	149
95	138	85	104	79	63	63	92	60	92	76	118	125	102	68	117	130
84	118	122	104	123	127	72	99	44	112	78	72	66	94	118	79	129
134	89	90	80	146	105	83	88	134	106	127	77	43	90	95	87	94
57	46	68	124	134	120	136	111	92	128	88	110	65	157	113	145	118
116	72	120	62	52	90	91	115	128	61	58	106	100	106	99	82	121
71	103	75	112	107	88	142	112	71	135	124	99	98	95	119	93	101
99	118	150	100	91	109	83	87	93	120	92	77	80	101	140	112	60
116	67	97	114	60	111	102	107	117	77	129	80	56	105	88	113	93
73	127	111	107	76	55	102	90	78	105	87	113	138	108	68	134	85
97	76	111	85	98	97	78	79	110	143	62	84	110	94	135	156	71
112	70	123	69	58	79	123	121	117	84	64	72	97	84	115	133	90
128	120	120	148	68	46	116	103	124	61	43	98	99	111	110	66	40
75	100	63	72	71	96	72	87	52	102	73	65	80	71	91	100	64
52	78	130	85	87	66	103	94	111	159	88	85	115	108	111	41	120
111	129	89	183	78	99	108	55	91	80	103	86	58	99	132	109	118
92	105	109	95	88	153	70	76	109	121	102	119	82	103	45	50	69
59	44	100	93	71	99	45	78	99	124	99	108	72	87	122	98	96
83	52	94	121	63	108	94	102	103	69	114	95	41	28	65	107	66
76	99	33	115	121	146	83	89	124	118	98	131	95	100	97	89	74
102	103	144	76	91	53	103	92	158	65	99	107	114	90	104	54	136
86	76	114	118	114	84	112	98	139	61	23	127	115	94	68	108	93
89	75	90	113	97	64	100	101	109	100	91	37	128	118	64	72	92
76	113	166	118	125	82	104	87	113	95	112	105	73	101	121	70	69
98	72	140	115	106	114	83	113	95	91	75	128	50	68	92	127	92
91	57	120	106	117	84	74	121	134	92	91	102	78	81	106	119	106
82	114	146	110	46	112	136	41	87	23	120	119	131	68	99	123	87
113	69	102	76	127	88	99	96	65	113	103	77	104	123	105	79	107
108	61	93	81	143	93	112	139	109	123	93	134	110	55	115	75	147
136	139	125	103	108	47	107	115	56	101	80	56	95	82	66	112	76
54	62	49	51	99	131	101	134	94	68	68	53	120	124	73	120	70
129	78	151	55	134	69	80	109	125	49	36	134	97	118	76	78	137
61	73	72	111	95	162	97	128	91	58	93	54	90	93	80	94	101
96	69	99	43	111	108	133	101	121	66	77	71	105	156	85	89	150
104	97	94	128	100	138	105	49	85	116	98	89	123	153	129	93	78
98	82	84	65	93	54	109	57	96	139	65	126	129	110	86	72	100
61	105	131	101	108	96	138	87	69	112	154	116	88	134	79	92	41
93	121	99	74	117	100	93	127	99	108	95	95	76	120	130	111	111
131	104	145	93	97	96	109	98	99	60	83	60	94	69	97	120	88
108	74	110	62	68	77	86	94	93	94	70	119	120	120	120	112	38
49	128	91	84	101	65	87	72	128	88	126	122	53	29	100	111	117
127	107	83	138	80	88	54	104	105	72	115	119	113	80	90	64	128
89	109	61	135	127	100	96	81	76	148	69	55	136	98	77	44	105
125	118	109	92	119	94	108	33	101	108	122	88	102	56	92	117	114
117	130	65	99	60	88	102	130	128	109	105	115	64	120	127	87	81
110	81	120	115	99	73	105	122	109	95	85	83	99	60	48	123	127
43	104	96	103	84	131	104	115	69	108	113	37	63	106	78	52	57
102	121	135	72	158	89	106	94	105	99	86	75	94	114	109	91	114
151	138	135	64	80	80	91	99	107	100	94	110	91	100	47	113	94
89	106	64	84	99	90	67	145	74	87	92	67	81	113	116	89	99
69	98	109	84	77	106	123	74	107	92	98	87	80	74	61	125	101
72	77	92	50	108	91	118	159	145	94	110	95	107	81	112	73	82
85	116	98	44	105	140	109	41	105	43	119	148	112	83	73	81	117
135	63	91	102	117	74	66	90	124	64	117	58	126	100	108	65	88
24	88	80	34	107	101	44	78	97	133	77	137	90	114	72	110	163
56	102	55	61	103	80	50	71	110	88	44	123	99	160	65	87	113
71	92	107	124	134	96	126	98	62	89	54	115	131	81	101	91	106
44	102	90	110	106	84	70	70	78	104	73	69	125	109	81	76	103
86	106	67	109	103	120	117	91	76	122	69	105	154	68	107	127	58
125	122	55	102	65	83	149	58	45	119	119	129	151	78	81	50	78
103	117	70	79	111	124	57	93	134	120	117	75	90	77	130	84	102
97	76	135	139	49	97	136	84	89	104	52	71	87	43	96	131	80
86	107	139	127	61	113	71	109	67	126	96	113	82	55	84	71	62
70	97	107	115	95	64	111	88	118	118	95	78	178	88	88	109	88
98	32	39	71	68	60	95	139	123	91	99	61	104	78	178	98	85
95	89	97	110	80	94	87	64	100	42	105	99	39	124	109	67	108
85	101	130	39	114	127	115	62	149	137	64	113	97	93	104	113	154
93	47	79	44	84	102	124	132	131	43	102	140	42	75	123	63	122
73	99	107	111	98	80	92	120	60	100	74	99	78	87	127	61	142
131	106	116	98	86	77	130	67	107	106	82	120	72	70	63	47	82
108	78	112	101	113	107	50	106	106	117	100	53	107	79	115	96	78
70	100	44	80	71	82	114	107	80	62	88	92	89	88	129		

2.1 THE ARITHMETIC MEAN AND THE STANDARD DEVIATION

When it is necessary to analyze a population of statistical measurements it often is desirable to calculate a single number which will be typical of the general level of magnitude of the measurements in the population. Logically, the first question is: What features should averages have in order to be typical of the data in some useful sense? Therefore, the following properties of averages are suggested as being either required of averages, or desirable:

(2.11) An average should be close, on the scale of measurement, to the point (or interval) of greatest concentration of the measurements in the population.

(2.12) It should be as centrally located among the numbers as is compatible with property (2.11).

(2.13) An average should be simple to compute if that is achievable under the circumstances.

(2.14) It should be tractable to mathematical operations so that useful theoretical information can be derived by means of mathematical methods.

(2.15) The average should be such that measures of the scatter of the data about the average can be obtained and also have properties (2.13) and (2.14).

A simple but crude average which sometimes is quite useful is the *midrange,* MR. It is defined as that number (not necessarily one of those being studied) which is halfway between the extreme numbers in the set being summarized. For example, the extreme ACE scores in Table 2.01 are 23 and 183. The difference is 160; hence

$$\text{MR} = 23 + 160/2 = 103, \text{ also } = \frac{23 + 183}{2},$$

because this is the number which is halfway between 23 and 183. Among the desirable properties of averages listed above, the midrange is centrally located (in the sense that it is midway between the extremes), and it often is in the region of the greatest concentration of the data. It also is easily calculated, but it does not possess the other properties listed. In addition, the midrange does not appear to be a very reliable average because its size depends on only two of the numbers.

An average called the *arithmetic mean* has been found to possess all the properties (2.11) to (2.15) to a rather high degree for a broad class of statistical populations. In addition, it is extremely useful in the analysis of sampling data, as will be shown later. Hence the arithmetic mean is a highly recommended average.

The arithmetic mean, μ (Greek letter mu), of N measurements: $X_1, \ldots, X_N$ is calculated by dividing the sum of the N measurements by N. Symbolically,

$$(2.11) \qquad \mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} (X_i)}{N}, \quad \text{or for brevity,} \quad \mu = \frac{\Sigma X}{N}.$$

To illustrate, suppose that $X_1 = 2$, $X_2 = 5$, $X_3 = 1$, $X_4 = 3$, and $X_5 = 4$; then the arithmetic mean is

$$\mu = (2 + 5 + 1 + 3 + 4)/5 = 3.*$$

Problem 2.11. Suppose that eight players are on the traveling squad of a basketball team, and their weights are 152, 170, 165, 185, 201, 174, 191, and 210 pounds. What is the arithmetic mean of these weights?

The first question which may occur to the student is: What are the X_i in this instance? It is a well-known assumption in arithmetic and algebra that the same sum is obtained for a given set of numbers no matter what the order of addition; that is, $3 + 6 + 15 = 24 = 6 + 15 + 3 = 15 + 3 + 6$, or any other possible order of addition. Likewise, in the present problem, it makes no difference which weight is symbolized as X_1, which as X_2, etc. It is convenient just to let X_1 = the first weight listed, X_2 = the second weight on the list, and so on. If that is done in this problem, $X_1 = 152$, $X_2 = 170$, $X_3 = 165$,

* Although the discussion in this chapter is chiefly devoted to methods and ideas appropriate to populations of data—which usually contain a large number of measurements—small groups of numbers will be used in examples and problems for the purpose of facilitating and shortening discussions. Obviously, most of these problems and examples resemble samples far more than populations. However, the methods introduced will apply to populations and will not necessarily be correct or efficient for sampling studies, as will be noted later in the book.

$X_4 = 185$, $X_5 = 201$, $X_6 = 174$, $X_7 = 191$, and $X_8 = 210$, all in pounds. Therefore,

$$\sum_{1}^{8} X_i = X_1 + X_2 + \cdots + X_8 = 152 + 170 + \cdots + 210 = 1448$$

so that $\mu = 1448/8 = 181$ pounds, which is the arithmetic mean of the weights.

Although the number 181 pounds gives a useful impression of the general weight size of the players of problem 2.11, it is obvious that the same mean weight could have been obtained for many other groups of eight weights, some of which might be considered to be quite different from those above. For example, each of the following sets of eight weights (in pounds) has $\mu = 181$:

Set 1. 185, 180, 181, 184, 182, 179, 177, and 180.
Set 2. 190, 190, 190, 182, 184, 183, 190, and 139.
Set 3. 172, 180, 165, 160, 175, 168, 180, and 248.

In Set 1, the extreme weights are but 8 pounds apart; three weights are higher than the mean, four are lower than the mean, and one is the same as the mean weight. This set differs from that of problem 2.11 chiefly by being more uniform. In Set 2, the difference between the extremes is 51 pounds, and every weight but one is higher than the mean. This set differs from that in problem 2.11 chiefly in the fact that the mean actually is not very representative of the weights of the squad members. Similar remarks hold for Set 3 except that seven of the eight weights are below the mean. To summarize: Sets 1, 2, and 3 differed from the example of Problem 2.11 in the amount of dispersion, or non-uniformity, among the numbers and in the manner in which it occurred. All sets of data had the same arithmetic mean.

A measure of the dispersion, or variation, of the measurements, X_i, about their arithmetic mean, μ, logically would be based upon the amounts by which the X_i are greater than or less than that mean. It is customary to symbolize those amounts by $x_i = X_i - \mu$, and to call the x_i the *deviations from the mean.* It is observed that when an X is smaller than the mean, the x is negative; when the X is larger than the mean, the corresponding deviation, x, is positive.

In the first numerical example of this chapter, $x_1 = -1$, $x_2 = +2$, $x_3 = -2$, $x_4 = 0$, and $x_5 = +1$. For problem 2.11, $x_1 = -29$, $x_2 = -11$, $x_3 = -16$, $x_4 = +4$, $x_5 = +20$, $x_6 = -7$, $x_7 = +10$, and $x_8 = +29$. It is observed that, at least in these instances, $\Sigma x = 0$.

The truth of the general theorem that the sum of the algebraic values of the x_i always is zero is established as follows. By definition and simple algebra, $\Sigma x = \Sigma(X - \mu) = \Sigma X - \Sigma(\mu)$; but $\Sigma X = N\mu$, and $\Sigma(\mu)$ also $= N\mu$ because this symbol requires that we add N terms obtained by letting i have values from 1 to N, inclusive. The μ stays constant for each i; therefore, $\Sigma(\mu) = N\mu$. Since $\Sigma x = N\mu - N\mu$, it is always equal to zero, as was to be shown.

As a consequence of the truth of the above theorem, a measure of the variation about the arithmetic mean cannot be based upon the algebraic sum of the x_i. Therefore, one of two actions should be taken: (a) Ignore the signs of the x_i and obtain their mean thereafter. Or (*b*), find some other relatively simple function of the x_i which has more of the desirable properties (2.11) to (2.15) than are obtained by method *a*. The latter procedure has proved to be the more successful and therefore will be considered first. As a matter of fact, it involves a function of the squared deviations, x_i^2.

The quantity

$$\sigma = \sqrt{\Sigma(x_i^2)/N}\,, \tag{2.12}$$

where σ, the Greek letter sigma, has been found by statisticians to be a good measure of the variability of a set of numerical measurements about their arithmetic mean. Just why it should be so useful cannot be shown to the student at this time, but it does have more of the desirable properties of measures of variation than any other such measure which has yet been devised. The quantity defined by formula 2.12 is called the *standard deviation* of the X_i about their mean, μ. It would be zero if all the X_i were equal; the more dispersed they are about the mean, the larger the standard deviation tends to be. For example, consider the weights of problem 2.11 and of Set 1. The former obviously are more dispersed and generally more variable than the latter. The two standard deviations are 18.0 and 2.4, respectively, which certainly is a concise way to point out that, although the mean weights of the two squads are the same, their dispersions about that mean are far from the same.

The square of the standard deviation, σ^2, is called the *variance of the X_i about μ*. There are some relatively advanced statistical procedures in which it is preferable to work with the variance instead of the standard deviation, but the latter will be used most of the time in this book.

From the definition of σ contained in formula 2.12 it appears that each x_i must be calculated and squared, but such is not the case. If

the x's are difficult to compute, the following results are useful. In view of the fact that $x^2 = (X - \mu)^2 = (X^2 - 2\mu X + \mu^2)$, it should be clear that $\Sigma x^2 = \Sigma(X^2 - 2\mu X + \mu^2) = \Sigma X^2 - 2\mu\Sigma X + \Sigma\mu^2$. But $\Sigma\mu^2 = N\mu^2$, as explained earlier, and $\mu = \Sigma X/N$; therefore, $-2\mu\Sigma X = -2(\Sigma X)^2/N$ and $\Sigma\mu^2 = +1(\Sigma X)^2/N$. It follows that $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N$. If this substitution is made into formula 2.12, the following alternative method for computing the standard deviation, σ, is obtained:

$$\sigma = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/N}{N}}. \tag{2.13}$$

For a numerical example considered earlier in this chapter, formula 2.13 becomes $\sigma = \sqrt{\dfrac{55 - (15)^2/5}{5}} = \sqrt{2}$. The variance is $\sigma^2 = 2$.

To further illustrate the uses to which σ^2 and σ can be put, consider again the ACE scores of Table 2.01. The arithmetic mean is 95.7. The standard deviation is calculated to be 26.1 (see problem 11 at the end of this section), with the extreme scores being 23 and 183. It is noted that 95.7 is a bit less than midway between the lowest and highest scores, but in general it is quite centrally located in that respect. To obtain a clearer picture of the dispersion of the scores between the extreme scores, and about the mean, the standard deviation will be found to be very useful in subsequent discussions. It is entirely possible for different sets of data to have essentially the same extremes but very different distributions with respect to the mean. The σ^2 and σ will help to describe these differences. This use of the variance and standard deviation is illustrated, in part, by the following discussion.

As the student can verify, approximately 67.1 per cent of the ACE scores in Table 2.01 lie within a distance (on the ACE scale) of 1σ below or above the mean, μ; that is, approximately 67.1 per cent of the 1290 scores are among the numbers from 70 to 121, inclusive. This fact can be put in the following brief form: $\mu \pm 1\sigma$ includes 67.1 per cent of the scores. In a strictly normal population the corresponding percentage is 68.3. Such information sometimes is considered useful in the summarization of sets of numbers like Table 2.01.

Likewise the interval $\mu \pm 2\sigma$ (which includes scores from 44 to 147, inclusive) contains 95.2 per cent of the 1290 scores in the table. If

this population were perfectly normal, that percentage would be 95.4. Also the interval $\mu \pm 3\sigma$ includes 99.8 per cent of the ACE scores, whereas a normal population would have 99.7 per cent of its members in that interval. The reader can determine how closely the population of Table 2.01 conforms to the normal requirement that 38.3 per cent of the measurements shall lie not more than one-half a standard deviation above or below the mean, μ.

More discussion of normality and of population distributions will come later; the point of the above discussion is that knowledge about the mean and the standard deviation is useful in the study of one of the most important types of populations of data.

PROBLEMS

1. Calculate the arithmetic mean and the standard deviation of the following numbers: 2, 3, 9, 7, 5, 4, 10, 6, 3, 1, and 5.

2. Make up three sets of numbers, each of which has $\mu = 7$.

3. Compute the x_i for problem 1 and verify that $\Sigma x_i = 0$.

4. Given the numbers 0, 8, 0, 1, 1, 1, 10, 2, 1, 1, 2, 3, 0, and 1, compute μ. Does μ seem to you to be a good average for these numbers? Why? *Ans.* 2.21.

5. Suppose that the mathematics grades for a certain class were 54, 95, 68, 71, 87, 75, 84, 63, 76, 81, 70, 90, 73, 77, and 61. Calculate μ and σ, using the individual x_i first and then using formula 2.13 for σ.

6. The following percentages of protein in samples of pasture grasses were made available by Dr. George Wise, formerly of the Department of Dairy Husbandry at Kansas State College. Compute μ, σ^2, and σ, given that $\Sigma X = 1423.33$ and that $\Sigma X^2 = 21{,}924.2025$.

22.59	15.26	15.63	13.52	10.82	9.29	12.25	21.07	18.83
16.52	15.54	16.17	10.03	15.71	12.79	13.11	14.85	11.45
11.97	11.07	15.26	9.31	12.30	13.04	14.19	12.94	14.36
15.02	11.15	12.08	15.41	8.56	9.09	13.07	12.51	18.91
14.28	14.54	13.68	11.78	14.22	13.07	14.27	10.27	11.01
11.66	8.19	6.75	14.48	15.98	14.36	15.24	14.48	14.05
15.02	15.41	10.02	9.96	12.34	16.26	10.19	14.20	12.56
9.74	14.34	13.07	12.33	11.57	15.48	11.74	9.39	6.47
25.09	23.23	16.75	10.62	16.30	17.29	20.63	13.76	11.88
10.57	8.37	26.20	26.74	22.02	20.60	20.36	15.83	14.11
22.42	19.47	17.98	20.32	14.83	13.03	10.31	8.95	11.57
16.03								

Ans. 14.23, 16.66, 4.08.

7. The following are bearings taken with a radio direction finder on a signal sent repeatedly from a fixed location. Compute their arithmetic mean and standard deviation as though these data constituted a population.

X: 9, 8, 6, 4, −10, 6, 7, 10, 8, 9, 7, 6, 8, 8, 10, 10, 8, 8, 10, 9, 10, 7, 7, 3, and 8 (degrees from north).

8. The following data are like those of problem 7, but taken on a different direction finder. Obtain the variance and the σ for these data given that $\Sigma X^2 = 116{,}830$, $\Sigma X = 1708$.

X: 66, 68, 69, 68, 71, 70, 66, 70, 68, 67, 68, 73, 68, 65, 72, 73, 68, 67, 69, 65, 64, 66, 67, 70, and 70 (degrees from north). *Ans.* 5.58, 2.36.

9. Work problem 8, after subtracting 60 degrees from each bearing. How much were μ and σ changed? How much were the x_i changed? What if only 50 degrees had been subtracted?

10. Use one-sixth of the range in problem 6 as an estimate of the standard deviation, and compare this estimate with the true standard deviation. *Ans.* 3.38.

11. Given that for Table 2.01 $\Sigma X = 123{,}445$, and $\Sigma X^2 = 12{,}693{,}988$. Calculate the arithmetic mean and the variance about the mean, μ.

12. Given the following six yields of Ponca wheat at Manhattan, Kansas, compute their mean after first subtracting 27 from each number. (Data provided by Department of Agronomy, Kansas State College.) Yield (bushels/acre): 27.2, 40.9, 46.0, 38.1, 43.8, 46.3.

Ans. 13.2 bushels per acre; therefore, true mean = 40.2.

13. The test weights corresponding to the bushel yields of problem 12 were as follows (data from same source): 59.3, 60.7, 60.6, 60.2, 61.9, 58.1. Calculate the midrange, the arithmetic mean, and the variance.

14. In problems 12 and 13, which of the types of measurement, yield or test weight, gives the more consistent results according to this evidence? Give reasons.

15. Write down every fiftieth score in Table 2.01, starting in the upper left-hand corner of the table and working from left to right. Compute the arithmetic mean of the sample thus obtained and compute the percentage error relative to the true mean, 95.7.

2.2 THE AVERAGE (OR MEAN) DEVIATION

A measure of the variation about the arithmetic mean based on the numerical values (signs are ignored) of the x_i was mentioned in the preceding section. If we were to set out to devise a simple and logical way to measure the dispersion of a group of numbers about some point, such as the arithmetic mean, we might well decide to use what is called the *average (or mean) deviation.* It is the arithmetic mean of the x_i each taken as a positive number regardless of its actual sign. For example, consider the weights of problem 2.11. The numerical deviations from the mean, μ, were found to be: 29, 16, 11, 7, 4, 10, 20, and 29 pounds. On the average—that is, considering the arithmetic mean as the average—the weights of those

basketball players differed from the mean weight of the group by

$$(29 + 16 + 11 + 7 + 4 + 10 + 20 + 29)/8$$

$$= 126/8 = 15.75 \text{ pounds.}$$

Then the average deviation for these weights is 15.75 pounds.

Symbolically the average deviation is defined by

$$AD = \frac{\Sigma|X - \mu|}{N}, \quad \text{or} \quad AD = \frac{\Sigma|x|}{N}, \tag{2.21}$$

where $|x|$ = a deviation from μ taken as a positive number whether the corresponding X was larger than μ or smaller than μ.

For the weights just used for illustration, $\sigma = 18.11$ pounds. The standard deviation is larger than the mean deviation, as is usual. The standard deviation is much more widely used than the mean deviation partly because it has many useful applications in sampling studies, which after all is by far the more fruitful and interesting field of statistical analysis.

2.3 OTHER AVERAGES

Another average which is simple to compute and of rather wide application for descriptive purposes is the median, symbolized as *md*. The median of a set of numerical measurements is intended to be a number such that one-half the numbers are less than or equal to the median, and the other half are greater than or equal to the median; that is, the median is exactly in the middle of the set of numbers in order of size, if such is possible.

It is necessary—either actually or effectively—to list the numbers in order of size before the median can be determined accurately. Such an ordered group of numbers is called an *ordered array*. Thus the numbers 1, 5, 2, 3, 0, 1, 8, and 10 do not form an ordered array, whereas these same numbers listed as 0, 1, 1, 2, 3, 5, 8, and 10 do constitute an ordered array.

With the definition of an ordered array established, it is convenient to define the median of N numbers: $X_1, X_2, \ldots, X_N$ as the $[(N + 1)/2]$th number in the array, starting with the lowest number in the array. It is noted that only if N is odd does such an ordinal number exist; but it is sufficient herein to define an "ordinal" number like 4.5 to be a number which is just midway in magnitude between the fourth and fifth numbers in the array. For example, for the

array used above, $N = 8$, so that $(N + 1)/2 = 4.5$. Hence the median is $md = 2.5$, a number midway in size between the 2 (which is the fourth number in the array) and the 3 (which is the fifth number in the array).

It can be seen, with a little study, that the median is an average which will be nearer the region of concentration of the numerical measurements in a population than the arithmetic mean if there are a few "stray" numbers at one end of the scale of measurement. For example, consider the following simulated annual salaries (in thousands of dollars) of college instructors in one department: 3.1, 3.5, 3.5, 3.6, 3.6, 3.6, 3.8, 3.8, 3.8, 3.9, 3.9, 4.0, 4.0, 4.0, 4.4, 4.8, 5.0, 6.5, 8.4, 8.7, and 8.8. For these data $N = 21$, $\Sigma X = 96.1$, $\mu = 4.7$, and $md = 3.9$. It is seen that eleven of that staff are receiving within \$300 of the median salary whereas only three are that close to the arithmetic mean. The arithmetic mean exaggerates the typical salary in a very real sense for all but the fortunate six at the top. In situations of this sort—which will be described later as skewed distributions when more data are involved—the median is a better average than the arithmetic mean when its purpose is to describe the typical measurement in the population.

If a fairly large group of numbers is to be summarized and the median is a desirable average to use, the midrange can be helpful in reducing the necessary labor. For example, the MR for Table 2.01 is 103; hence we can hope that the median has about the same size. On this assumption we can count the scores greater than or equal to 100 and thereafter determine exactly the 645.5th number in the array without excessive labor. It thus is found that $md = 97$.

There are three other averages which will be considered, and which will find occasional application to numerical measurements. One is the *mode* (MO). The mode is defined to be that measurement which occurs in a given set of numbers with the greatest frequency, if such a number exists. For example, the mode of the set 5, 8, 9, 10, 10, 10, 11, 13, and 15 is MO = 10. If some number in a group of data decisively occurs with the greatest frequency, the mode may well be the average to use; but such is rather rarely the case.

The *geometric mean* of $X_1, \ldots, X_N$ is defined as the Nth root of the product of these numbers. Symbolically,

$$\text{GM} = \sqrt[N]{X_1 X_2 X_3 \cdots X_N}. \tag{2.31}$$

Under most circumstances it is easier to compute the geometric mean from the relation

$$\log (\mathrm{GM}) = \frac{1}{N}\Sigma(\log X). \tag{2.32}$$

As an illustration consider the numbers 2, 5, 8, and 15. By definition GM = the fourth root of the product $(2)\cdot(5)\cdot(8)\cdot(15)$; but, using logarithms to the base 10, one has log GM = (log 2 + $\cdots$ + log 15)/4 = 0.7698. The antilog 0.7698 is approximately 5.9, which is the geometric mean of the given set of numbers. The geometric mean is useful in the calculation of certain index numbers, in studies of biological growth, and, in general, whenever the statistical array indicates that a geometrical series is involved. Obviously the geometric mean is not used if any $X = 0$ or if the product under the radical is negative.

The last average to be considered herein is called the *harmonic mean.* It also is used only in specialized circumstances, but the possession of some information about it will help to round out the reader's knowledge regarding statistical averages.

The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of a given group of numbers; or

$$\mathrm{HM} = \frac{1}{[\Sigma(1/X_i)]/N} = \frac{N}{\Sigma(1/X_i)}. \tag{2.33}$$

For example, if the X's are 3, 8, 2, 5, and 2 the denominator is $\Sigma(1/X_i) = 1/3 + 1/8 + 1/2 + 1/5 + 1/2 = 1.6583$, approximately; hence HM = 5/1.6583 = 3.02. One use of the harmonic mean comes when rates of some sort are involved. Consider this problem. A man drives the first 50 miles of a trip at 50 mph, and the second 50 miles at a rate of 60 mph. What is his average rate for the trip? By the usual definition, the average rate is obtained by dividing the total distance traveled by the total time taken to go that distance. The distance traveled was 100 miles. The first 50 miles took one hour, and the second 50 miles took five-sixths of an hour; hence the total time was 11/6 hours. Therefore, the average rate of speed was 100/(11/6) = 600/11 = 54 and 6/11 mph. The harmonic mean of 50 and 60 also is 54 and 6/11 mph; that is, the required average rate is just the harmonic mean of the two rates in this instance. It is noted that the distance traveled was the same for each rate of speed.

Now suppose that a person drives for one hour at 50 mph and then the second hour at 60 mph. What is the average rate of speed during

this trip? The total distance traveled is 110 miles, and it took 2 hours; therefore the average rate is 55 mph. But that is just the arithmetic mean of the two rates. It is seen that when time (hours) was fixed in the problem, the appropriate average was the arithmetic mean; when the distance (miles) was fixed and time was variable according to the speed of travel, the appropriate average was the harmonic mean.

In general, the proper average to use in any particular situation either will be determined at the outset by previous practices in the particular sort of work, or it can be determined by a bit of preliminary study of the matter. Hence no attempt will be made to lay down rules. However, it should be apparent to the reader that when a body of data is to be summarized statistically there may be several possible choices of averages and also of measures of variation. We should be fully conscious of this fact when we compute averages, or when we interpret those averages computed by someone else.

PROBLEMS

1. The following numbers are salaries (in dollars) in a public school system before World War II: 1300, 1500, 1300, 1350, 1600, 1250, 1400, 1350, 1800, 4500, 1450, 3000, 2200, 1250, 1300, 1550, 1700, 1600, 1350, 1400, 1450, 1750, 1500, 1600, and 1400. Calculate the arithmetic mean and the median, and state which average you consider the more typical of these salaries.

2. Suppose, in problem 1, that the following raises in salary were given: \$2200 to \$3000, \$3000 to \$3500, \$4500 to \$5000, \$1800 to \$2200, \$1750 to \$2200; and all others are given a \$100 raise. The salaries of problem 1 add to \$41,850, and the total of the raises is \$4650, approximately 11 per cent of \$41,850. Is it then fair to state that those teachers received an 11 per cent increase in salary, on the average?

3. Compute the geometric mean of 76.3 and 85.1.

4. Sometimes the median can be used as an average when numerical measurements are not employed. For example, some radio direction finder networks rate their bearings as to their quality, ranging from *A*, (best) through *B*, *C*, *F*, and *P*. If the median does not turn out to be indeterminate (as by falling between two different letters) it may be useful in describing the average quality of the readings. Obtain the median quality of the following quality ratings: *A*, *C*, *C*, *B*, *A*, *B*, *C*, *B*, *P*, *F*, *C*, *C*, *F*, *B*, *C*, *A*, and *B*. *Ans. C.*

5. What is the modal quality rating for problem 4?

6. Compute the geometric mean of the salaries in problem 1 to the nearest dollar. *Ans.* \$1591.

7. Suppose that peaches were bought in three different areas for \$3 per bushel, \$2 per bushel, and \$4 per bushel, respectively. Suppose also that \$24 was spent for peaches at each price level. What was the average price paid per bushel?

8. Do as in problem 7 except to consider that 10 bushels of peaches were bought at each price. *Ans.* \$3 per bushel.

9. Suppose that 5 bushels of the \$3 peaches, 10 of the \$2 peaches, and 15 bushels of the \$4 peaches were purchased. What was the average price per bushel?

2.4 FREQUENCY DISTRIBUTIONS

To introduce the method of constructing frequency distributions, and to show what sort of information can be derived from them, reference is made to the numbers of Table 2.01. It is possible by means of problem 11, section 2.1, to calculate that $\mu = 95.7$ and $\sigma = 26.1$. These statistical constants furnish some useful information about the population of scores, but they fail quite badly to summarize them adequately. For example, a person who made a score of 120 could not be told accurately how he compared with the others taking this same test, and that information usually is important in the use of such tests. One way to obtain this sort of information is to construct frequency distributions and graphs which display the outstanding features of the population.

Two types of summaries of distributions will be considered both numerically and graphically: a frequency and a relative cumulative frequency (or *r.c.f*) distribution. Both distributions will be described by means of a grouping of the individual scores into convenient score classes, even though such frequency distributions could be made without grouping the members of the population into classes. The scores then lose their individual identities and become members of ten to twenty groups. The data become more manageable, and little accuracy is lost in the process. To illustrate, consider Table 2.01 again. The extreme scores have been noted previously to be 23 and 183 so that the range is 160. If the range is divided by 10 a quotient of 16 is obtained. Classes of that length would give the minimum acceptable number of classes; hence for convenience in tallying (as shown below) the class interval will be taken as 15. Table 2.41 was constructed starting with the lowest score at 10 purely because it was convenient and the lowest class included the lowest ACE score in Table 2.01. The actual tallying of the data is shown, as is the summarization of the tallies into a frequency (f) for each class. In Table 2.42 a more concise form of the frequency distribution is shown along with the *r.c.f.* distribution. The latter distribution gives the decimal fraction of the ACE scores which were less than or equal to the upper limit of the corresponding score class at the left. For example, practically one-third (actually .332) of the scores were at or below a score of 84, according to Table 2.42.

TABLE 2.41

FREQUENCY DISTRIBUTION TABLE FOR THE DATA OF TABLE 2.01, SHOWING TALLYING

Score Class		*f*
175–189	///	3
160–174	////	4
145–159	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ///	33
130–144	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 /	81
115–129	𝍸 /	186
100–114	𝍸 ///	278
85– 99	𝍸 //	277
70– 84	𝍸 ////	209
55– 69	𝍸 //	132
40– 54	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ////	69
25– 39	𝍸 𝍸 𝍸	15
10– 24	///	3
	Total	1290

To assist in the drawing of conclusions from Tables 2.41 and 2.42, the information they contain is presented in graphic form in Figure 2.41. Conclusions such as the following can be drawn from that figure and the tables pertaining to Table 2.01:

(2.41) Only about 7 per cent of the students made a score less than 55. This information can be read directly from Table 2.42 and verified approximately from the *r.c.f.* curve of Figure 2.41. Also, approximately 50 per cent of the students made a score of 98 or more, a fact which corresponds closely with the fact that the exact median is 97. Information of this sort can be obtained from Figure 2.41 by reading horizontally from *r.c.f.* = .50 over to the *r.c.f.* curve and then vertically downward to the scale of ACE scores.

TABLE 2.42

THE RELATIVE CUMULATIVE FREQUENCY DISTRIBUTION OF THE DATA IN TABLE 2.01

Score Class	*f*	*c.f.*	*r.c.f.*
175–189	3	1290	1.000
160–174	4	1287	.998
145–159	33	1283	.995
130–144	81	1250	.969
115–129	186	1169	.906
100–114	278	983	.762
85– 99	277	705	.546
70– 84	209	428	.332
55– 69	132	219	.170
40– 54	69	87	.067
25– 39	15	18	.014
10– 24	3	3	.002
Total	1290		

(2.42) With specific reference to the student mentioned above who made a score of 120, it is learned that about 82 per cent of the students did no better; or only about 18 per cent beat him. Hence he should be considered to be quite high in aptitude and intelligence relative to those who took that same test, and would be expected to do rather well in college.

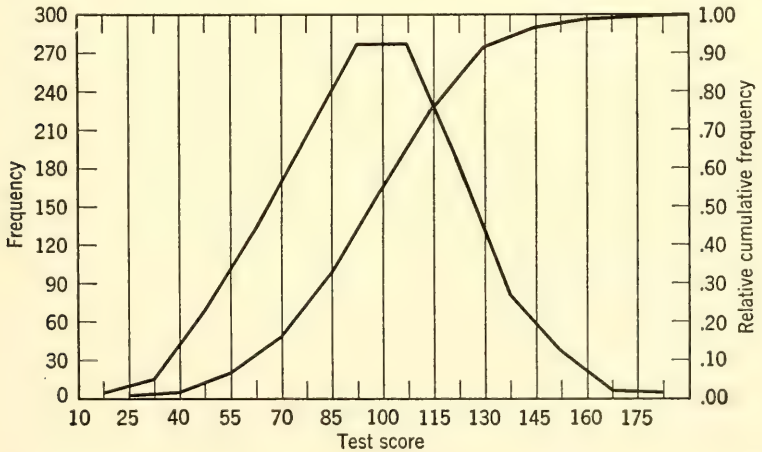


Figure 2.41. Frequency distributions of the ACE test scores listed in Table 2.01. Data furnished by the Counseling Bureau of Kansas State College.

Other similar information can be read from the above tables and graphs. Moreover, it will be shown later that the *r.c.f.* graph can be employed to determine quartile, decile, and even percentile limits if the graph is drawn with sufficient care. In addition it will be shown in the next section that in some situations good approximations to μ and σ can be computed from the frequency distribution table. In brief, the frequency and *r.c.f.* tables and graphs serve as visual guides and as sources of good approximations. If more precise information is needed or desired (and it seldom is), one can analyze the individual observations.

The graph of the *r.c.f.* distribution sometimes is called an *ogive.*

PROBLEMS

1. Following are some numbers of house flies counted on individual dairy cows which had been sprayed with a 3 per cent solution of Thanite in 40 oil: 35, 37, 41, 103, 174, 7, 11, 32, 23, 7, 6, 3, 14, 23, 23, 36, 25, 27, 3, 3, 13, 14, 14, 6, 15, 9, 11, 21, 9, 12, 3, 15, 19, 29, 26, 1, 8, 4, 9, 7, 12, 5, 1, 3, 5, 60, 11, 6, 4, 7, 22, 28, 5, 3, 6, 15, 1, 2, 11, 4, 27, 1, 0, 0, 19, 6, 2, 3, 4, 13, 5, 12, 11, 14, 45, 4, 38, 5, 17, 27, 39, 33, 13, 9, 8, 33, 19, 6, 12, 32, 11, 35, 18, 11, 25, 23, 45, 30, 4, 4, 15, 15, 16, 11, 16, 18, 32, 49, 129, 7, 21, 26, 76, 40, 5, 7, 5, 7, 4, 62, 91, 133, 61, 59, 20, 26, 10, 12, 6, 7, 8, 8, 2, 24, 21, 51, 110, 11, 6, 4, 4, 5, 5, 5, 13, 3, 6, and 7. Construct frequency and relative cumulative frequency distributions for these data, estimate the median, and decide whether μ or *md* is the preferable average for these data. Is this a skewed distribution? Use class intervals: 0–8, 9–17, etc.

2. Graph the distributions asked for in problem 1.

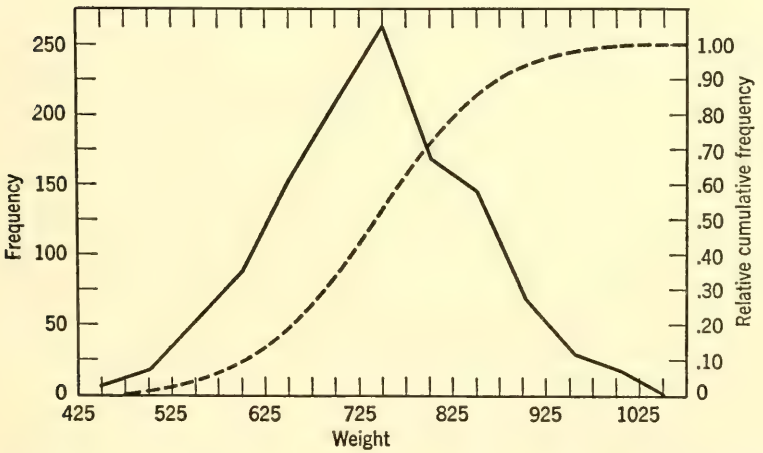
3. Compute the arithmetic mean and the median of the counts in problem 1. Compare them, and draw appropriate conclusions. $\Sigma X = 3031$, $\Sigma X^2 = 163{,}439$.

4. Estimate from the ogive (*r.c.f.* curve) for problem 1 what percentage of the fly counts lie between 5 and 25, inclusive. Check your calculation by actually counting in problem 1. *Ans.* Graph, 46; by count, 45.

5. Use the following frequency distribution table of 8-week weights (in grams) of male White Rock chickens raised at the Kansas State College Poultry Farm and the accompanying graphs to: (*a*) estimate μ, (*b*) determine what percentage of the weights exceeded 800 grams, (*c*) determine the range covered by the "middle" 50 per cent of the weights, that is, excluding the upper and lower 25 per cent of the weights.

6. Construct the frequency and the relative cumulative frequency distributions for the following counts, which are similar to those numbers in problem 1: 6, 8, 13, 36, 48, 65, 34, 24, 49, 24, 40, 18, 20, 34, 87, 28, 14, 30, 24, 53, 57, 93, 36, 80, 33, 48, 57, 98, 73, 135, 21, 40, 32, 58, 4, 20, 30, 33, 20, 22, 28, 11, 23, 46, 41, 41, 44, 23, 18, 41, 48, 81, 80, 70, 5, 2, 13, 21, 21, 171, 1, 7, 10, 5, 2, 17, 9, 35, 6, 8, 10, 23, 19, 3, 25, 16, 131, 19, 19, 24, 12, 10, 4, 5, 2, 14, 17, 18, 10, 8, 4, 0, 4, 12, 14, 111, 17, 33, 3, 2, 7, 10, 17, 4, 5, 2, 48, 4, 11, 31, 18, 32, 26, 3, 18, 19, 101, 10, 10, 3, 27, 14, 29, 24, 13, 26, 31, 5, 20, 16, 13, 6, 7, 32, 17, 25, 6, 8, 5, 24, 13, 8, 7, 2, 1, 3, 26, 38, 44, 3, 5, 6, 22, 28, 16, 22, 8, 19, 12, 3, 24, 10, 8, 33, 20, 29, 3, 15,

Weight Class	Frequency, f	Cumulative Frequency, $c.f.$	Relative Cumulative Frequency, $r.c.f.$
1025–1074	1	1217	1.000
975–1024	16	1216	.999
925– 974	29	1200	.986
875– 924	66	1171	.962
825– 874	148	1105	.908
775– 824	169	957	.786
725– 774	265	788	.647
675– 724	210	523	.430
625– 674	155	313	.257
575– 624	85	158	.130
525– 574	51	73	.060
475– 524	17	22	.018
425– 474	5	5	.004
Total	1217		



18, 20, 8, 13, 17, 27, 23, 23, 10, 25, 13, 10, 12, 10, 6, 6, 14, 24, 61, 25, 26, 21, 12, 15, 18, 19, 26, 21, 11, 0, 0, 8, 34, 66, 32, 7, 8, 23, 20, 24, 62, 8, 15, 19, 33, 20, 51, 11, 20, 13, 27, 15, 10, 16, 16, 5, 4, 24, 30, 37, 26, 17, 14, 15, 6, 3, 22, 53, 54, 74, 1, 10, 12, 22, 49, 52, 31, 7, 20, 23, 28, 56, 2, 6, 6, 30, 30, 38, 1, 2, 4, 21, 51, 14, 5, 17, 21, 28, 9, and 7.

7. Construct the relative cumulative frequency distribution for the data of the preceding problem, and read from it the value of the median. Check that result with the value obtained from an ordered array of those data.

8. Use the graph of the relative cumulative frequency distribution for the counts in problem 6 to determine what percentage is less than or equal to ten flies per cow. *Ans.* 29 per cent.

9. Within what extremes did the lowest one-fourth of the counts listed in problem 6 lie? The middle one-fourth? The highest one-fourth?

10. Take any newspaper with at least one hundred bond or stock quotations and make frequency and relative cumulative frequency distributions of those prices.

2.5 CALCULATION OF THE ARITHMETIC MEAN AND THE STANDARD DEVIATION FROM FREQUENCY DISTRIBUTION TABLES

If the frequency distribution table has class intervals of equal lengths, approximate values can be computed for μ and σ with a considerable saving in labor as compared to their computation from the individual measurements. The method of computation involves the sole assumption that the numbers grouped into each class actually were at the midpoint of their class. Although that assumption is not strictly correct, the individual discrepancies usually balance out so well that the net error is unimportant in practice. If it should be decided that some additional accuracy is needed, Sheppard's corrections for grouping can be employed. (See, for example, Kenney, *Mathematics of Statistics,* Part One, D. Van Nostrand.)

Table 2.51 presents methods for computing μ and σ which follow directly from the definitions of these quantities if all the data in a class are considered to be at the midpoint of the class. For example, the data in Table 2.51 would be considered to be 22.5, 22.5, 17.5, 17.5, 17.5, 17.5, 17.5, 17.5, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 7.5, 7.5, 7.5, 7.5, 7.5, 2.5, and 2.5 each midpoint appearing precisely that number of times indicated by the class frequency, f. The student should check the fact that the sum of these products is 317.5, which is shown in Table 2.51 as the total of the column headed "$f \cdot z$." The symbol z is employed to denote the midpoint of the class interval. For convenience and for uniformity of procedure, the midpoint of a class of data measured on a continuous scale is defined to be the lower limit of the class (as recorded in the table) plus one-half the length of the class interval. Also, the length of the class interval, for such data, is defined as the numerical difference between any two successive left- or right-hand end points of classes. Thus the midpoint of the class "20–24.999 ..." is $20 + (1/2)(5) =$ the z for this class interval.

The reader should note that there will be circumstances in practice which will justify a different determination of the midpoint, z. For

example, if some objects have been weighed to the nearest pound it is reasonable to suppose that an interval written as 20–24 actually means 19.5 to 24.5. If the class intervals are so written, the above-stated rules apply. The length of the class interval will be 5 as before, but the midpoint will be computed as $19.5 + (1/2)(5) = 22.0$ instead of 22.5, as it would be if computed on the assumption that the interval started at 20.

If we are summarizing data which can only be integers, a class interval such as 20–24 should include only the numbers 20, 21, 22, 23, and 24. It then is reasonable to take $z = 22$. The length of the class interval should be taken as 5 again so that the numerical distance between midpoints will coincide with the length of the class interval, which seems to be a reasonable requirement. The proper procedure for other methods of measurement can be figured out along the lines just outlined.

TABLE 2.51

ILLUSTRATION OF A METHOD OF CALCULATING μ AND σ FROM THE DATA IN A FREQUENCY DISTRIBUTION TABLE WITH CLASS INTERVALS OF EQUAL LENGTHS

Class Interval	Midpoint z	Frequency f	$f \cdot z$	$z - \mu$	$f \cdot (z - \mu)^2$
20–24.9	22.5	2	45.0	9.8	192.08
15–19.9	17.5	6	105.0	4.8	138.24
10–14.9	12.5	10	125.0	− 0.2	0.40
5– 9.9	7.5	5	37.5	− 5.2	135.20
0– 4.9	2.5	2	5.0	−10.2	208.08
	Totals	25	317.5		674.00

$$\mu = \frac{\Sigma f \cdot z}{\Sigma f} = \frac{317.5}{25} = 12.7; \quad \sigma = \sqrt{\frac{\Sigma f \cdot (z - \mu)^2}{\Sigma f}} = 5.19, \text{ approximately.}$$

Another, and easier, method of computing μ and σ from a frequency distribution table with equal class intervals is illustrated in Table 2.52 along with a partial demonstration of the generality of the method. The procedure involves the same assumption made above and produces exactly the same values for μ and σ. However, in this method the class interval is employed as the computational unit, with the result that the sizes of the numbers needed in the process are smaller than those of Table 2.51. This makes the computations simpler.

The procedures outlined in Table 2.52 can be justified as follows. It is obvious that the midpoints, z_i, of Table 2.51 can be rewritten as follows:

$$2.5 = 12.5 - 2(5),$$

$$7.5 = 12.5 - 1(5),$$

$$12.5 = 12.5 + 0(5),$$

$$17.5 = 12.5 + 1(5),$$

$$22.5 = 12.5 + 2(5).$$

Then

$$\begin{aligned}\Sigma(f\cdot z) &= 2[12.5 + 2(5)] + 6[12.5 + 1(5)] + 10[12.5 + 0(5)] \\ &\quad + 5[12.5 - 1(5)] + 2[12.5 - 2(5)]; \\ &= (2 + 6 + 10 + 5 + 2)(12.5) \\ &\quad + [2(2) + 6(1) + 10(0) + 5(-1) + 2(-2)](5);\end{aligned}$$

or, a bit more generally,

$$\Sigma(f\cdot z) = (\Sigma f)(12.5) + \Sigma(f\cdot d)(I),$$

if $d = +2$ for the top class of Tables 2.51 and 2.52, $d = +1$ for the next class down, etc., until $d = -2$ for the bottom class of each of those tables. (The symbol I stands for the length of the class interval.) Therefore,

$$\mu = \frac{\Sigma(f\cdot z)}{\Sigma(f)} = 12.5 + \frac{\Sigma(f\cdot d)(I)}{\Sigma(f)} \tag{2.51}$$

is the approximation to μ which can be obtained from the frequency distribution table. It should be clear that some other midpoint besides 12.5 could have been used without changing the answer obtained. Hence if there is any best choice of a midpoint to use as a base point (with $d = 0$), that choice must rest on its leading to simpler computations. Generally, the d should be taken as zero for the class with the greatest frequency. If the distribution is quite non-symmetrical, it is advisable to shift the choice one way or the other so that the positive and negative fd's will be more nearly balanced. This rarely will be more than two classes from the one with the greatest frequency, f.

It can be seen in Tables 2.51 and 2.52 that when $d = 0$ for the class with the largest frequency (f) the resulting arithmetic involves smaller numbers than for the other methods. It should be noted, again, that all three of the methods illustrated give exactly the same answers; the only differences lie in the ease of computation.

TABLE 2.52

ILLUSTRATION OF A SIMPLIFIED METHOD FOR COMPUTING μ AND σ FROM A FREQUENCY DISTRIBUTION TABLE WITH EQUAL CLASS INTERVALS

Class Interval	Mid-point z	Frequency f	Method A (d taken 0 for interval with greatest frequency) d	$f \cdot d$	$f \cdot d^2$	Method B (d taken 0 for interval with lowest frequency) d	$f \cdot d$	$f \cdot d^2$
20–24.9	22.5	2	+2	4	8	0	0	0
15–19.9	17.5	6	+1	6	6	−1	− 6	6
10–14.9	12.5	10	0	0	0	−2	−20	40
5– 9.9	7.5	5	−1	−5	5	−3	−15	45
0– 4.9	2.5	2	−2	−4	8	−4	− 8	32
	Totals	25		+1	27		−49	123

$$\mu = (z \text{ for class with } d = 0) + \frac{\Sigma(f \cdot d)}{\Sigma(f)}(I) = 12.5 + (1/25)(5) = 12.7.$$

$$\sigma = (I)\sqrt{\frac{\Sigma(f \cdot d^2) - [\Sigma(f \cdot d)]^2/\Sigma(f)}{\Sigma(f)}} = (5)\sqrt{\frac{27 - (1)^2/25}{25}} = 5.19.$$

The derivation of the formula shown for σ is more difficult than that for μ, as might be expected, but it can be obtained by elementary algebra, formula 2.13, and by expressing each z in terms of the one for which $d = 0$. This derivation will be left as an exercise for the ambitious student.

The methods just described can be applied to obtain satisfactory approximations to the arithmetic mean and the standard deviation of the ACE scores in Table 2.01, a task which clearly would be quite laborious if formulas 2.13 and 2.11 were to be employed directly on the 1290 numbers in that table. The d are taken as zero for the class with a frequency of 277 (Table 2.42) instead of the class with $f = 278$ because they are essentially the same size and the distribution is a bit non-symmetrical (or skewed) in the direction of the lower ACE scores. The general result is to have smaller d's with the larger fre-

quencies, and hence to make the computations somewhat easier. Following are the required calculations based on Table 2.42; it is assumed that the scores are necessarily integers.

z	f	d	$f \cdot d$	$f \cdot d^2$
	3	6	18	108
	4	5	20	100
	33	4	132	528
	81	3	243	729
	186	2	372	744
	278	1	278	278
92	277	0	0	0
	209	−1	−209	209
	132	−2	−264	528
	69	−3	−207	621
	15	−4	− 60	240
	3	−5	− 15	75
	$\Sigma(f) = 1290$		$\Sigma(f \cdot d) = +308$	$4160 = \Sigma(f \cdot d^2)$

By the formulas previously used,

$\mu = 92 + (308/1290)(15) = 95.6$ compared to the true mean of 95.7.

$\sigma = (15)\sqrt{\dfrac{4160 - (308)^2/1290}{1290}} = 26.7$ compared to the true value of 26.1.

In view of the fact that the scores were integers, these approximations certainly would be considered satisfactory, and the time and labor saved by these methods are considerable.

The distribution of the population of ACE scores is rather symmetrical, that is, there is a region of high frequency about halfway between the extremes, and the frequency of occurrence of scores away from this region diminishes at about the same rate as scores are considered equally far above and below the region of highest frequency. This distribution is shown in Figure 2.41. With this type of distribution the arithmetic mean is an excellent average to use as a part of the description of the population.

Other distributions may be non-symmetrical, or skewed. For such populations the median often serves as the more descriptive average. As a matter of fact, the difference between the sizes of the arithmetic mean and the median is an indication of the degree of skewness or lack of symmetry, in the frequency distribution. If the distribution

is perfectly symmetrical (no skewness) the arithmetic mean and the median are equal. The more skewed the distribution, the farther apart the median and this mean may become.

PROBLEMS

1. Compute the arithmetic mean of the following numbers by Method A of Table 2.52:

X: 24, 8, 7, 14, 21, 10, 12, 14, 17, 9, 11, 5, 15, 16, 8, 2, 13, 18, 12, 3, 15, 4, 16, 19, and 11.

Use class intervals 2–5.9 . . . , etc., to 22–25.9 . . . .

2. Compute the arithmetic mean and the standard deviation for problem 1 exactly, and compare with the values obtained by the methods of Table 2.52.

Ans. $\mu = 12.2$, $\sigma = 5.5$; they are 12.8 and 5.3 by table.

3. Put the numerical measurements of problem 1, section 2.4, into a frequency distribution table with class intervals of equal lengths, and compute the standard deviation of those counts.

4. Do as in problem 3 for the data of problem 6, section 2.4. *Ans.* 23.7.

5. Calculate the mean and standard deviation for the hypothetical data in the following table. Also, compare six times the standard deviation with the range as nearly as it can be derived from the table.

Class Interval	Frequency
28–29.9 . . .	5
26–27.9 . . .	16
24–25.9 . . .	29
22–23.9 . . .	41
20–21.9 . . .	50
18–19.9 . . .	45
16–17.9 . . .	32
14–15.9 . . .	20
12–13.9 . . .	9
10–11.9 . . .	3
Total	250

6. Graph the relative cumulative frequency distribution for problem 5 and read from it the percentage of the measurements which exceed 23. Which exceed the arithmetic mean. Which lie between the mean and 23.

Ans. 28%, 50%, 22%.

7. What is the median for problem 5 as read from the *r.c.f.* curve? Which is the modal class? Would you expect the mode and the median to differ by as much as two units; or less than two units? Give reasons.

8. Graph the following actual or estimated age distributions of the United States population and draw appropriate conclusions regarding apparent trends during the decades covered. Consider top class as 0–4 and bottom one as 75–79.

(Data from *Current Population Report,* Population Estimates, August 10, 1950, Bureau of the Census.) Numbers are thousands.

Age Class	1940	1950	1960
Under 5 years	10,542	16,580	13,121
5–9	10,685	13,959	15,693
10–14	11,746	11,349	17,439
15–19	12,334	10,561	13,860
20–24	11,588	11,585	11,274
25–29	11,097	12,161	10,725
30–34	10,242	11,439	11,771
35–39	9,545	10,960	12,211
40–44	8,788	10,061	11,377
45–49	8,255	9,231	10,713
50–54	7,257	8,254	9,583
55–59	5,868	7,440	8,469
60–64	4,760	6,210	7,205
65–69	3,748	4,611	5,980
70–74	2,561	3,282	4,428
75 years and over	2,655	3,716	5,083

9. Change the top and bottom classes in problem 8 to 0–4 and 75–79, and then compute μ and *md*. Which do you consider the better average here?

10. Given the following frequency distribution of the minimum annual temperatures in 116 cities of Kansas, compute the arithmetic mean and the standard deviation:

Temperature Interval	f	Temperature Interval	f
−17.4 to −15.0	3	−32.4 to −30.0	17
−19.9 −17.5	7	−34.9 −32.5	1
−22.4 −20.0	17	−37.4 −35.0	2
−24.9 −22.5	24	−39.9 −37.5	0
−27.4 −25.0	24	−42.4 −40.0	1
−29.9 −27.5	20		—
			$\Sigma(f) = 116$

Ans. −25.7°, 4.4°.

11. Given the following table of average lengths of growing season for ninety-five Kansas counties, compute the mean length of growing season, and also the median length. Which would be the most descriptive average here?

Class Interval	f	Class Interval	f
198–202 days	3	168–172 days	15
193–197	8	163–167	9
188–192	13	158–162	7
183–187	18	153–157	2
178–182	8		—
173–177	12		$\Sigma(f) = 95$

2.6 PERCENTILES, DECILES, AND QUARTILES

The standard deviation about the arithmetic mean, the range, the average deviation, and the comparative magnitudes of the median and the mean (all presented earlier) provide useful information regarding the dispersion of the numerical measurements in a group of data which is being analyzed. However, there are some circumstances in which it is desirable to divide the ordered array into segments each containing a stated percentage of all the numbers in the set. More specifically, it may be convenient to partition a large body of data into four, ten, or one hundred subgroups, each containing approximately the same number of measurements from the set, and with the subgroups corresponding to successive segments of the array. The subgroups will be called *quartiles* if four divisions are employed, *deciles* if there are ten subgroups, and *percentiles* if there are one hundred subgroups.* The aim in stating the upper limit of the first quartile, for example, is to designate a number such that one-fourth the numbers in the array are less than or equal to that upper limit.

Although the upper and the lower limits of the quartiles, deciles, and percentiles could be read from a carefully drawn *r.c.f.* curve if the data are sufficiently numerous, it is desirable to have precise definitions for them. This could be done in a variety of ways, not essentially different, so that certain convenient and reasonably standard definitions will be adopted rather arbitrarily.

Before general rules and methods for determining the limits on the quartiles, deciles, and percentiles are considered attention is called to the following two arrays and to some general problems inherent in the determination of such subgroups as quartiles:

Set 1. 1, 2, 3, 4, 6, 8, 8, 9, 10, 10, 11, 12, 15, 18, 18. $N = 15$.
Set 2. 1, 2, 4, 7, 9, 9, 11, 11, 12, 15, 15, 18, 20, 24, 25, 27. $N = 16$.

Suppose that we wish to divide these sets of numbers into four subgroups, each containing equally many numbers, if possible, and coming as close as possible to equality in other instances. Two facts are

* It seems to the author that the term percentile should refer to an *interval* which includes approximately one per cent of all the measurements. However, most textbooks use this term to designate only one end point of what is called a percentile herein. Similar remarks apply to the terms decile and quartile. Since we usually speak of a score being *in* a percentile rather than *at* it, usage seems to support the point of view taken herein.

immediately clear. (*a*) When N is not a multiple of four, we cannot define four groups each containing one-fourth of N measurements; and (*b*) repetitions of numbers will pose a problem in some instances because numbers of equal size logically must be in the same subgroup, and yet to put them there sometimes will cause one subgroup to contain more than its stated proportion of all the measurements.

It will be convenient first to describe the method to be used to determine percentile limits because deciles and quartiles can be defined in terms of percentiles. The general aim in defining percentiles is to divide the ordered array into 100 subgroups, each of which contains one per cent of the numbers in the set, as nearly as this is possible. This result will be accomplished by defining the upper limit of the pth percentile to be the $\left[\frac{p}{100}(N+1)\right]$ th number in that array. For example, if $N = 1290$, as in Table 2.01, the upper limit of the ninetieth percentile is the $[\frac{90}{100}(1291)]$th, or the 1161.9th, number in the array or along its scale of measurement. Such an "ordinal" number as 1161.9 will be defined to be the number which is nine-tenths of the way between the 1161st and the 1162nd numbers from the bottom of the array. It is seen from Table 2.42 that there are 1169 numbers less than or equal to 129. With this information it is found that the 1161st and the 1162nd scores in order of size are 128 and 129, respectively. Hence, the 1161.9th number along the scale of the ACE scores is 128.9, which, then, is the upper limit of the ninetieth percentile. The lower limit of this percentile is just the upper limit of the eighty-ninth percentile. By definition, this is the [89(1291)/100]th number along the array of the ACE scores. Since $89(1291)/100 = 1148.99$, the lower limit of the ninetieth percentile is a number which is .99 of the way between the 1148th and 1149th scores from the bottom of the array. The 1148th score is 127, whereas the 1149th score is 128; hence the lower limit of the ninetieth percentile is 127.99. It follows then that the ninetieth percentile contains scores of 128 only. Actual enumeration discloses that there are 13 scores of 128, which is as close to one per cent of 1290 as is possible with integers. Such close agreement with the ideal will not be attained with most of the percentiles, especially in the neighborhood of the mean and the median, because there will be repetitions of scores which will cover more than one percentile. It may be better when much of this occurs to be content with the coarser subgroups given by deciles or even quartiles.

The upper limit of the first decile is the same as the upper limit of the tenth percentile, and similarly for the other deciles. The upper limits of the first, second, and third quartiles are the same as the upper limits of the twenty-fifth, fiftieth, and seventy-fifth percentiles. It should be clear that the median is the upper limit of the second quartile.

It is traditional to designate the upper limits of the first and third quartiles as Q_1 and Q_3, respectively, even though the term quartile may be used differently from the way they are used in this book, as was mentioned earlier in a footnote.

PROBLEMS

1. Following are the average temperatures for July in Topeka, Kansas, from 1901 to 1930, inclusive, in degrees Fahrenheit: 86.6, 77.0, 77.6, 75.0, 74.1, 74.8, 78.7, 76.0, 78.0, 79.4, 78.8, 79.9, 81.8, 80.2, 74.0, 81.9, 80.4, 78.0, 81.6, 76.8, 79.8, 76.4, 79.0, 75.2, 78.6, 79.0, 76.6, 78.3, 79.0, and 82.4. Obtain Q_1, md, and Q_3.

2. Determine and interpret the limits of the second decile for the data of problem 1. Also compute the median. *Ans.* 74.88 to 76.08; $md = 78.65$.

3. What are the limits of the third quartile of the data of problem 6 of section 2.1?

4. What are the limits of the first quartile for the fly counts given in problem 6, section 2.4? *Ans.* 0 to 8 inclusive.

5. Calculate the limits on the ninth decile for the counts of problem 1, section 2.4. What information can you derive from these limits?

6. Use Figure 2.61 on page 40 to determine the approximate sizes of Q_1, md, and Q_3 for the birth weights recorded in Table 2.61. What information about the birth weights do these numbers give? *Ans.* 66, 81, 95 grams.

7. Construct a frequency distribution table and a graph for the 4-day gains of Table 2.62 and compute the mean gain.

8. Determine the limits on the 10th percentile of the 4-day gains of Table 2.62 and interpret these numbers statistically. *Ans.* -30 to -1.7, inclusive.

9. Construct a relative cumulative frequency distribution table for the birth weights listed in Table 2.61, using the class limits indicated in Figure 2.61.

10. Suppose that a student entering college takes the following tests: a general psychological test, a reading test, a mathematics aptitude test, a social science aptitude test, and a physical science aptitude test. If his respective percentile ratings are 90, 87, 50, 92, and 63, what advice would you give him regarding a choice of a curriculum, assuming that you have faith in these tests? Explain your reasoning.

11. Determine the lower limit of the upper (tenth) decile for the ACE scores of Table 2.01.

TABLE 2.61

Birth Weights of Female (F) and Male (M) Guinea Pigs Born during a Particular Experiment

(Data obtained from H. L. Ibsen, Kansas State College.)

Jan., F	65.3	106.0	100.7	52.0	81.6	83.9	89.6	Nov., M
77.6	73.3	106.2	74.4	60.3	93.6	90.0	87.2	118.9
100.2	80.0	84.7	63.0	53.9	43.9	92.2	83.7	108.4
66.7	57.2	69.4	65.0	73.1	84.6	79.2	98.2	75.9
62.7	64.3		79.5	70.0	81.5	87.0	105.5	98.6
72.7	77.1	May, F	85.5	77.3	63.7	115.3	82.2	122.3
82.9	63.4	93.5	72.4	67.4	64.3	68.0	123.6	99.6
59.4	57.7	84.5	90.7	58.2	54.6	52.5	86.8	89.4
		99.0	85.2	75.6	57.5		80.6	79.3
Jan., M	March, M	77.5	97.3	62.2	59.3	Sept., F	90.0	106.7
87.4	94.0	77.8		57.0	62.2	81.8	69.6	95.5
97.3	80.5	76.6	June, M	41.1	74.2	57.4	109.8	105.1
97.6	84.3	91.1	112.0	66.6	61.7	70.2		119.1
107.3	73.7	57.3	44.5	41.5	64.7	53.0	Oct., F	130.4
86.7	79.1	96.7	36.1	50.7	63.2	80.7	60.9	102.7
58.9	92.6	112.4	60.3	48.6	61.6	77.4	92.9	100.6
75.3	82.1	79.8	63.0	48.0		81.2	91.2	93.5
46.1	90.0	100.3	69.1	43.3	Aug., M	80.9	96.4	134.0
	56.2	100.8	66.1	51.6	102.7	76.8	50.4	136.3
Feb., F	63.2	91.0	53.8	65.8	117.2	64.0	76.2	113.5
77.3	61.8	84.6	82.1	45.8	63.7	96.5		115.6
75.0	109.6	88.0	63.0	65.9	55.0	87.3	Oct., M	74.9
72.3	76.5	69.0	62.5	63.4	75.2	89.2	124.9	80.0
77.6	68.9	75.6	79.4	67.3	85.2	109.0	107.0	98.2
105.5	67.9	83.9	94.6	46.6	76.4	91.2	91.7	89.7
98.8	57.8	72.1	74.0	49.0	47.4	121.1	119.2	82.2
88.7	78.8	97.4	82.0	49.9	66.6	91.5	109.6	106.7
76.5	73.9	104.3	87.6	55.3	82.0	87.6	107.8	
56.7	75.5	67.0	80.5		79.6	78.8	99.6	Dec., F
90.1	68.2	64.1	53.6	Aug., F	66.5	91.0	101.9	110.6
138.4	65.0	94.6	78.0	53.3	62.9	112.2	68.1	97.9
51.4	94.2	70.3	124.5	88.8	73.6	94.7	72.0	76.8
133.4	59.8	83.2		74.3	56.9	87.8		94.2
	44.4	56.2	July, F	45.0	88.3	76.0	Nov., F	89.3
Feb., M		50.6	37.0	85.9	92.2	67.7	94.0	87.9
96.5	Apr., F		54.7	85.8	58.4	91.2	90.8	81.5
105.0	89.8	May, M	51.6	89.0	45.6	93.1	117.7	79.6
80.8	95.5	110.0	65.7	71.2	98.0	84.8	103.5	67.0
76.5	82.2	112.7	53.7	121.9	92.0	99.9	88.2	52.5
104.9	69.2	109.4	76.1	120.4	94.7		85.7	45.6
104.2	67.2	113.4	54.3	78.3	95.7	Sept., M	95.7	87.3
88.2	79.3	115.6	47.3	76.1	98.3	68.6	95.6	70.5
104.0	68.9	102.7	43.3	110.8	99.3	72.7	78.3	
105.0	75.5	110.6	42.1	91.3	72.5	53.0	105.3	Dec., M
83.1	75.2	80.3	40.8	99.9	98.4	84.0	108.3	72.3
84.2	68.1	102.6	65.8	83.8	50.1	67.1	103.7	111.6
94.5	84.5	77.8	38.7	72.4	70.2	116.2	108.3	94.7
98.0	102.6	102.5	47.3	55.9	33.5	70.3	88.0	114.0
80.0	77.6	82.6	54.6	56.5	30.0	73.2	110.3	93.6
67.2	63.9	59.7	68.9	73.8	68.4	63.2	96.9	92.2
		70.0	59.9	67.8	74.0	68.9	105.9	88.0
March, F	Apr., M	80.5	56.5	84.8	75.4	65.9	97.8	136.9
88.0	110.3	77.5	43.6	67.0	77.2	90.8	99.5	94.5
90.0	96.8	61.5		78.3	75.2	98.5	95.1	87.5
99.3	51.8	113.5	July, M	76.6	75.8	86.6	102.0	94.7
50.2	117.8	127.2	97.8	82.0	79.2	84.1	92.4	84.6
84.5	64.4		54.6	80.1	77.7	90.8	101.7	66.5
82.6	71.2	June, F	38.6	88.9	86.7	87.1	109.0	
105.2	69.8	34.9	70.8	77.8	81.2	107.5		
67.8	67.2	57.1	49.5	88.2	34.2	102.1		

TABLE 2.62

Four-Day Gains of Female (F) and of Male (M) Guinea Pigs Described in Table 2.61

(Gains in grams. Negative sign denotes loss of weight.)

Jan., F	4.8	5.7	7.3	15.9	17.3	19.1	Oct., M	0.6
	3.7	3.8	0.6	11.3	31.5	21.1	15.2	22.0	18.9
	3.8	− 1.1	− 0.7	8.1	26.0	25.8	25.6	20.4	2.9
	1.7		4.6	0.5	19.2	14.4	12.0	17.6	6.8
	2.5	Apr., F	20.4	−12.8	22.5	18.1	1.5	5.5	31.2
		− 4.2	24.8	6.7	5.5	22.0	12.1	12.5	31.2
Jan., M		− 2.8		13.0	21.3	23.5	12.9	21.2	28.7
	4.1	− 3.1	June, F	7.5	8.0	7.8	10.9	15.8	28.1
	6.9	0.9	−13.6		20.0	4.8	9.7	−30.0	9.3
	3.9	7.3	6.1	July, M	21.4	21.2	0.3	−22.1	14.7
	5.8		21.7	21.2	19.2	17.6	− 5.3	−23.3	21.4
	5.3	Apr., M	8.3	− 6.0	11.4	6.7			21.1
	0.9	9.8	14.5	− 0.6	11.8	12.6	Sept., M	Nov., F	2.8
		− 1.3	12.1	17.3	12.1	30.9	−26.3	− 0.8	11.4
Feb., F		− 1.9	15.2	8.5	12.7	13.6	4.5	− 0.4	
	4.4	− 6.1	20.4	7.3	16.6	21.7	3.7	−21.7	Dec., F
	0.2		13.5	18.6	10.8	19.9	14.8	25.6	13.6
	− 0.2	May, F	15.7	15.8	12.0	21.1	26.4	8.4	13.6
	4.5	− 3.3		13.9	8.7	18.7	10.8	−10.3	10.4
	6.3	3.6	June, M	14.9	13.6	4.8	9.5	2.6	12.0
	17.3	5.4	8.9	10.7	16.0	5.8	13.7	4.2	−10.7
	4.8	19.9	11.6	12.4	12.8	6.0	14.7	4.5	− 7.9
	10.0	10.1	15.4	12.1	13.8	17.7	7.2	24.6	− 5.4
	− 1.6	7.9	7.7	15.4	20.0	13.5	5.7	23.3	− 8.0
		9.0	20.9	12.3	16.8	16.1	18.7	8.6	−14.3
Feb., M		5.0	20.6	15.2	− 1.7	20.9	18.3	7.3	− 5.4
	−29.2	0.7	15.4	13.7	− 3.1	27.5	− 2.3	7.1	
	1.3	1.8	17.8	13.4	− 2.6	16.3	22.8	15.0	Dec., M
	4.4	10.1	7.4	14.1	− 4.8	−21.5	8.9	−16.5	13.8
	5.1	9.4	20.3	4.9	− 2.3		28.5	26.1	15.2
		12.1	8.9	12.5	13.2	Sept., F	23.4	8.6	20.7
March, F		7.2	25.4	10.6	10.4	− 8.1	13.0	11.0	18.1
	5.1	3.8	25.1	− 1.6	17.4	10.6	14.8	13.7	15.6
	13.6	3.9	20.5	8.9	19.8	−18.9	− 0.6	− 0.8	9.8
	3.3	11.2	17.8	19.5		9.2	10.1	5.7	15.8
	4.5	0.8	16.8	9.5	Aug., M	12.4	14.1	13.2	36.1
	7.7	3.2		7.5	20.6	− 9.7	16.4	16.8	3.9
	−12.3		July, F	9.4	2.8	10.8	29.4		2.1
	− 1.9	May, M	8.4		9.1	14.7	11.1	Nov., M	2.4
	0.1	20.7	15.9	Aug., F	10.1	10.4	5.2	29.5	
	5.8	20.4	14.6	11.0	11.5	12.5	6.9	26.0	
	11.4	18.5	8.0	2.4	6.5	9.2	10.7	− 3.3	
	−12.7	18.7	8.5	28.4	28.2	20.8	20.0	0.7	
		12.5	16.0	21.9	21.6	12.5		20.3	
March, M		5.3	16.8	28.5	24.3	12.5	Oct., F	9.1	
	3.5	2.5	15.0	21.5	14.5	20.9	17.6	5.8	
	13.1	7.3	4.3	16.9	19.4	16.1	25.8	0.6	
	9.7	9.7	3.8	13.0	18.4	8.9	19.7	6.2	

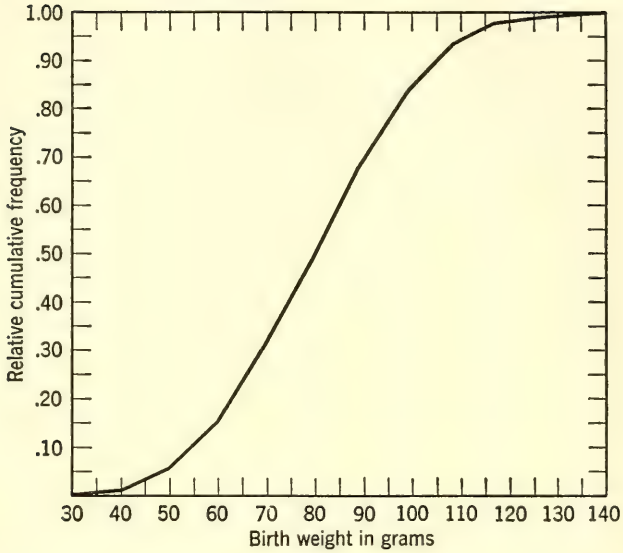


Figure 2.61. Graph of the relative cumulative frequency distribution for the data of Table 2.61.

2.7 THE COEFFICIENT OF VARIATION

There is considerable need for a measure of *relative* variability in one set of numbers as compared with another when the units of measure, or the levels of magnitude of measurement, are quite different. The standard deviation, the mean deviation, and the range are expressed in the same units as those in which the data were taken; so they obviously reflect the general size of those units.

Suppose that one bushel of a particular sort of crop weighs 60 pounds, on the average. Then the frequently used unit "pounds per 1/1000 acre" is but 0.06, or 6 per cent, of the size of the unit "bushels per acre." Hence to convert yields in pounds per 1/1000 acre into bushels per acre would require multiplication by $16\frac{2}{3}$. To see what effect such a procedure has on the standard deviation, consider the general case of two measurements X and kX, where the k is any constant. For example, X could be the number of pounds per 1/1000 acre so that $k = 16\frac{2}{3}$. By formula 2.13, $\sigma_X = \sqrt{\dfrac{\Sigma X^2 - (\Sigma X)^2/N}{N}}$ is the standard deviation of X. For the measurement, kX,

$$\sigma_{kX} = \sqrt{\frac{\Sigma(kX)^2 - (\Sigma kX)^2/N}{N}} = \sqrt{\frac{k^2\Sigma X^2 - k^2 \cdot (\Sigma X)^2/N}{N}} = k \cdot \sigma_X.$$

In other words, the standard deviation of the yields in bushels per acre is $16\frac{2}{3}$ times that of the same yields expressed as pounds per 1/1000 acre. Hence, even though σ is an excellent and widely used measure of the variability exhibited by a group of numerical measurements, its size does depend directly upon the units of measure involved, and also upon the level of magnitude of those measurements.

To illustrate the point regarding the level of magnitude of measurement, suppose that one were interested in knowing if the weights of thirty-year-old males in Manhattan, Kansas, were more (or less) variable than the weights of twelve-year-old boys in that city. Suppose also that the average weight of the men is known to be approximately twice that of the youths. The analysis just presented shows that if the boys' weights were each to be doubled so they would be on a level comparable to that of the men, their standard deviation automatically would be doubled too. It does not seem reasonable that doubling all of the X's in a set of measurements should change their fundamental variability relative to another set of measurements; hence there is need for a measure of variability which would not be so affected. The *coefficient of variation* is that sort of measure of *relative* variability.

It is easy to see that the mean of kX is k times the mean of X because $\mu_{kX} = \Sigma(kX)/N = k\Sigma(X)/N = k\mu_X$. Therefore, the ratio of the standard deviation to the arithmetic mean will be a measure of relative variability in a useful sense because

$$\frac{\sigma_{kX}}{\mu_{kX}} = \frac{k\sigma_X}{k\mu_X} = \frac{\sigma_X}{\mu_X}$$

regardless of the size of $k(\neq 0)$. It is customary to express this ratio of the standard deviation to the arithmetic mean as a percentage, and to define the *coefficient of variation* (CV) by

$$\text{CV} = 100\sigma/\mu. \tag{2.71}$$

To illustrate formula 2.71 from previously discussed data, the student can verify that, for ACE scores, CV = 27.3; for the birth weights of guinea pigs, CV = 25.2; and for problem 5, section 2.5, CV = 18.9, each as a per cent. A person acquainted with ACE scores might then observe that the scores at Kansas State College during 1947 were relatively more variable than the national scores, which (it is supposed for illustration) had CV = 20 per cent. Concerning the birth weights, we might learn that some other group of these animals has a standard deviation of only 15 grams, and hastily (and erroneously)

conclude that they were more uniform in weight than those whose weights are reported in Table 2.42. However, if the second group has a mean weight of $\mu = 50$ grams, it then is apparent that CV = 100(15/50) = 30 per cent. Hence Professor Ibsen's guinea pigs had less variability in weight at birth than the other group of guinea pigs when account is taken of the fact that they were generally heavier.

PROBLEMS

1. Compute the coefficient of variation for each of the following and draw appropriate conclusions:

X (N. Y. Curb Issues): 4, 3, 88, 1, 108, 42, 1, 25, 18, 5, 3, 6, 2, 22, and 70;
Y (Bond Quotations): 88, 115, 104, 113, 119, 80, 66, 40, 31, 101, 48, 43, 100, 84, and 15.

2. Using the X_i as -2, 5, 8, 3, 1, 0, -2, 4, 3, and 6, and using $k = 2$, demonstrate the $\sigma_{kX} = k \cdot \sigma_X$.

3. Suppose that a group of measurements of the yield of corn in a certain area of Iowa had $\mu = 70$ bushels per acre, with $\sigma = 10$ bushels, whereas an area in Kansas, growing the same variety of corn and employing the same agronomic method of culture, gave yields with $\mu = 40$ bushels per acre and $\sigma = 8$ bushels per acre. Are the yields in that part of Iowa relatively more variable than those in Kansas, less variable, or about the same, according to these data?

4. Suppose that during a certain period of years the prices of a certain commodity averaged $1.25, with standard deviation of 25 cents. The prices of this same commodity during another period averaged but 80 cents, with a standard deviation of 10 cents. During which period were the prices of this commodity relatively more stable? Give reasoning.

5. Following are simulated breaking strengths of samples of concrete (in hundreds of pounds per square inch): 40, 65, 50, 33, 48, 57, 60, 52, 50, 46, 70, 55, 51, 41, 49, 53, 56, 44, 47, 50, 46, 53, and 55. Compute the coefficient of variation.

6. In problems 7 and 8, section 2.1, for which direction finder were the readings relatively less variable? Would that result bear on the choice of one instrument over the other if such a choice were to be made?

7. Use the data of Table 2.61 to determine if the birth weights of female guinea pigs born during January and February were relatively more or less variable than those of males born during the same period.

8. Solve as in problem 7 for June and July considered together.

9. Use Table 2.62 to determine if the four-day gains of the males born during January, February, and March were relatively more or less uniform than those of females born during the same period.

10. Solve as in problem 9 for animals born during October, November, and December.

2.8 SOME OF THE PROBLEMS CREATED WHEN ONLY A SAMPLE OF A POPULATION IS AVAILABLE FOR STATISTICAL STUDY

Suppose that we wished to study ACE scores of college students but could not afford the time or the expense required to analyze *all* their scores, and hence took only a portion of them, say 50. Although an economy of time and money will be obtained, several new problems will be created.

First, how should the 50 students be chosen for the sample? Ideally, they should be representative, in all important respects, of the whole group which is being sampled. But this cannot be ascertained without studying the ACE scores of the whole group—and then no sampling would be needed. If the first 50 on an alphabetical list were to be taken, the MacIntoshes, McTaverishes, Swensons, and Swansons never would be chosen; and they might differ fundamentally from those who would be chosen. If the first 50 who came into the counseling bureau were taken as a sample, they might differ as regards ACE scores from those who came in later, or who never came into the bureau at all. In view of these and similar dangers of acquiring a biased sample from such procedures, it is necessary to devise a sampling method such that every eligible student has an equal and independent opportunity to be chosen in the sample. The net result of these requirements is to make it true that every possible sample of the chosen size (50 in the example above) will be equally likely to be drawn. This is the fundamental requirement of *random sampling.*

There are various ways to draw a random sample of 50 from among 1290 members of a population. One would be to assign each person who took the ACE test a different number, place these numbers on pieces of cardboard, and draw 50 of them at random (in the popular sense) from a bowl containing all of the pieces of cardboard. If the scores in the population are recorded in rows and columns, as in Table 2.01, we can assign numbers to the rows and to the columns, and then draw a row number and a column number at random as before. These two numbers together will uniquely designate a score for the sample. If this is done 50 times—ignoring any repeats of exactly the same row-column combination—this sample also will be a random sample because every possible set of 50 scores among the 1290 in the population will have had an equal opportunity to have been drawn.

The following random sample of 50 scores from Table 2.01 was obtained by the second method described above:

131, 66, 117, 117, 145, 71, 118, 99, 128, 111, 95, 78, 88, 55, 86, 89, 97, 98, 87, 80, 100, 76, 124, 89, 79, 101, 89, 156, 111, 98, 103, 68, 110, 76, 99, 100, 102, 61, 50, 125, 92, 106, 63, 117, 124, 87, 95, 100, 58, and 99.

These particular measurements were obtained by chance from among many possible different sets of 50. This fact suggests that the theory of probability is needed in the analysis of sampling data.

It is found in the usual manner that the mean and the standard deviation for the sample above are 96.28 and 22.55, respectively. The range of scores in this sample is 106, the median is 98, and the coefficient of variation is 23.4 per cent. It is known that these statistical measures are not likely to be exactly the same as the population parameters, but it is to be hoped that they are not far from those values.

Another sample was drawn in the same manner as the sample just described. The following were calculated for this second sample: mean = 99.22, standard deviation = 27.30, range = 144, median = 100.5, and the coefficient of variation is 27.5 per cent. It is noted that each of these statistical measures is different for the two samples, yet only the ranges differ by a large percentage. It is typical of random samples that they usually differ from each other in several respects because the particular members of such samples are in the sample by chance. It also is true that the sizes of such statistical measures as the sampling mean will follow some predictable pattern over considerable sampling experience. If this were not true, nothing much could be learned from sampling. It will be seen in later chapters that probability theory is needed to study these matters.

To illustrate the effect of the type of population on the results obtained from random sampling, consider two samples drawn from the data in problem 1 at the end of section 2.4. For convenience, samples of 10 numbers each were drawn even though this is a somewhat larger fraction of the population than was taken from Table 2.01. These samples were obtained by considering that the fly counts were numbered serially from left to right, starting with the top row. There are 148 fly counts in this population; hence a sample of 10 was drawn by effectively drawing 10 numbers at random from among the numbers 1, 2, 3, . . . , 148. When these 10 ordinal numbers were drawn, the corresponding fly counts were obtained by counting in

the manner indicated above. Following are the summaries of the two samples:

Sample 1	Sample 2
Mean = 23.2, median = 19	Mean = 13.4, median = 10
Standard deviation = 16.2, range = 45	Standard deviation = 9.7, range = 28
CV = 69.8 per cent	CV = 72.4 per cent

For these two samples, each of the five statistical measures is different again. Moreover, considering the fact that these fly counts are generally smaller numbers than the ACE scores, the relative differences are much larger between the two samples than was true for the ACE scores. For example the mean for one sample of the counts is almost twice the size of the mean of the other sample. Much the same is true of each of the other measures except the coefficient of variation. This is an illustration of the fact that a statistical study of samples requires some information about the frequency distributions of the populations sampled. Hence this matter, and probability, must be studied before more can be done about the analysis of sampling data. These are the aims of chapters 3 and 4.

REVIEW PROBLEMS

1. The effectiveness of penicillin in controlling bacterial growth can be measured by the "inhibition zone" produced when a standard amount of penicillin is properly added to a plate of agar containing the type of bacterial growth one wishes to study. Following are 54 such determinations arranged in 9 groups of 6 tests each. (From an article by Jackson W. Foster and H. Boyd Woodruff, *Journal of Bacteriology,* August, 1943.) Calculate the arithmetic mean of each set of tests, and then compute the standard deviation of these nine means.

Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9
28.1	28.5	28.0	27.5	29.0	28.0	29.0	28.5	28.0
28.0	28.0	28.2	28.0	28.0	28.0	29.0	28.2	29.0
27.5	28.0	27.5	28.1	28.3	28.0	27.3	28.2	28.6
27.8	28.0	28.0	28.5	27.3	28.0	27.0	27.5	27.7
28.0	27.5	28.0	28.0	29.0	28.1	29.0	28.0	27.5
27.5	28.0	28.0	28.2	28.0	27.5	29.0	28.0	29.0

2. Determine the range for each set of data in problem 1, and compute the standard deviation for the set with the greatest range.

Ans. 0.6, 1.0, 0.7, 1.0, 1.7, 0.6, 2.0, 1.0, 1.5; 0.88.

3. If a list of the farm acreages in a certain county in Kansas forms a statistical array of numbers from 35 to 4000; and if $\mu = 600$ and $md = 350$ acres:

(*a*) Which average would you think might be more typical of the size of farm in that county?

(*b*) Would you expect the high point of the frequency distribution to be about over the mean, to its right, or to its left? Give reasons.

4. Determine the median and also the upper limits of the first and third quartiles for the data of problem 1 when all results are considered as one group of data. *Ans.* 28.0, 27.95, 28.22.

5. Take any newspaper which gives quotations from New York bond prices and make frequency and *r.c.f.* distributions for all the closing prices as listed for that particular day.

6. Graph the distributions for problem 5. Then determine graphically the proportion of the prices which exceed 100. Check this result by actual count.

You are given the following information as a basis for working problems 7 to 11, inclusive. These data are from the Ohio Psychological Tests given to 602 students at Kansas State College during 1945. The scores are represented by the symbol X in the following summary, and are given only as integers:

Score Class	f	Score Class	f
At least 111	3	55–61	70
104–110	9	48–54	74
97–103	16	41–47	72
90– 96	21	34–40	65
83– 89	32	27–33	49
76– 82	49	20–26	12
69– 75	59	Less than 20	5
62– 68	66		—
		Total	602

For ungrouped data:
$\Sigma X = 36{,}000$
$\Sigma X^2 = 2{,}400{,}000$

7. Make the *r.c.f.* distribution and graph it.

8. Compute approximations to μ and σ^2 after changing the top and bottom classes to 111–117, and 13–19, respectively. *Ans.* 58.6, 423.0.

9. What percentage of the students had scores above 100? Between 50 and 75? What range of scores is included *between* the 80th and 90th percentiles?

10. Graph the frequency distribution and state which average you would employ to convey the best impression of the level of performance by these students on this standard test.

11. If 30 scores were to be selected at random from these 602, how many of them would you expect to fall at or below 40?

12. Given that $N = 25$, $\Sigma Y = 110$, and $\Sigma Y^2 = 600$ for any particular set of measurements, calculate the coefficient of variation. *Ans.* 48.9 per cent.

13. Given the following frequency distributions for a certain group of prices of bonds, construct graphs of these distributions. Briefly describe the sorts of information which can be obtained from such figures, and give several illustrations.

Price Class	f	*r.c.f.*	Price Class	f	*r.c.f.*
At least 121	5	1.00	85–90.99	70	.25
115–120.99	10	.99	79–84.99	30	.11
109–114.99	40	.97	73–78.99	15	.05
103–108.99	90	.89	67–72.99	7	.02
97–102.99	130	.71	Less than 67	3	.01
91– 96.99	100	.45		—	
			Total	500	

14. Calculate the median and the lower limits of the second quartile and of the 85th percentile of the data in problem 13. *Ans.* 98.14, 90.99, 107.32.

15. What percentage of the bonds of problem 13 had prices below 100?

16. Within what limits do the prices of the "middle" 60 per cent of the bonds lie (that is, excluding the lower and also the upper 20 per cent of the prices in the array)? *Ans.* 88.85 to 105.99.

17. The figures recorded below are the batting averages of all American Association players who were at bat at least 100 times, as reported by the *Kansas City Star* on July 29, 1951. The averages are arrayed from highest to lowest. Make a frequency distribution table, and compute μ and σ.

Player	Average	Player	Average	Player	Average
Walker	.389	Federoff	.295	Lerchen	.266
Cerv	.354	Thorpe	.292	Daugherty	.265
Crowe	.335	Segrist	.291	Markland	.262
Thompson	.335	Mangan	.290	Marchio	.262
Sullivan	.332	Beard	.288	Montag	.262
Wright	.323	Milne	.287	Scherbarth	.262
Richter	.322	Bollweg	.286	Atkins	.260
Katt	.322	Lyons	.284	Henley	.260
Clarkson	.322	Unser	.283	Lund	.259
Dandridge	.317	Barnacle	.282	Basso	.256
Whitman	.315	Pendleton	.280	Antonello	.256
Benson	.314	Brancato	.280	Hoak	.254
Reed	.312	Deal	.279	Marshall	.254
Mordarski	.311	Chapman	.279	Fernandez	.252
Hoderlein	.310	Marquis	.278	Kropf	.250
Mavis	.310	Tipton	.278	Turner	.250
Carey	.309	Zauchin	.277	Ruver	.248
Campbell	.309	McQuillen	.277	Conway	.247
Courtney	.309	Ozark	.276	Lu Cadello	.243
O'Brien	.307	Dallasandro	.276	Aliperto	.237
Broome (L)	.307	Stevens	.274	Natisin	.235
Saffell	.305	Thomas	.274	McAlister	.234
Kalin	.304	Cole	.274	House	.231
Mozzali	.304	Wright	.274	Teed	.230
Cassini	.301	Olmo	.270	Thomson	.223
Merson	.300	Klaus	.269	Rocco	.212
Repulski	.300	Gilbert	.268	Morgan	.179
Broome (C)	.295	De La Garza	.267	Okrie	.165

18. Referring to problem 17, write numbers from 1 to 84, inclusive, on pieces of pasteboard, mix them well, and then draw 5 at random. Consider each number drawn as the rank of a person's batting average in the above list, starting at the top. Do this 10 times, and record the range of batting averages in each set of 5 so drawn. Compare the average range with the standard deviation obtained in problem 17. If another such set of batting averages had twice as large a standard deviation, what general effect do you think this would have on the range?

REFERENCES

Dixon, Wilfrid J., and Frank J. Massey, Jr., *Introduction to Statistical Analysis,* McGraw-Hill Book Company, New York, 1951.

Freund, John E., *Modern Elementary Statistics,* Prentice-Hall, New York, 1952.

Kenney, John F., *Mathematics of Statistics,* Part I, Second Edition, D. Van Nostrand Company, New York, 1947.

Neiswanger, W. A., *Elementary Statistical Methods,* Macmillan Company, New York, 1943.

Waugh, Albert E., *Elements of Statistical Method,* Second Edition, McGraw-Hill Book Company, New York, 1943.

CHAPTER 3

Elementary Probability

Several relatively small populations of data have been studied because it is not feasible to use large groups of data in the classroom. Quite commonly, populations actually involve a very large number of numerical measurements; so large, in fact, that their number can be considered as infinite without doing appreciable violence to the subsequent analyses. Obviously, no more than a portion (sample) of the measurements in an infinite population can be obtained for study. Sampling theory requires certain probability considerations and some definite assumptions regarding the distribution of the measurements in the population (as noted in section 2.8). Hence it is appropriate to consider some of the more basic and widely used frequency distributions before attacking the problems of sampling. That is done in this and the following chapter.

Probability is involved whenever the occurrence, or non-occurrence, of any anticipated event is dependent to some degree upon chance. An "event" can be any sort of occurrence or non-occurrence which has been specified in advance. In the classroom, red and green marbles might be placed in a sack, thoroughly mixed, and one drawn out without looking into the sack. The drawing of a green marble could be considered as the event E in this instance. Likewise, if a bridge deck is thoroughly shuffled and one card drawn at random from it, the appearance of the ace of spades then might be the event E.

Another wide application of probability in everyday life lies in the determination of the premiums for life insurance policies and annuities. If a man aged 35 years purchases an annuity which will pay him \$100 per month starting at age 60 if he is alive, there are three major matters to be considered: (a) interest on the money involved, (b) the probabilities that the man will live to receive each successive payment, and (c) operating expenses and a fair profit for

the company. Whether or not such a person does live to receive a particular payment must be regarded as a chance event and, therefore, requires some use of the theory of probability. Public opinion polls regarding political matters, buyers' preferences, and foreign affairs involve chance in the selection of the persons who are to be interviewed. The reader should be able to think of many other everyday events in which the theory of probability is involved.

3.1 THE DETERMINATION OF PROBABILITIES

Before a method is presented for determining the probability that an event E will occur under specified conditions it is useful to distinguish between what will be called *single events* and *classes of events*. For the purposes of this book this distinction can be made by means of examples. Suppose that two dice are placed in a can, shaken vigorously, and rolled out upon a flat, hard surface. Many "events" can occur with each die, but just six usually are of interest: a 1, 2, 3, 4, 5, or a 6 appears on the upper face of each die when it stops rolling. How the dice were turned when they were thrown, where on the surface they came to rest, or how many turns they made while in motion are ordinarily of no interest. Moreover, it would be at least impracticable, if not impossible, to relate those phenomena to the number of dots on the upper face of a die. Hence the six possible events which will be considered herein are the appearance of a 1, 2, 3, 4, 5, or a 6 on the upper face of each die. Since these events cannot be further decomposed, we shall refer to them as *single events*. If, with each die, these faces tend to appear with equal relative frequencies over many trials, the dice are each said to be unbiased. It is with *single events occurring with equal relative frequencies* that we shall be primarily concerned in the subsequent discussion. If both dice are considered simultaneously and an event is considered to consist of a number on one die and a number on the other die, thirty-six single events are possible because any one face on the first die can appear with any of the six faces on the other die. Each possible pair of faces defines an observable event.

If attention is turned to the sum of the numbers of dots appearing on the upper faces of two dice which have been thrown simultaneously, any one of eleven different sums is possible. The different possible sums define eleven *classes of single events* (occurring with equal relative frequencies). For example, the class of events (com-

posed of such single events), "sum = 7," contains the following single events:

1 on die 1, 6 on die 2;	6 on die 1, 1 on die 2;
2 on die 1, 5 on die 2;	5 on die 1, 2 on die 2;
3 on die 1, 4 on die 2;	4 on die 1, 3 on die 2.

The class, "sum = 2," includes but one single event because there is but one way that it is possible to get a sum of 2. The class, "sum = 3," includes two single events: a 1 on die 1, a 2 on die 2; or a 2 on die 1, a 1 on die 2. The class, "sum = 4," includes three single events, etc., until all thirty-six of the possible single events have been put into one of the eleven classes of events.

We could define other classes of events among the thirty-six single events possible when two unbiased dice are tossed. For example, we could have class 1 = "sum = 7" and class 2 = "sum is not = 7." There are six single events in class 1 and thirty single events in class 2.

The preceding discussion has brought out the fact that *single events* and *classes of events* differ in one important respect. The single events are expected to occur with equal relative frequencies over many trials under the specified conditions, whereas the classes of events consist of groupings of single events, and hence would be expected to occur with relative frequencies which depend upon the numbers of single events in the classes.

Upon the basis of the preceding discussion, a useful method for determining probabilities can be devised for instances in which the single events occur with unequal relative frequencies. Suppose that under certain specified conditions any one of N possible single events can occur and that they form an exhaustive set; that is, some one of these single events must occur on any trial under the specified conditions. Assume also that the single events are grouped into s non-overlapping classes of events, with n_1 in class 1, n_2 in class 2, ..., and with n_s in class s. Then the probability that the single event which actually does occur on one future trial will belong to class i (i varies from 1 to s) is given by

$$P(E_i) = n_i/N. \tag{3.11}$$

As an illustration of the use of formula 3.11 consider the dice problem discussed above in which thirty-six single events are possible. Certain classes of events, the single events which each class includes, and the probabilities associated with each class of events are given in Table 3.11.

TABLE 3.11

SINGLE EVENTS AND CLASSES OF EVENTS INVOLVED WHEN TWO UNBIASED DICE ARE THROWN, AND THE PROBABILITIES ASSOCIATED WITH THOSE CLASSES OF EVENTS

Classes of Events	Single Events Die 1	Single Events Die 2	Number of Single Events	Probabilities for Classes
Sum = 2	1	1	1 (= n_1)	1/36
Sum = 3	1	2	2 (= n_2)	2/36
	2	1		
Sum = 4	1	3	3 (= n_3)	3/36
	3	1		
	2	2		
Sum = 5	1	4	4 (= n_4)	4/36
	4	1		
	2	3		
	3	2		
Sum = 6	1	5	5 (= n_5)	5/36
	5	1		
	2	4		
	4	2		
	3	3		
Sum = 7	1	6	6 (= n_6)	6/36
	6	1		
	2	5		
	5	2		
	3	4		
	4	3		
Sum = 8	2	6	5 (= n_7)	5/36
	6	2		
	3	5		
	5	3		
	4	4		
Sum = 9	3	6	4 (= n_8)	4/36
	6	3		
	4	5		
	5	4		
Sum = 10	4	6	3 (= n_9)	3/36
	6	4		
	5	5		
Sum = 11	5	6	2 (= n_{10})	2/36
	6	5		
Sum = 12	6	6	1 (= n_{11})	1/36
			$\Sigma n_i = N = 36$	

Other classes of single events could be defined, of course, such as the two classes: "sum ≤ 5" and "sum > 5." From Table 3.11 it is apparent that 10 of the 36 possible single events produce sums which are less than or equal to 5, whereas the remaining twenty-six single events yield sums which are greater than 5. Therefore, in this case, $N = 36$, $s = 2$, $n_1 = 10$, and $n_2 = 26$; so that $P(\text{sum} = 5) = n_2/N = 26/36 = 13/18$, or $= .72$.

Two useful facts are derivable from formula 3.11:

(3.12) $0 \leq P(E) \leq 1$ because no n_i can be larger than N; and,

(3.13) $P(E) + P(\text{not } E) = 1$ because $n_i/N + (N - n_i)/N$

$$= N/N = 1.$$

Other laws follow from formula 3.11. Two of the more important theorems will be proved and illustrated. Suppose that E_1 and E_2 denote two mutually exclusive classes of events; that is, single events in classes E_1 and E_2 cannot occur simultaneously on any one trial. Suppose also that there are n_1 and n_2 single events in classes E_1 and E_2, respectively. If a total of N single events is possible, the probability that an event in either class E_1 or class E_2 will occur on one random trial is, by definition,

(3.14) $$P(E_1 \text{ or } E_2) = (n_1 + n_2)/N = n_1/N + n_2/N$$

$$= P(E_1) + P(E_2).$$

The same reasoning and algebra are sufficient to show that for r classes of events: $E_1, E_2, \ldots, E_r$ with n_i single events in class E_i ($i = 1$ to r), the probability that some *one* of the mutually exclusive events $E_1, E_2, \ldots, E_r$ will occur on one random trial is given by

(3.15) $$P(E_1, E_2, \cdots, \text{or } E_r) = P(E_1) + P(E_2) + \cdots + P(E_r).$$

This result is known as the *Law of Total Probability for Mutually Exclusive Events.*

To illustrate formula 3.15, suppose that a sack contains 10 green, 15 red, 5 white, and 20 purple marbles, all identical save for color. What is the probability that a colored marble will be drawn on one future random trial? Let E_1 stand for the drawing of any one of the green marbles. There are 10 green marbles, and each is equally likely to be drawn; hence there are 10 single events in the class E_1. Also, let E_2 represent the drawing of any red marble, E_3 stand for the drawing of a white marble, and E_4 equal the drawing of any purple

marble. Then $n_1 = 10$, $n_2 = 15$, $n_3 = 5$, and $n_4 = 20$; hence $P(E_1,\ E_2,\ \text{or}\ E_4) = (10 + 15 + 20)/50 = 10/50 + 15/50 + 20/50 = .90$.

In the discussion leading to the Law of Total Probability the events $E_1, E_2, \ldots, E_r$ were assumed to be mutually exclusive, that is, only one of those events could occur on any one random trial. Suppose now that E_1 and E_2 are independent events in the sense that each can occur simultaneously on one trial without interfering or helping with the occurrence of the other in any way. For example, the obtaining of a 6 on one die and a 5 on another die on a single throw of the pair of dice is an illustration of independent events. If E_1 and E_2 are independent events, they can occur together in $n_1 \cdot n_2$ combinations of single events because each of the n_1 single events in E_1 can occur with each of the n_2 single events in E_2. The classes of events, E_1 and E_2, will each belong to a general class of events, which will be supposed to contain N_1 and N_2 single events, respectively. Therefore, the total number of combinations of single events possible on random trials now is $N_1 \cdot N_2$. Of those possible single events, $n_1 \cdot n_2$ will belong to both E_1 and E_2. Therefore, the probability that an even in E_1 will occur simultaneously with an event in E_2 is given by

$$P(E_1 \text{ and } E_2) = (n_1 \cdot n_2)/(N_1 \cdot N_2) = (n_1/N_1) \cdot (n_2/N_2)$$
$$= P(E_1) \cdot P(E_2).$$

As an illustration of the above discussion and results, suppose that a game consists in throwing a penny and an unbiased die simultaneously, with the thrower winning if he throws a head on the penny along with a 5 or a 6 on the die. Let E_1 represent throwing a head on the coin, and E_2 throwing a 5 or a 6 on the die. A single event now consists of a particular result on the coin plus a particular result of throwing the die. The coin can turn up either of two ways, the die any of six ways; hence there are $2(6) = 12$ combinations of events, each equally likely to occur on any one trial. In these circumstances, $n_1 = 1$, $n_2 = 2$, $N_1 = 2$, and $N_2 = 6$; hence

$$P(H \text{ and a 5 or a 6}) = \frac{(1)(2)}{(2)(6)} = (1/2)(2/6) = 1/6.$$

The reasoning and algebra above can be extended easily to prove that, if the occurrence of events in classes $E_1, E_2, \ldots,$ and E_r are independent and can occur in n_i out of N_i ways, respectively ($i = 1$ to r), the probability of the simultaneous occurrence of these r events

on one future random trial can be obtained from the following formula:

$$(3.16) \qquad P(E_1, E_2, \cdots, \text{and } E_r) = P(E_1) \cdot P(E_2) \cdots P(E_r).$$

This result is known as the *Law of Compound Probability for Independent Events.*

A similar and more general law than (3.16) can be established for situations involving dependent, rather than independent, events. Suppose that the occurrence of event E_2 depends on another event E_1 having occurred previously, or that it is useful to regard E_2's occurrence as depending on the prior happening of E_1, perhaps for the sake of convenient computation. For example, suppose that a bridge deck is to be well shuffled, and then two cards drawn successively and at random without replacing the first card drawn. What is the probability that the second card to be drawn will be an ace? It should be apparent that the answer depends somehow on the outcome of the first draw from the deck so that the second event is *dependent* upon the first event.

To attack the problem rather generally, suppose that n single events are possible under a given set of circumstances and that an event E_1 is associated with n_1 of these n single events. Assume also that an event E_2 occurs on n_{12} of the events on which E_1 also occurs. Then the probability that both E_1 and E_2 will occur on one trial is $P(E_1E_2) = n_{12}/n$, which can be rewritten in the following way:

$$(3.17) \qquad P(E_1E_2) = \frac{n_{12}}{n} = \left(\frac{n_1}{n}\right)\left(\frac{n_{12}}{n_1}\right) = P(E_1) \cdot P(E_2/E_1),$$

where $P(E_2/E_1)$ is the probability that E_2 will occur after it is known that E_1 has occurred.

The probability law expressed in (3.17) actually includes the law of (3.16) as a special case. If E_2 is independent of E_1, the number of single events on which E_2 can occur will be the same regardless of the prior occurrence or the non-occurrence of E_1; hence the probability $P(E_2/E_1)$ will be just $P(E_2)$, and the formula 3.17 becomes (3.16).

Problem 3.11. What is the probability that two successive aces will be drawn from a well-shuffled bridge deck if the first card drawn is not replaced before the second draw is made?

The number of aces available for the first draw is 4, and any of 52 cards might be drawn; hence $P(E_1) = 4/52$. On the second draw

—assuming that an ace was drawn on the first draw—there are 3 aces among 51 cards remaining in the deck. Hence, $P(E_2/E_1) = 3/51$; then, by formula 3.17, $P(A, A) = (4/52) \cdot (3/51) = 1/221$.

There are many situations in which those chance occurrences which would be considered as the single events do not occur with equal relative frequencies. For example, a coin may be biased so that the heads side turns up more frequently than the tails side. Under such conditions, we cannot assign a probability of 1/2 to the occurrence of heads (and likewise for tails) and employ the simple arguments used above. However, we can think of determining the appropriate probabilities for these single events by empirical means, that is, by many actual trials under the specified conditions. For example, we could toss the coin in question many times and then use the observed proportion of heads as an approximation to the true probability, p. Thereafter, formulas 3.15, 3.16, and 3.17 can be used.

An interesting and instructive application of the probability methods introduced above can be made to the study of human blood groups. If the red blood corpuscles of one individual's blood are mixed with the blood serum of another person (as in transfusions), one of two general results will be observed to follow: the red corpuscles will disperse evenly through the recipient's blood as though in their own serum, or they will form clumps of cells. The latter reaction is called agglutination, and it is so undesirable that there is considerable interest in preventing its occurrence. To that end, bloods are classified according to certain systems. One such system is based on the known existence of factors A and B each (or both) of which may be either present or absent from any person's blood. The following four blood groups are based on the A and B factors:

(1) type O: neither A nor B present in the blood;
(2) type A: factor A present but not factor B;
(3) type B: factor B present but not factor A;
(4) type AB: both of the factors A and B are present in the blood.

There are several interesting features about the A and B factors in blood. (1) They are inherited essentially in accord with simple Mendelian laws of inheritance, a circumstance which requires measures of probability. (2) Various racial or geographic groups tend to differ from each other in the proportions carrying the A and/or B factors, thus providing a source of some additional evidence about racial origins. (3) The mode of inheritance of the A-B groups can be used in genetic studies and in some legal problems.

Each person is considered to have in the cells of his body twenty-four pairs of chromosomes, one member of each pair having come from each parent. On these chromosomes are carried genes which are believed to govern the inheritance of various human characteristics. The genes which determine the presence or the absence of the A and B factors in the blood are carried on one of the twenty-four pairs of chromosomes. Attention here is centered solely on the chromosomes of that pair, one of which came from the father, the other from the mother. Moreover, attention is to be fixed upon one specific gene position on each such chromosome, namely, that position occupied by the genes which cause the presence or the absence of the A and B factors. If the gene at this position produces neither the A nor the B factor in the blood, it is marked (diagrammatically) as an *O* gene. Similarly there are *A* and *B* genes so that the blood type can be indicated by showing what genes the two chromosomes carry. Symbolically, there are the following four blood types:

O/O = type O; A/O or A/A = type A; B/O or B/B = type B; and

A/B = type AB.

The information presented in the preceding paragraphs makes it possible to predict the proportions of the various blood groups among the progeny of any particular combination of parents, provided that a large number of such parents and children are involved. Suppose that one parent has blood of type AB and the other has type O blood. Then the possible blood types which can occur among their offspring are as follows:

	Father		Mother
Parents	A/B	×	O/O
Genes passed on	*A* or *B*		*O* or *O*
Possible offspring	B/O B/O A/O A/O		

Of the four possible pairings of chromosomes from the father and the mother, two produce type B blood in the child because only the gene for the factor B is carried on the chromosomes. Similarly, the other two possible pairings of chromosomes produce type A blood in the children. There is no reason to doubt the usual hypothesis that each of these four possible pairings occurs the same percentage of the

time in the long run of many such matings; hence the individual pairings are considered to be single events. It follows that the probability that any specified future child of these parents will have type B blood is $P(\text{B}) = 2/4 = 1/2$. Similarly, $P(\text{A}) = 1/2$. There is no possibility that these parents will produce a child of either type O or type AB; hence $P(\text{AB}) = P(\text{O}) = 0/4 = 0$.

In view of the fact that the inheritance which a child receives from its father is an independent event with respect to its inheritance from its mother, formula 3.16 can be applied. This gives the following simple solution for the probability that a child with blood type B will be produced:

$$P(\text{B from father}) = 1/2;\ P(\text{O from mother}) = 1;\ \text{therefore, } P(\text{B from father and O from mother}) = (1/2)(1) = 1/2.$$

Since that is the only way a child with B-type blood can be produced by these parents, that is the solution to the problem.

If one parent is type O and the other type A, something must be known or assumed regarding the specific type A, that is, A/O or A/A. If one parent is A/O and the other is type O/O, $P(\text{A}) = P(\text{O}) = 1/2$. No other type is possible. But if the parent with type A blood is actually A/A, all children will be A/O.

If it is known only that the parents are of types A and O, respectively, and if it assumed that type A is equally frequently A/O and A/A, the probability that any particular future child will be type A is

$$\begin{aligned} P(\text{type A child}) &= P(\text{A/O parent}) \cdot P[\text{A child/(A/O parent)}] \\ &\quad + P(\text{A/A parent}) \cdot P[\text{A child/(A/A parent)}] \\ &= (1/2) \cdot (1/2) + (1/2) \cdot (1) = 3/4 \end{aligned}$$

by the probability laws of (3.15) and (3.17).

Table 3.12 was derived by the above methods under the assumption that type A is equally frequently A/A and A/O; and similarly for type B. (Actually this assumption is unrealistic, but it is convenient here.) The reader should verify several of the probabilities in this table, noting particularly where the assumption regarding the relative frequency of A/A and A/O among type A parents (or likewise for type B) affects the calculations.

TABLE 3.12

TYPES AND PROPORTIONS OF OFFSPRING FROM THE INDICATED MATINGS

Mother's Type	Father's Type	Types of Progeny	Probability of That Type of Progeny
O	O	O	1
	A	O	1/4
		A	3/4
	B	O	1/4
		B	3/4
	AB	A	1/2
		B	1/2
A	O	O	1/4
		A	3/4
	A	O	1/16
		A	15/16
	B	O	1/16
		A	3/16
		B	3/16
		AB	9/16
	AB	A	1/2
		B	1/8
		AB	3/8
B	O	O	1/4
		B	3/4
	A	O	1/16
		A	3/16
		B	3/16
		AB	9/16
	B	O	1/16
		B	15/16
	AB	A	1/8
		B	1/2
		AB	3/8
AB	O	A	1/2
		B	1/2
	A	A	1/2
		B	1/8
		AB	3/8
	B	A	1/8
		B	1/2
		AB	3/8
	AB	A	1/4
		B	1/4
		AB	1/2

Another method of classifying human bloods is based on the M and N factors, which are inherited independently of the A and B factors; that is, the genes for M and N are on a different pair of chromosomes from that which carries the gene for A and B. Apparently, both M and N are never both absent. There are, therefore, just three types: M, N, and MN if we ignore the subtypes which have been discovered recently. The following symbolism will be employed in the discussion of the M-N blood types:

$$\text{M/M} = \text{type M}, \quad \text{N/N} = \text{type N}, \quad \text{and} \quad \text{M/N} = \text{type MN}.$$

The inheritance of these types can be studied in the manner already established for the A-B blood groups.

In view of the fact that the three M-N types are classifications which are independent of a person's A-B type, the two blood groupings considered simultaneously make it possible to distinguish $3 \times 4 = 12$ different blood types even without bothering with the subdivisions of the A-B and M-N groups, which are serologically determinable.

Problem 3.12. If a woman's blood belongs to types O and MN, and her husband's blood is AB and N, what are the possible blood types for their first child, and what is the probability associated with each type?

The mother can pass on to her child *one* of the pairs of genes O, M or O, N because her genetic constitution as regards blood types is (O/O) (M/N). Likewise, the father can transmit either A, N or B, N to his offspring. Therefore, the possible gene combinations in the child are: O, M with A, N; O, M with B, N; O, N with A, N; and O, N with B, N. This makes the following four classifications of bloods possible for their first child: A-MN, A-N, B-MN, and B-N, each being expected to occur equally frequently. Hence, each of these four classifications has a probability of one-fourth occurrence in their first child. No other type can occur.

Two other general blood groups will be mentioned: one involves the Rh factor, the other is based on the P factor. Each is inherited independently of the other and of the A-B and M-N types. Recent discoveries of subdivisions of these groups will be ignored. Therefore, P+ and P− groups will be recognized, and also Rh+ and Rh− types. Genetically, PP and Pp are P+, and RhRh and Rhrh are Rh+, leaving pp as P− and rhrh as Rh−.

The information given previously on blood groups can be summarized as follows:

Blood Typing System

I(A-B)	II(M-N)	III(P)	IV(Rh)
O/O = group O	M/M = type M	P/P	Rh/Rh
		or	or
A/O	N/N = type N	P/p = type P+	Rh/rh = type Rh+
or			
A/A = group A	M/N = type MN	p/p = type P−	rh/rh = type Rh−
B/O			
or			
B/B = group B			
A/B = group AB			

It is seen that there are $4 \times 3 \times 2 \times 2 = 48$ mutually exclusive and serologically distinguishable blood classifications, even with the simplified groups discussed herein. By using all the known subgroups of blood types, there are many more distinguishable and mutually exclusive classifications of human bloods. The availability of these classifications has been helpful in legal cases involving disputed parentage, heirship claims when alleged maternity is doubted, identification of blood stains, genetic studies, anthropological investigations, and the identification of corpses when other methods have failed.

Apparently the chief use of blood types in legal cases occurs when one can prove the impossibility of an allegation. For example, an O-type father and an AB-type mother cannot (under the information set forth above) have an O-type child. Or, as another case, if an accused person has blood stains on his clothing and claims that they resulted from his having had a nosebleed, the finding that his blood is A, M whereas the stains are A, N would disprove his claim.

PROBLEMS

1. What is the probability that a wife with type A blood and a husband with type B blood will have three children whose blood types all are O?

2. Suppose that a husband has type A blood and that his wife's blood group is AB. What is the most likely type of blood for their first child under the assumptions made in Table 3.12? What type is impossible? *Ans.* A, O.

3. If the parents of problem 2 claim five children all of blood type B, would you doubt the blood types or, perhaps, the alleged parentage? Give probability argument.

4. If a name is to be selected at random from among those persons who were residents of the United States in 1950 and who then were between the ages of 35 and 74, inclusive, estimate the probability that the person so chosen will be aged 50 to 59, inclusive. (See problem 8, section 2.5 for the appropriate table.)
Ans. .26.

5. Compute the probability of throwing either a sum of seven or of eleven on one throw of two unbiased dice by enumerating the single events in these two classes of events. Verify your answer by applying the Law of Total Probability.

6. If five unbiased coins are to be flipped simultaneously, calculate the probability that there will be a 3:2 division of heads and tails, either way. *Ans.* 5/8.

7. Verify the probabilities given in the second and tenth lines of Table 3.12 by listing all the possible combinations of chromosomes. Where does the matter of single events come into these calculations?

8. Use the laws of Total and Compound Probabilities to solve problem 7.

9. Suppose that two bags—identical in appearance—contain, respectively, 20 red and 30 blue marbles; and 40 red and 10 blue marbles. If one bag is to be selected at random and then one marble withdrawn from that bag, what is the probability that it will be red? That it will not be red?

10. If three unbiased dice are to be thrown once, what is the probability that a sum of 4 will be thrown? A sum of at least 4? *Ans.* 1/72, 215/216.

11. If the throw described in problem 10 is to be made twice, what is the probability that a sum of 4 will be thrown both times? What is the probability that exactly one sum of 4 will be thrown on the two throws?

12. Suppose that two babies have been born almost simultaneously in a certain hospital, and that one of the families subsequently claims that the babies were interchanged either willfully or accidentally. The blood classes of the babies and of the parents are as follows:

Mr. Timoféef is A, MN, P+, and Rh+;
Mrs. Timoféef is B, N, P+, and Rh−;

Mr. Brown is B, M, P+, and Rh−;
Mrs. Brown is O, N, P−, and Rh−.

The child the Timoféefs now have is O, MN, P−, and Rh+. The child the Browns now have is O, MN, P+, and Rh−. Have the babies been interchanged? Or is it impossible to tell from this information? Give reasons.
Ans. No interchange has occurred.

13. Suppose that a few days after a wealthy man has died a woman claims that a certain girl is her daughter and that the deceased was the father. Also suppose that the following facts about blood classes have been established:

(1) The deceased's blood was B, M, Rh+, and P+.
(2) The deceased had a son whose blood was in group O and also was Rh−.
(3) The alleged mother's blood is A, MN, Rh+, and P+.
(4) The girl's blood is O, M, Rh−, and P−.

What conclusions can you draw about the paternity of the girl? Justify your statements with probability evidence based on the following assumptions: (1) For a person whose blood is B it is assumed that the chances are two out of three that the specific type is B/O if no other pertinent information is available

to change this assumption. (2) Similarly, for type A. (3) The probability that a person who tests Rh+ is specifically Rh/rh is 2/3, and similarly for P+ if there is no other information available which would change the probability.

14. Suppose that a tortoise (land turtle) is wandering at random on a 50 by 50-foot lawn enclosed by a fence. He is equally likely to be on any particular square foot of lawn one could designate in advance. What is the probability that at any specified future time he will be within 10 feet of the fence? If it is known that he is not within 10 feet of the south fence, what is the probability that he is not within 10 feet of any of the four fences? *Ans.* 16/25, 9/20.

15. Ignoring the refinements, the Rh factor is inherited as described above. The discovery of this factor in 1940 led to an explanation of one type of infant mortality, erythroblastosis. In a large majority of the cases, the father is Rh+, the mother is Rh−, and the child is Rh+. Only a fraction of the cases wherein the child is Rh+ and the mother is Rh−, which are potentially erythroblastotic, actually result in trouble; but why some do and others do not is not presently known. Obviously, the father could be either Rh/Rh or Rh/rh, but the mother must be rh/rh. Assume that the population of potential parents is divided for each sex as follows:

30 per cent RhRh, 60 per cent Rhrh, and 10 per cent rhrh

What is the expected proportion of potential erythroblastotics among their children?

3.2 PERMUTATIONS AND COMBINATIONS

Probability has been calculated in such a way that two numbers need to be determined: (*a*) the number of single events in the class of events whose probability of occurrence is being determined, and (*b*) the total number of single events possible under the prescribed conditions or in the mathematical model. For example, the probability of throwing a sum of seven with two unbiased dice is the ratio of the number of single events which give a seven to the total number of ways a sum can be produced. In this instance it is easy to determine those two numbers, but it is not usually easy. The determination of the necessary two numbers often is greatly facilitated by the use of the mathematical concepts, *permutations* and *combinations*. In the process of introducing these concepts, it is convenient to develop certain useful formulas in terms of abstract letters. Thereafter, these symbols will be employed to represent persons, heads and tails on a coin, physical objects, etc.

A set of letters, such as ABC, can differ from another set of the same number of similar marks in one, or both, of two ways: the same letters may appear in a different order, or exactly the same letters may not be present in both sets. For example, ABC and ACB are different orderings of the same three letters, whereas ABC and BCD

involve different letters. Two sets of three (or of n) letters are said to be different *permutations* of letters if they differ in either the order in which the letters are arranged, or if some different letters are involved in the two sets. Two groups of n letters are considered to be different *combinations* of letters only if some different letters are included in the two sets. A reordering of the same letters forms a new permutation but not a new combination. Hence, just one combination can be formed from n given letters if all the letters are used at once. This question then arises: How many different permutations can one form from n letters, using all n letters each time?

The process of constructing a permutation consists in determining a first, a second, ..., and finally an nth letter. The first letter is chosen from among n, the second from among the remaining $(n-1)$ letters, the third from among the $(n-2)$ then remaining unchosen, etc., until finally only one letter is left for the nth choice. These n choices can be made in $n(n-1)(n-2)(n-3)\ \ldots\ (2)(1)$ ways, which is then the number of different permutations possible with n letters if all n of them are used in each permutation. To illustrate, suppose that there are three letters: A, B, and C. The following outline shows how the choices can be made:

For first letter,	A,	or B, or	C.
For second letter,	B or C	A or C	A or B.
For third letter,	C or B	C or A	B or A.

Therefore the permutations are: ABC, ACB, BAC, BCA, CAB, CBA, and there are $3(2)(1) = 6$ of them.

It is convenient to denote the product $n(n-1)(n-2)\ \ldots\ (2)(1)$ by the symbol $n!$, and to call it n *factorial*. Hence, if $P_{n,n}$ is adopted as the symbol for the number of permutations of n marks arranged in sets of n, we have

$$P_{n,n} = n! \tag{3.21}$$

as the formula for computing the number of such permutations. If $n = 3$, as above, $P_{3,3} = 3! = 3 \cdot 2 \cdot 1 = 6$, as before.

More often, it is necessary to make up permutations of marks in which only r of the n marks are used at any one time. For example, we might be choosing a batting order of 9 men from a squad of more than 9 men. To see how the process goes, suppose that it is required to make up all the possible two-letter permutations from the letters A, B, C, and D. There are 4 choices for the first letter and 3 choices for the second; therefore, there are $4(3) = 12$ possible choices of two

from among four. In general, the symbol $P_{n,r}$ stands for the number of r-letter permutations which can be formed from n letters. Then, for $n = 4$ and $r = 2$, $P_{4,2} = 4(3) = 12$, or, in general,

$$P_{n,r} = n(n-1)(n-2)\cdots(n-r+1). \tag{3.22}$$

It is possible and useful to express $P_{n,r}$ in terms of factorials. To do so, we deliberately create the factorials from formula 3.22 by multiplying and dividing by $(n-r)(n-r-1)\ldots(2)(1)$ to make the numerator into $n!$ and the denominator into $(n-r)!$. The final result is

$$P_{n,r} = \frac{n!}{(n-r)!}. \tag{3.23}$$

From the definitions of permutations and combinations it follows that every set of r letters can be formed into but one combination, using all r of the letters; whereas, r letters can be formed into $r!$ permutations. Hence, it is concluded that there are $r!$ times as many permutations of n marks taken r at a time as there are different combinations of n letters taken r at a time. Therefore, if the symbol $C_{n,r}$ is adopted to indicate the number of possible combinations of n letters taken in groups of r letters each, the formula for that number is whichever of

$$C_{n,r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} \quad \text{or} \tag{3.24}$$

$$C_{n,r} = \frac{n!}{r!(n-r)!} \tag{3.25}$$

we wish to employ.

Problem 3.21. In how many different orders can 4 cars be parked among 6 consecutive parking places along a curb?

It should be clear in this situation that the order in which the cars are parked makes a difference because the different orders are distinguishable, and would be considered as different by a policeman checking parking. Therefore, this is a problem in numbers of permutations and can be worked by either formula 3.22 or 3.23. By formula 3.23

$$P_{6,4} = \frac{6!}{(6-4)!} = \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{2\cdot1} = 360.$$

Problem 3.22. How many different (as to cards held) 5-card poker hands are possible from the usual 52-card deck?

In view of the fact that the order in which the cards were dealt does not affect the actual cards held, this is a problem in numbers of combinations of 52 objects taken 5 at a time; hence the number of poker hands is $C_{52,\,5} = (52!)/(5!47!) = 53{,}040$ after common factors in numerator and denominator are divided out and the remaining factors multiplied together.

Problem 3.23. What is the probability that 5 cards dealt from a well-shuffled poker deck will all be spades?

Two numbers need to be determined before formula 3.11 can be applied: (1) the total number of 5-card hands which are all spades, and (2) the total number of 5-card hands of any sort which possibly could be dealt from the deck. In view of the fact that the order in which the cards were dealt is unimportant, this is a matter of finding numbers of combinations, namely, $C_{13,\,5}$ and $C_{52,\,5}$. Therefore, the required probability is

$$P(\text{all spades}) = C_{13,\,5}/C_{52,\,5} = .0005, \text{ or 1 chance in 2000.}$$

Problem 3.24. What is the probability that 5 cards dealt from a well-shuffled poker deck will include exactly 3 aces?

Three aces can be chosen from among the 4 available in $C_{4,\,3}$ or 4 ways. Likewise, $C_{48,\,2} = 1128$ is the number of different pairs of cards which do not include any aces. All possible 5-card hands with exactly 3 aces must necessarily be the same as all the possible ways to put *some* 3 aces with one of the 1128 pairs of cards which are not aces; hence there must be $4(1128) = 4512$ different 5-card hands which include exactly 3 aces. Therefore, the probability of being dealt such a hand is

$$P(\text{exactly 3 aces, 2 non-aces}) = 4512/C_{52,\,5}$$
$$= .0016, \text{ or 1 chance in 625.}$$

PROBLEMS

1. In how many ways, which differ as regards the persons in particular chairs, can 4 men and 4 women be seated around a dinner table, with men and women seated alternately?

2. Suppose that there are 10 persons in a room, and that they have the following blood types: 1 is AB, 3 are A, 2 are B, and 4 have type O blood. If 2

are chosen at random what is the probability that they will have the same type of blood? *Ans.* 2/9.

3. Suppose that a baseball team has 4 men who can bat in any of the first 3 positions, 5 who can bat in any of the fourth, fifth, and sixth positions, and 7 who can bat in any of the last three positions. How many possible batting orders are there?

4. Assume that 7 insecticides are to be tested as to their effectiveness in killing house flies. If each spray is to be tested against every other spray once in a separate test, how many tests will this require? *Ans.* 21.

5. Suppose that a housewife buys 3 cans of peaches, 6 cans of apricots, and 4 cans of pears; and suppose that her child tears off all the labels on the cans. If the housewife needs 2 cans of fruit for dinner, what is the probability that the first 2 cans chosen will contain the same kind of fruit?

6. How many 13-card bridge hands are there with no card higher than 8? *Ans.* 37,442,160.

7. If 7 unbiased coins are flipped simultaneously, how many single events are there in the class: 3 heads, 4 tails?

8. Compare the coefficients of $(x+y)^5$ with $C_{5,5}$, $C_{5,4}$, $C_{5,3}$, $C_{5,2}$, $C_{5,1}$, and $C_{5,0}$ given that $0! = 1$.

9. What is the probability that 5 cards dealt from a well-shuffled poker deck will include 3 queens and 2 aces? Three queens and at least 1 ace?

10. What is the probability that 13 cards dealt from a well-shuffled bridge deck will include exactly 8 honor cards (honor cards are 10, J, Q, K, and Ace)? *Ans.* .040.

11. In how many ways can 6 boxers be paired off for 3 bouts being held simultaneously?

3.3 REPEATED TRIALS UNDER SPECIFIED CONDITIONS

Situations involving the numbers of occurrences and non-occurrences of an event E on repeated trials under the same original conditions are of particular interest in statistical analysis. The principles involved will be seen to be important to the study of frequency distributions, and to sampling studies.

The probability problems created when trials are repeated under fixed conditions can be illustrated by means of mathematical models of these problems. Suppose that a coin is flipped n times and the number of heads noted. On such a set of repeated trials any number of heads is possible from 0 to n, that is, there are $(n+1)$ possible classes of event: 0 heads, n tails; 1 head, $(n-1)$ tails; 2 heads, $(n-2)$ tails; ... ; n heads, 0 tails. Each class of events includes some number of single events (if the coin is unbiased) from 1 to whatever the maximum size of $C_{n,r}$ is for the given n. For example,

if an unbiased coin is tossed 5 times, there are 6 classes of events, and the specific single events in the class $4H$, $1T$ are

Toss				
First	Second	Third	Fourth	Fifth
T	H	H	H	H
H	T	H	H	H
H	H	T	H	H
H	H	H	T	H
H	H	H	H	T

It is seen that $C_{5,\,4} = \dfrac{5!}{4!1!} = 5 =$ the number of single events in the class which includes exactly 4 heads. It also should be observed that the outcome of each toss is an independent event relative to the outcome of any other toss; hence the probability of the first result listed above, $THHHH$, is $(1/2)(1/2) \cdots (1/2) = (1/2)^5$. With this unbiased coin, that also is the probability for any of the other single events in this class of events. Therefore, the probability of an event in the class $4H$, $1T$ is $C_{5,\,4}(1/2)^4(1/2)^1$, in which the exponent 4 refers to the number of heads and the exponent 1 refers to the number of tails. The reader can verify the fact that for any specified number of heads from 0 to 5 the probability of exactly r heads is $C_{5,\,r}(1/2)^r \times (1/2)^{5-r}$, where r takes any value from 0 to 5.

In general, if n unbiased coins are to be flipped (or one such coin be flipped n times) the probability of the appearance of any specific number of heads, say r, is

$$P[r \text{ heads}, (n - r) \text{ tails}] = C_{n,\,r}(1/2)^r(1/2)^{n-r}. \tag{3.31}$$

To extend this result a bit, let an event E have a constant probability, p, of occurrence on each of n repeated trials. Then the probability that E will occur on exactly r of the trials [and fail on the other $(n - r)$ trials] is given by the following formula:

$$P[r\ E\text{'s}, (n - r) \text{ not-}E\text{'s}] = C_{n,\,r}(p)^r(1 - p)^{n-r}. \tag{3.32}$$

The student can verify that this formula becomes (3.31) if $p = 1/2$, $E = H$, and (not-E) $= T$.

One more generalization can be obtained regarding formulas 3.31 and 3.32 by considering the expansions of the two binomials

$(1/2 + 1/2)^n$ and $(q + p)^n$, in which $q = 1 - p$. To see these generalizations, consider the following binomials expansions:

$$\begin{aligned}(1/2 + 1/2)^2 &= 1(1/2)^0(1/2)^2 + 2(1/2)^1(1/2)^1 + 1(1/2)^2(1/2)^0,\\ &= C_{2,0}(1/2)^0(1/2)^2 + C_{2,1}(1/2)^1(1/2)^1\\ &\quad + C_{2,2}(1/2)^2(1/2)^0,\\ &= P(0H, 2T) + P(1H, 1T) + P(2H, 0T).\end{aligned}$$

That is, the successive terms of the expansion of $(1/2 + 1/2)^2$ are given by formula 3.31 if $r = 0$, 1, and 2, successively; and those three terms give the probabilities for the three possible classes of events in terms of the number of heads appearing. The generalizations for 3, 4, ..., or n tosses should be apparent. For the more general situation in which the probability of the occurrence of an event E is constantly p under repeated trials,

$$\begin{aligned}(q + p)^2 &= 1(p)^0(q)^2 + 2(p)^1(q)^1 + 1(p)^2(q)^0,\\ &= P(0\ E\text{'s}, 2\ \text{not-}E\text{'s}) + P(1\ E, 1\ \text{not-}E)\\ &\quad + P(2\ E\text{'s}, 0\ \text{not-}E\text{'s});\end{aligned}$$

and again it should be apparent that these successive terms correspond to formula 3.32 for $r = 0$, 1, and 2, successively.

PROBLEMS

1. What is the probability that if 6 unbiased pennies are tossed simultaneously, exactly 3 heads will appear?

2. What is the probability that at least 3 heads will appear under the conditions of problem 1? *Ans.* 21/32.

3. If one parent is Rhrh and AO, and the other parent is rhrh and BO, what is the probability that both their first two children will be Rh— and AB?

4. Suppose that a sample of 100 bolts is taken from a very large batch which contains exactly one-half of 1 per cent of unacceptable bolts. What is the probability that at least 2 bolts in the sample will be unacceptable? *Ans.* .09.

5. If 5 bolts among the 100 in the sample of problem 4 are found to be unacceptable products, what would you conclude about the hypothesis that only one-half of 1 per cent were faulty in the whole batch? Give reasons.

6. Write out the series for $(x + y)^4$ and show that the coefficients are numbers of combinations, $C_{4,r}$, with $r = 0$ to 4.

7. Suppose that the teams listed on a football parlay card are so handicapped that you actually have a 50-50 chance on each team you pick. What is the probability that you will pick exactly 9 winners out of 10? Would this probability justify odds of 25 to 1 for this accomplishment? What about odds of 250 to 1 for getting 10 out of 10 correct?

8. If a pair of unbiased dice is to be thrown 6 times in succession, what is the probability that exactly 3 sevens will be thrown? What would you think the most likely number of sevens would be? *Ans.* .054.

9. If a certain manufacturing process is producing machine parts of which 10 per cent have some serious defect, what is the probability that all of the 10 parts chosen at random will be acceptable (that is, have no serious defect)? How many would you have to take in the sample before the probability of all being acceptable will be no greater than .05?

10. Graph $f(p) = (1 - p)^{10}$, and relate this graph to problems like problem 9.

3.4 MATHEMATICAL EXPECTATION

The discussions earlier in this chapter have involved the occurrences of chance events as a result of what have been termed "trials" under specified conditions. The outcome of a trial is described in one of two general ways: (*a*) Something happens a certain number of times on a specified *number* of trials, or (*b*) we simply note whether or not an event E has, or has not, occurred and associate with that occurrence some value, say a financial loss, as in insurance. With either type of situation it may be important to be able to predict what will be the average outcome of trials under the stated conditions, over the long run of experience. For example, an insurance company needs to know what amounts it should expect to have to pay out in death benefits during a particular period of time, one year, for instance.

In case *a*, the prediction needed is to be presented in the form of an *expected number* of occurrences of an event E on a set of n future trials. A formula for this expected number can be justified heuristically as follows. If the probability of E is p, the p is just the fraction of the time that E should occur over many trials. Hence, if there are to be n trials, it is reasonable to say that the expected number of occurrences of E on n trials is

$$\text{Expected number} = E(r) = p \cdot (n). \tag{3.41}$$

Problem 3.41. If 6 unbiased coins are to be tossed simultaneously, what is the expected number of heads?

In this circumstance $p = 1/2$ and $n = 6$; hence the expected (or long-run average) number of heads is $E(r) = (1/2)(6) = 3$. Actually, our intuition would lead to the same conclusion.

Problem 3.42. Suppose that an insurance company has insured 50,000 persons who are each 30 years old, and that records from past experience show that 6/1000 of such persons die before reaching the age of 31. What is the expected number of deaths during the first year of the insurance contract?

For this situation $p = .006$, $n = 50{,}000$; therefore the mathematically expected number of death benefits among those thirty-year-olds is $E(r) = .006(50{,}000) = 300$. The reader will realize that it would be unsound financially for the company to be prepared to pay only 300 death benefits because *this* time the number of deaths might be considerably higher. All that is being said is that over a period of years of such calculations the average number of deaths among thirty-year-olds in this same insurance class will be very close to 300.

When chance occurrences are of the type b described above there may be associated with the occurrence of E some value, say a financial gain or loss. Then we may wish to predict the loss or gain to be expected on the average under the given conditions. For example, suppose that you are going to roll a pair of unbiased dice and are to be paid 60 cents if you get a sum of 7. How much should you pay to play such a game if you just wish to break even? Obviously you will receive either 60 cents or zero cents after each game; but over many games what will be your average winnings per game? That is the amount you can pay and break even. Because the probability of throwing a sum of 7 is 1/6 you expect, mathematically, to win 60 cents on about one-sixth of your throws and to win zero cents on the other five-sixths of the throws. Hence, the mathematical expectation logically is

$$(1/6)\cdot(60 \text{ cents}) + (5/6)\cdot(0 \text{ cents}) = 10 \text{ cents}.$$

Therefore you can expect to break even in the long run if you pay 10 cents to play each game.

The game just described can be extended to include a reward of 90 cents if you throw a sum of 11 on the two dice. In this circumstance you can win in either of two mutually exclusive ways, that is, you can throw a 7 or an 11. Therefore attention is centered on three classes of events and the corresponding rewards:

A sum of 7 with a reward of 60 cents,
a sum of 11 with a reward of 90 cents, and
a sum other than 7 and 11 with a reward of 0 cents.

Therefore, over a large number of games you will tend to win 60 cents on one-sixth of the throws, 90 cents on one-eighteenth of the throws, and 0 cents on the other seven-ninths of the throws. Hence your mathematical expectation on this game is $(1/6)(60 \text{ cents}) + (1/18)(90 \text{ cents}) + (7/9)(0 \text{ cents}) = 15$ cents, because that is the

average winnings per game and is, therefore, the amount you could pay to play this game and expect to break even.

The preceding ideas and methods can be generalized and symbolized in the following manner. Let all the single events possible under a specified set of conditions be grouped into s mutually exclusive classes of events. Let x_i be the reward, loss, or in general the "value" of the occurrence of an event in the ith class; and let p_i be the probability that an event in the ith class will occur on any designated future trial under the stipulated conditions. Finally, let $E(x)$ stand for the total mathematical expectation under the given conditions. It follows from the reasoning outlined above that

$$E(x) = p_1x_1 + p_2x_2 + \cdots + p_sx_s = \sum_{i=1}^{s} (p_ix_i). \tag{3.42}$$

PROBLEMS

1. Suppose that you are 20 years of age and that you are to inherit \$10,000 at the age of 30 if you are alive then. What is the expected value of this inheritance if you have a probability of .92 of living to be 30 years of age? (This probability is derived from the American Experience Mortality Table.)

2. It is approximately true that brown and blue eye colors are inherited in a manner similar to that explained for the A-B blood groups. If b/b = blue eye color and either B/b or B/B = brown eye color, what is the expected number of blue-eyed children among 500 from parents who are B/b and b/b, respectively? *Ans.* 250.

3. Answer the same question as in problem 2 for parents who are both B/b.

4. If in each three-month period 1 car in 20 of the type which you drive has an accident costing an average of \$75 for repairs, how much insurance against such a loss should you pay each quarter if you allow the company 15 per cent beyond mathematical expectation for handling the business, and if you ignore interest on your money? *Ans.* \$4.31.

5. How much would one be justified mathematically in wagering against one dollar that on 10 throws of two unbiased dice a sum of 7 will appear less than 3 times?

6. Suppose that you have the choice of receiving \$10,000 at age 65 if you are alive then, or of taking a cash payment now. From a purely mathematical point of view and ignoring interest on money, what should the size of the payment be if your probability of living to be 65 is .56? *Ans.* \$5600.

7. Suppose that a concession at a fair offers a 50-cent prize if you pay 10 cents for 3 throws and knock down all of a stack of milk bottles on the 3 throws. Suppose also that you have 1 chance in 10 to knock down the bottles. If the operator of the concession has to pay \$75 per day for the privilege of doing business there, how many customers must he have per day in order that he can expect (mathematically) to make some money?

8. Suppose that a person who is 40 years of age is to receive \$1000 on each of his sixtieth, sixty-first, and sixty-second birthdays if he is alive to receive them. Also suppose that interest on money is to be ignored. Given that his proba-

bilities of living to those successive ages are .74, .72, and .70, what should this person pay for an annuity of this sort if the company is allowed 15 per cent for overhead? *Ans.* $2484.

REVIEW PROBLEMS

1. What is the difference between a population of numerical measurements and a sample of such measurements?

2. What was Political Arithmetic? With what sorts of problems were the political arithmeticians mainly concerned?

3. Who was Student and how was his work connected with present types of statistical problems?

4. Expand $(1/3 + 2/3)^4$ into a series and state specifically what probability is given by each term if $p = 2/3$ is the probability that a certain loaded penny will turn up heads on any particular future throw. Describe the mathematical procedure needed here to define the single events.

5. What is a frequency distribution? A relative cumulative frequency distribution?

6. Suppose that a college freshman has earned the following percentile ratings on the indicated tests: (*a*) general intelligence, 90; (*b*) achievement in social sciences, 65; (*c*) achievement in physical sciences, 92; (*d*) achievement in mathematics, 95. What can you say about the student's probable future success in courses in chemistry, physics, mathematics, history, and sociology if it is assumed that the tests are trustworthy and if no serious personal problems interfere?

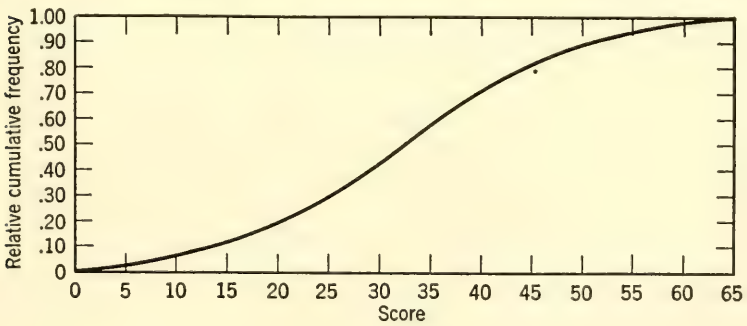
7. Given that for a set of numerical measurements $X_1, X_2, \ldots, X_{50}$, $\Sigma X = 95$ and $\Sigma(x^2) = 2.06$, calculate the coefficient of variation.

8. Calculate the geometric, arithmetic, and harmonic means of 1/2, 2, and 8 and discuss the choice of the best average for these numbers. *Ans.* 2, 3.5, 8/7.

9. Suppose that the following probabilities regarding football games have been determined reliably: A to beat B, $p_1 = 2/3$; C to beat D, $p_2 = 1/2$; and E to beat F, $p_3 = 5/6$. What are the odds that A, D, and E all win?

10. Given that the graph below is the *r.c.f.* curve for a certain group of scores, determine from it the median, the upper limit of the third quartile, and the upper limit of the sixth decile. Also interpret these results statistically, with some indication of the uses to which such information can be put.

Ans. 31.5, 41.0, 35.0.



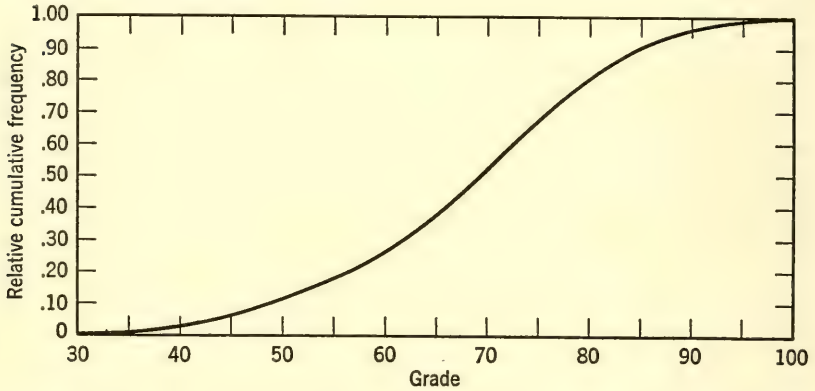
11. What proportion of the scores summarized in problem 10 lay between 40 and 60 inclusive? What proportion exceeded 60?

12. What is the probability that on 15 flips of an unbiased penny one will get either 7 or 8 heads? Will get neither 7 or 8 heads? *Ans.* .39, .61.

13. Suppose that for a certain strain of chickens the probability that a late-feathering chick will be hatched from any egg selected at random is 1/16. What is the expected number of such late-feathering chicks among 800 newly hatched chicks?

14. If a pair of true dice is rolled 60 times, what is the mathematically expected number of sevens? Of either sevens or elevens? Of sums greater than 9? *Ans.* 10, 13⅓, 10.

15. Assume that the semester grades in a large chemistry class have the ogive graphed below. If the letter grades are to be distributed as follows: 7 per cent A, 20 per cent B, 46 per cent C, 20 per cent D, and 7 per cent F, what are the grade ranges covered by each letter grade?



16. What is the median numerical grade for the data of problem 15 above? What are the upper limits of the quartiles? *Ans.* 69; 59.5, 69, 77, 100.

17. Suppose that 6 unbiased pennies are to be tossed simultaneously. What is the probability that *no more* than 2 will show heads? That *at least* 2 will turn up heads?

18. Assume that the true odds on each of 3 horses to win a particular race are determined to be as follows: horse A, 3:2; horse B, 1:3; and horse C, 1:9. What is each horse's probability of winning? What is the probability that some one of these 3 horses will win? *Ans.* .60, .25, .10; .95.

19. Given that for three separate statistical populations of data: $\mu_1 = 25$, $\sigma_1 = 4$; $\mu_2 = 50$, $\sigma_2 = 5$; and $\mu_3 = 100$, $\sigma_3 = 13$. Which group of data would you consider as relatively the more variable? Give specific statistical evidence to back your answer.

20. Compute the mean deviation and the standard deviation for the following data: 13, 9, 10, 17, 15, 20, 11, 5, 2, 10, 14, 13, 19, 21, 16, 8, 14, 6, 3, 29, 16, 17, 15, 15, 18, and 2. Which measure of variation do you think best describes the dispersion of these data about their arithmetic mean? Give reasons. You are given that $\Sigma X = 338$, $\Sigma X^2 = 5406$. *Ans.* 4.92, 6.24.

21. According to Figure 2.41, in which quartile would you place a score of 85? In which decile would this score fall?

22. If all students whose ACE scores (Table 2.01) fell among the lower 15 per cent, approximately, of all scores made were to be advised to consider seriously dropping out of college, what would be the highest score whose recipient would receive such advice? *Ans.* 67.

23. From Figure 2.41 determine approximately the percentage of those scores which were not more than one times the standard deviation either greater than or less than the mean, μ.

24. In the game of "craps" two unbiased dice are thrown successively by the same person. He wins if: (*a*) he throws a sum of either 7 or 11 on his first throw; or (*b*) he throws a 4, 5, 6, 8, 9, or 10 on the first throw and repeats his number on a subsequent throw before he throws a 7. What is his probability of winning within 2 throws? *Ans.* 97/324.

25. In a certain gambling game you are paid 15 for 1 if you throw a 1 and a 2 (either order) on two unbiased dice. On 1800 games on each of which the player pays one dollar, what is the expected percentage profit for the house relative to the amount taken in?

REFERENCES

Arley, Niels, and K. Rander Buch, *Introduction to the Theory of Probability and Statistics,* John Wiley and Sons, New York, 1950.

Kenney, John F., *Mathematics of Statistics,* Part II, Second Edition, D. Van Nostrand Company, New York, 1947.

Levy, H., and L. Roth, *Elements of Probability,* Oxford University Press, London, 1936.

CHAPTER 4

The Binomial and Normal Frequency Distributions

The discussions and illustrations of Chapter 2 involved situations in which groups of measurements (usually numerous) had been taken under specified conditions, and we had in mind only an efficient summarization of the data. The ACE scores in Table 2.01 were cited as an example. In a sense, we simply took what we got and thereafter applied statistical methods to reduce a bulk of data to a more comprehensible form without losing any essential information. More generally, however, populations of numerical measurements must be studied by means of samples because so many measurements are involved that it is not feasible, efficient, or even possible to obtain and to analyze the whole of the population.

Two different types of populations will be considered. In one type, the chance variable will be a qualitative one such as male or female, dead or alive, own an automobile or do not own an automobile. The population will consist of individual members, each falling into one of just two classes according to the qualitative designation adopted. The other type of population to be considered will be based upon a variable which is measured along a continuous scale, such as the weight of an individual, the volume of a gas, or the bushel yield of a variety of wheat.

As regards populations in which a qualitative variable is used, attention herein will be confined to what is called a *binomial population* because each member of the population falls into one of only two classes. The proportion of a binomial population which belongs to one of the two classes will be measured by the fraction p, leaving the fraction falling into the other class to be $1 - p = q$. For example, if all the babies born in New York City during a given year were to be classified as male or female, p might be the fraction who were males. If $p = .51$, then $q = 1 - .51 = .49$. The sex would be the qualitative variable mentioned above, and has but two "values":

male and female. If a baby were to be chosen at random from among those born in the specified year, its classification as male or female would be a member of this binomial population. Under the above assumptions, the probability that such a selection will turn out to be male is $p = .51$.

If n repeated observations, or trials, are made on a binomial population in which the proportion p is staying fixed, and if attention is fixed upon the *number* of individuals in each of the two classes, these numbers are variable from one set of n trials to another. For example, it was noted in Chapter 3 that the probability that r males, say, and $n - r$ females would be observed is given by $C_{n,r}(p)^r(1 - p)^{n-r}$. In other words, r is a chance variable. The relative frequencies with which r will have the values $0, 1, 2, \ldots$, and n after a great many sets of n random trials from a binomial population constitutes a *binomial frequency distribution.* This distribution will be of more direct interest to us than the binomial population in itself because the binomial frequency distribution describes results which are obtained in the process of sampling a binomial population.

There are many types of populations for which the random variable is of the second type discussed at the beginning of this chapter, namely, a measurement referred to a continuous scale, such as weight. Probably, the most important populations of this sort are those called *normal populations.* It will be convenient to describe this type of population by means of a mathematical formula for its frequency distribution. This will be done in a later section.

It seems obvious that we cannot possibly learn much by sampling a population which cannot be clearly and concisely described; hence there is need for a mathematical description, or classification, of populations. We choose to study types of populations by means of their frequency distributions because that—or something equivalent —constitutes the fullest description we can obtain for a particular population. As noted above, the discussion in this chapter will be devoted to two of the most important types of frequency distributions: one, *the binomial,* is appropriate to qualitative measurements of a certain kind; the other, *the normal,* typifies continuous numerical measurements of types quite frequently met in practice. Between these two theoretical distributions, a great many of the uses of statistical analysis will be introduced.

4.1 THE BINOMIAL FREQUENCY DISTRIBUTION

As was stated above, a binomial population and the corresponding binomial frequency distribution are involved when every single event which can occur under prescribed conditions must belong to one of two classifications. This fact corresponds to the meaning of the prefix *bi-* in the word *bi*nomial. For example, if you take out a term insurance policy for a period of 10 years you and the company are interested in your subsequent classification as "dead" or "alive" before, or at, the end of the 10 years. Of course, the company insures many persons and regards them as a group, some of whom will be classifiable as "dead" and the remainder as "alive" at the expiration of the 10-year term. What the insurance company and its clients need to know, then, is this: Given a group of n persons insured for a 10-year term, what are the probabilities associated with each of the possible numbers of "dead" and "alive" insured persons during the 10-year period of the insurance contract? For any specific n the relative frequency—over a great deal of experience—of the occurrence of 0, 1, 2, 3, . . . , $(n - 1)$, or all n "dead" after 10 years will be the *binomial frequency distribution* mentioned above. It is upon the basis of this sort of information that insurance premiums are calculated.

Suppose, for simplicity, that a company has insured 10 persons who are 30 years of age for a 10-year term. What can the company expect to pay out in death benefits? It is obvious that at the end of the 10-year period any one of 11 events may have occurred. There can be 0, 1, 2, . . . , or all 10 classified as "dead." Also, over the experience of many such groups of contracts for 10-year periods those 11 possible outcomes will occur with unequal relative frequencies which depend both on the number, such as $n = 10$, and on the probability of death for persons in this age interval. Clearly, this binomial frequency distribution depends on n and on p = probability of death between the ages of 30 and 40 years.

No one can state theoretically what the probability of death is for any particular person during the age period of 30 to 40 years; but tables have been compiled from experience which give the best available estimate of the desired probability. For example, the *American Experience Table of Mortality* indicates that the average probability is approximately one-tenth that an insurable person (determined by examination before the company will insure) now 30 years of age will die before he is 40 years of age.

If 10 persons are insured under conditions to which the *American Experience Mortality Table* applies, it follows from the discussion of section 3.3 that the 11 corresponding probabilities of occurrence of these numbers of deaths are given by the successive terms of the following binomial series:

$$(.9 + .1)^{10} = (.9)^{10} + 10(.9)^9(.1)^1 + 45(.9)^8(.1)^2 + \cdots + (.1)^{10}$$

It is not necessary to devise some game with $p = .1$ and discover from experience that a fraction $(.9)^{10}$ of the trials will show no occurrences of the event E because the only interest is in the relative number of

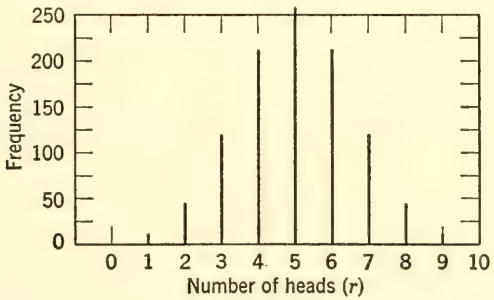


Figure 4.11. Graph of the binomial frequency distribution for $p = 1/2$ and $n = 10$.

occurrences, and that is what the probability gives. Hence, the above series gives the frequency distribution of the eleven possible classes of events.

To re-illustrate the discussion of the preceding paragraphs with an example which the reader can reproduce easily and, in addition, to show how to graph a binomial frequency distribution, attention again is called to a mathematical model. Suppose that an unbiased coin is to be flipped 10 times and the number of heads is to be recorded after each set of 10 throws. In these circumstances, $n = 10$, $p = 1/2$, and $q = 1 - p = 1/2$; hence the successive terms of the following binomial series give the probabilities for 0, 1, 2, 3, . . . , or 10 heads on any future set of 10 throws: $(1/2)^{10} + 10(1/2)^9(1/2)^1 + 45(1/2)^8 (1/2)^2 + \cdots + (1/2)^{10}$; or $1/1024 + 10/1024 + 45/1024 + \cdots + 1/1024$. In view of the fact that each of the denominators is 1024, we obtain a useful and simpler expression for the relative frequency of occurrence of 0, 1, 2, . . . , or 10 heads on 10 throws, by using only the numerators. From them a graph can be constructed to depict the relative frequency for each possibility, as is done in Figure 4.11.

This figure also can be described as the graph of the binomial frequency distribution when $n = 10$ and $p = 1/2$.

It is apparent that the actual form of a binomial frequency distribution depends upon two numbers, n and p. If $p = 1/2 = 1 - p$,

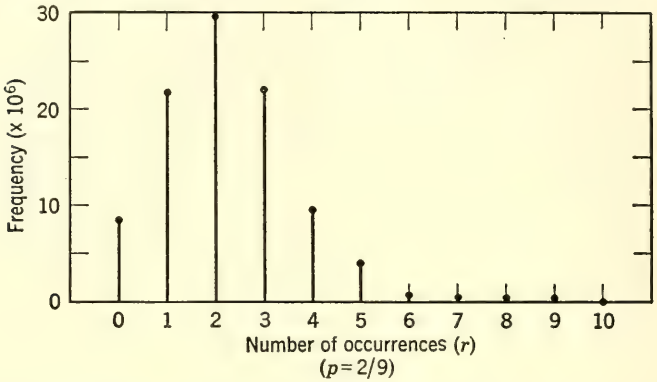


Figure 4.12*A*. Graph of the binomial frequency distribution with $p = 2/9$ and $n = 10$.

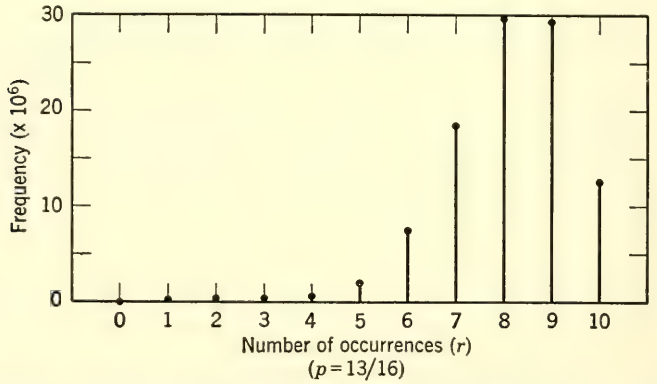


Figure 4.12*B*. Graph of the binomial frequency distribution with $p = 13/16$ and $n = 10$.

the graph is symmetrical, as in Figure 4.11. If $p > 1/2$, the event E is more likely to occur than to fail to occur; hence the higher ordinates of the graph will be toward the right-hand side of the graph. If $p < 1/2$, the reverse situation is expected. These remarks are illustrated in Figures 4.12*A* and *B*. For Figure 4.12*A*, $p = 2/9$; and for Figure 4.12*B*, $p = 13/16$. In both cases $n = 10$. The series for the binomials $(7/9 + 2/9)^{10}$ and $(3/16 + 13/16)^{10}$ were employed in the

constructions of these figures, using only the numerators of the terms as explained above.

The *r.c.f.* distribution for a binomial situation is discontinuous—as is expected—and involves successive ordinates, each at least as large as the preceding one to its left on the graph. Such a graph is shown in Figure 4.13. If we were to draw a smooth curve through the tops of the ordinates, it would have the same general appearance as the *r.c.f.* curves drawn in Chapter 2.

Fundamentally the frequency and *r.c.f.* tables corresponding to Figures 4.11 and 4.13 are as shown in Table 4.11. The meaning and

TABLE 4.11

FREQUENCY AND *r.c.f.* DISTRIBUTIONS FOR THE BINOMIAL DISTRIBUTION DEFINED BY $p = q = 1/2$, $n = 10$. TOTAL FREQUENCY TAKEN = 1024, THE SUM OF THE NUMERATORS OF THE SERIES FOR $(1/2 + 1/2)^{10}$

Number of Occurrences of E r	f	*c.f.*	*r.c.f.*
10	1	1024	1.000
9	10	1023	.999
8	45	1013	.989
7	120	968	.945
6	210	848	.828
5	252	638	.623
4	210	386	.377
3	120	176	.172
2	45	56	.055
1	10	11	.011
0	1	1	.001
	$\Sigma(f) = 1024$		

use of Table 4.11 are fundamentally the same as for similar tables in Chapter 2, but some differences should be noted. The major difference arises from the fact that the class "intervals" now are just isolated points on a scale of measurement appropriate to r. For example, 5½ per cent (0.055) of the observed values of r (over a very large number of observations on r) will be at or below $r = 2$. However, these observed numbers of occurrences of E will be 2's, 1's, and 0's *only:* there is no such r as 1.6, for example. Another difference between Table 4.11 and similar tables in Chapter 2 is that the former is a theoretical table which fits any situation for which $p = 1/2$ and $n = 10$. The frequency tables in Chapter 2 were relevant only to the

particular situation which produced the data summarized in a given table.

We might wish to know what the median r is for a binomial distribution. By Table 4.11, 37.7 per cent of the numbers are seen to be 0's, 1's, 2's, 3's, and 4's. If 5's are included, the percentage runs past 50 (needed for the median) to 62.3; therefore the median r must be 5. It is not some decimal fraction between 4 and 5 because no such numbers even exist on the scale of measurement of r.

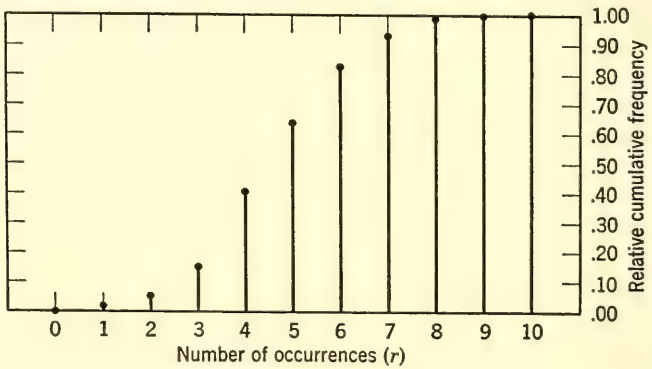


Figure 4.13. The *r.c.f.* distribution for the binomial distribution with $p = 1/2$ and $n = 10$.

The median of the binomial distribution just considered also can be obtained from the *r.c.f.* distribution of Figure 4.13 by reading horizontally from the point where *r.c.f.* = .50 until we come to the first ordinate on the left, which is high enough to be intersected by the horizontal line from *r.c.f.* = .50.

It is interesting to compare a frequency distribution which was obtained by actual trials with that which would be expected mathematically under the specified conditions. This is done approximately in Table 4.12 for a situation in which 5 pennies were flipped 2000 times. It was assumed that the pennies were unbiased, although this is known not to be strictly true for any actual coin. It should be apparent from previous discussions that the mathematically expected proportions of the 6 possible combinations of heads and tails listed in column 1 of Table 4.12 are 1:5:10:10:5:1. The resulting expected numbers of occurrences of each of the possibilities are given to the nearest whole number under the heading "Exp." in columns 3, 5, 7, and 9.

TABLE 4.12

Comparison of Observed and Expected Frequencies of Heads (H) and Tails (T) When 5 Pennies (Assumed Unbiased) Are Flipped 100, 500, 1000, and 2000 Times

Combination of H and T r, $(n - r)$	100 Throws		500 Throws		1000 Throws		2000 Throws	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
$5H$, $0T$	4	3	16	16	33	31	63	62 *
$4H$, $1T$	10	16	71	78	135	156	293	312
$3H$, $2T$	33	31	151	156	287	313	600	625
$2H$, $3T$	31	31	163	156	342	313	641	625
$1H$, $4T$	18	16	86	78	168	156	332	312
$0H$, $5T$	3	3	13	16	35	31	71	63 *

* Actually each of these numbers is 62.5 but was rounded off this way to keep the sum of the observed and expected frequencies equal.

After the 2000 trials involving 10,000 tosses the ratio of heads to tails is 0.94 to 1. Hence there apparently is a weak but definite tendency for tails to appear more frequently than heads; that is, p is not exactly equal to 1/2. Methods will be described in Chapter 5 for deciding when a coin, say, is biased, and for estimating the degree of bias.

If the observed frequencies in any column of Table 4.12 are taken as the f and the r is listed merely as 5, 4, 3, 2, 1, and 0, we have an observed frequency distribution, as in Chapter 2. If the expected frequencies (which follow a mathematical law) are used as the f column and r again is listed as 5, 4, 3, 2, 1, and 0, we have a theoretical frequency distribution of the sort being discussed in this chapter.

In view of the existence of a general mathematical expression for the binomial frequency distribution (as in formula 3.32), we might be curious to know if such statistical measures as the arithmetic mean and the standard deviation can be determined just from the n and p which determine the distribution. This is, in fact, true, as will be shown partially below.

The discussion of mathematical expectation given in Chapter 3 included the information that the arithmetic mean of the number of occurrences of an event E over many trials coincides with the ex-

pected number, $E(r)$, for any designated future trial. Hence it already has been found from experience and intuition that for the binomial situation $\mu = np$. This result can be established for any binomial frequency distribution, but such will not be done herein.

It also can be shown by somewhat more difficult mathematics that the standard deviation for a binomial frequency distribution is given by $\sigma = \sqrt{npq}$; consequently, given n and p, we can compute the mean and standard deviation very easily. This will be found to be helpful later in this chapter.

Sometimes when dealing with binomial distributions it is advantageous to work with the fractional number of occurrences, r/n, rather than with the actual number, r. In this case, the arithmetic mean is p, rather than np; and the standard deviation is $\sqrt{pq/n}$, instead of $\sqrt{npq}$. To illustrate the use of these formulas both for r and for r/n consider again the insurance example above in which $n = 10$ and $p = .1$. Under these circumstances the mean r is $np = 10(.1) = 1$, and the standard deviation is $\sqrt{npq} = \sqrt{10(.1)(.9)} = 0.949$, approximately. The mean r/n is $p = .1$, and the standard deviation of the fraction dead is $\sqrt{pq/n} = \sqrt{(.1)(.9)/10} = 0.095$, approximately. If the number n were sufficiently large that it would be practical, such information as that just derived might be useful to an insurance company in anticipating the average number (or fraction) of death benefits it could expect to pay, and in making sufficient allowance for chance deviations from those average numbers so that adequate funds would be available to pay death benefits.

PROBLEMS

1. Use the coefficients of $(1/2)^r$ in the series for $(1/2 + 1/2)^6$ to graph the binomial frequency distribution appropriate to sets of 6 trials with an event whose probability of occurrence is constantly $p = 1/2$.

2. Under the conditions of problem 1, what is the probability that the event will occur at least 4 times on 6 trials. *Ans.* 11/32.

3. Graph the frequency distributions for the binomial with $p = 1/2$ and $n = 4$, 8, and 12, successively. Compute the μ and σ in each instance and locate on the scale of r: $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$.

4. Graph the binomial frequency distribution for $p = 1/4$, $n = 4$, and read from it the probability that r will be 2, 3, or 4.

5. Check the result obtained in problem 4 by constructing the *r.c.f.* graph and reading the answer from this graph.

6. Graph the binomial frequency distribution for $p = .7$, $n = 4$, and determine the probability that on one random set of 4 trials E will occur at least μ times, where μ = arithmetic mean.

7. Flip 3 pennies 80 times, recording the number of heads after each toss of the 3 pennies. Then compare the observed and the mathematically expected numbers of occurrences in each of the four possible classes of events in terms of number of heads.

8. Perform the operations of problem 7, except to compare the observed and the theoretical values of the arithmetic mean.

9. Suppose that a large group of fruit flies consists of members who are classified as either "normal" or "sooty." Among 10 of these flies selected at random, 3 were found to be "sooty." How frequently would that result be obtained if half the flies in the population are "sooty"? How frequently if 25 per cent are "sooty"?

10. Suppose that under the conditions of problem 9, 100 flies are chosen at random and 30 (same percentage as in problem 9) are "sooty." Answer the same questions as in problem 9.

Ans. 23 times in 1,000,000; 12 times in 1 billion.

11. Construct the *r.c.f.* distributions for problem 3 and then determine the median r in each distribution.

12. Suppose that there are two political parties interested in a certain college election, and that 60 per cent of the eligible voters are Progressives and 40 per cent are Independents. If a random sample of 10 persons is taken, what is the probability that a plurality of them will be Independents, even though they constitute the minority party? *Ans.* .17.

13. Referring to problem 12, how large must the sample be before the probability is less than 1/4 that there will be more Independents than Progressives in the sample?

14. Suppose that 6 persons out of 10 selected at random in a certain city favor a particular flood-control policy. What is the probability of such a result when only 45 per cent of those in that city actually favor that policy? *Ans.* .16

15. For problem 9 determine the median number of "sooty" flies among 10. What is the probability that the actual number observed will exceed the median?

4.2 THE NORMAL FREQUENCY DISTRIBUTION

A frequency distribution for a population of numerical measurements is intended to display in some manner the density with which the measurements are distributed along the scale on which they are measured. Such a frequency distribution indicates the region (along the scale of measurement) in which the measurements tend to be most numerous, and also shows the way in which they are dispersed about that region of concentration. The reader should see that these are the same two general matters of concern considered in Chapter 2. Averages were used to measure general level of performance (as on ACE scores), and measures like the standard deviation, mean deviation, range, and quartiles were employed in the description of the dispersion of the data along the scale of measurement. This sort

of information is essential to any adequate description of a population, and also is vital when considering sampling problems.

In order that we may be able to perform certain useful statistical analyses it usually is necessary to assume (after investigation) that the data conform to some general type of frequency distribution, such as the binomial frequency distribution considered in the preceding section. In that section a formula was used to determine the frequency distribution for a binomial population when the basic information (n and p) was available. The formulas and the procedures for their use are appropriate for discontinuous measurements which fall into only two categories, such as heads and tails.

Likewise, we need a mathematical formula which is appropriate when continuous measurements (such as weights, heights, and ages) can be expected to conform to what is called a *normal frequency distribution.* Mathematicians long ago derived the necessary formula, in fact, it has been derived several different ways, all of which—as rigorous derivations—are inappropriate to this book. However, it is possible here, and useful, to show how the normal distribution is related to the binomial frequency distribution.

As the number of trials (n) is increased the number of ordinates which graphically represent the binomial frequency distribution also becomes greater. As the n increases the discontinuity of the distribution may become less important and less noticeable for many practical purposes. This matter is illustrated in Figures 4.21*A*, *B*, and *C*, for which $p = 1/2$ and $n = 5$, 20, and 100, respectively. In Figure 4.21*A* our eyes have to search a bit for the actual form of the distribution; for $n = 20$, the points rather definitely follow a certain symmetrical curve quite well; and for $n = 100$, the points of the graph dot out a symmetrical bell-shaped curve quite clearly. To put the matter another way, if the instructor were to ask each member of the class to draw a smooth curve which seemed to the student to fit the points of the figures best, there would be considerable hesitation and disagreement about Figure 4.21*A*, much less trouble with Figure 4.21*B*, and practically unanimous accord concerning the curve needed for Figure 4.21*C*.

The student will realize that the labor involved in the construction of figures 4.21*A*, *B*, and *C* becomes increasingly great as n varies from 5 to 100. In view of the fact that Figures 4.21*B* and *C* are closely approximated by continuous curves, we might hope that a relatively simple formula for a continuous curve might be employed instead of $C_{n,\,r} \cdot p^r q^{1-r}$, or instead of a summation involving this formula. For-

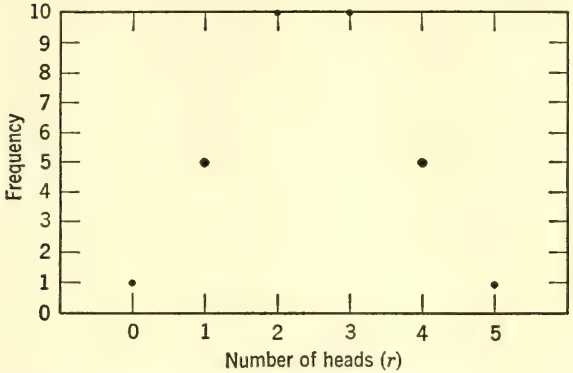


Figure 4.21*A*. The binomial frequency distribution with $p = 1/2$ and $n = 5$.

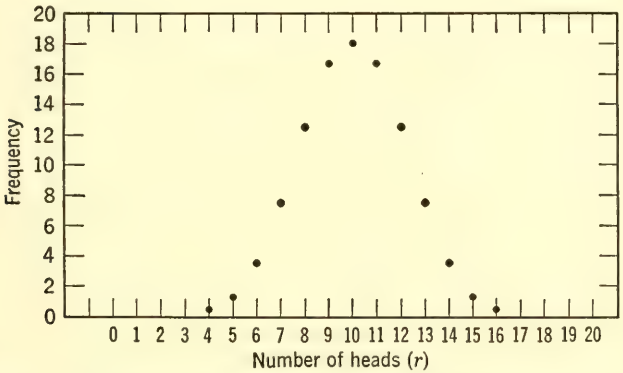


Figure 4.21*B*. The binomial frequency distribution with $p = 1/2$ and $n = 20$.

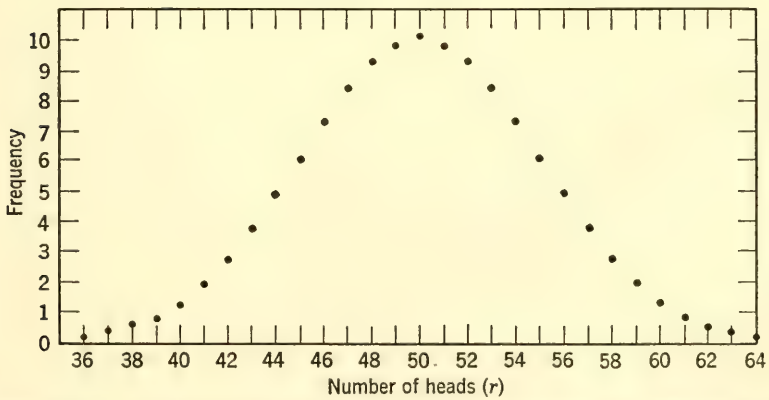


Figure 4.21*C*. The binomial frequency distribution for $p = 1/2$ and $n = 100$.

tunately, it can be shown mathematically that if n is fairly large and p is not far from 1/2, the numbers obtained from $C_{n,\,r} \cdot p^r(1-p)^{n-r}$ by setting r successively equal to 0, 1, 2, $\cdots$, and n are much the same as those obtained from $(1/\sqrt{2\pi}\cdot\sigma)\cdot e^{-(X-\mu)^2/2\sigma^2}$, wherein X replaces r, $np = \mu$, $\sigma = \sqrt{npq}$, and e = the base for natural logarithms. In particular, if $p = 1/2$ so that $\mu = n/2$ and $\sigma = \sqrt{n}/2$, it is found that the approximation is very close for $n = 20$ or more. Table 4.21 shows the approximation when $n = 20$.

TABLE 4.21

ILLUSTRATION OF THE GOODNESS WITH WHICH THE NORMAL FREQUENCY CURVE FITS THE BINOMIAL FREQUENCY DISTRIBUTION WHEN $p = q = 1/2$ AND $n = 20$

r or X	Binomial	Normal	Error	r or X	Binomial	Normal	Error
0	.000	.000	....	11	.160	.161	.001
1	.000	.000	....	12	.120	.120	.000
2	.000	.000	....	13	.074	.073	.001
3	.001	.001	.000	14	.037	.036	.001
4	.005	.005	.000	15	.015	.015	.000
5	.015	.015	.000	16	.005	.005	.000
6	.037	.036	.001	17	.001	.001	.000
7	.074	.073	.001	18	.000	.000	....
8	.120	.120	.000	19	.000	.000	....
9	.160	.161	.001	20	.000	.000	....
10	.176	.178	.002				

If the relative frequencies calculated the two ways shown in Table 4.21 are plotted on a common set of axes, Figure 4.22*A* is obtained. Graphically, the normal frequency distribution fits this binomial distribution almost perfectly at the points where the binomial distribution exists.

The sum of all the relative frequencies (ordinates) for the binomial frequency distribution is 1 because it is the sum of the probabilities for all of the $(n + 1)$ mutually exclusive events which are possible under the specified conditions. Likewise the sum of all the ordinates of the normal curve *at the points* where $X = 0, 1, 2, \ldots, 19$, and 20 will add to approximately 1. If rectangles of width 1 and heights

$$y_i = \frac{1}{\sqrt{10\pi}}\, e^{-(X_i - 10)^2/10}$$

where $i = 0$ to 20, inclusive, are constructed as in Figure 4.22*B*, their *total area* also is approximately 1. Moreover, the total area of the

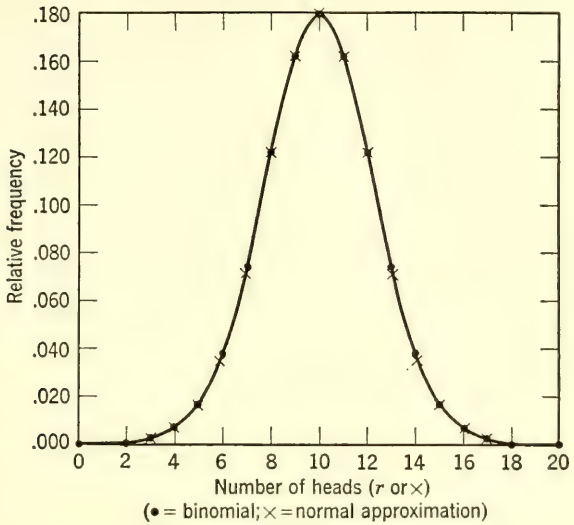


Figure 4.22*A*. The normal curve fitted to the binomial frequency distribution with $n = 20$ and $p = 1/2$.

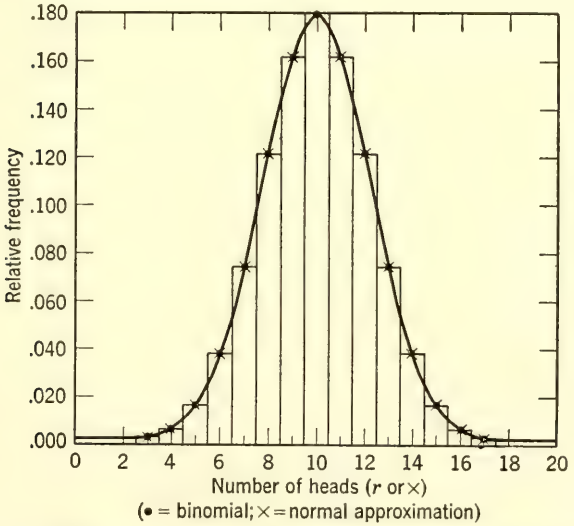


Figure 4.22*B*. Illustration of the relationship between the area under the normal curve and the probabilities which can be derived from a binomial distribution with $p = 1/2$ and $n = 20$.

rectangles is approximately the same as the area under the normal curve, as the reader can verify visually.

With the preceding remarks in mind, consider the following two facts: (*a*) To obtain the exact probability that r will have one of the values from a to b, inclusive, we need to sum the ordinates, $C_{n,r}(1/2^n)$ for $n = 20$ and $r = a, a + 1, a + 2, \ldots b$. (*b*) The operation described in *a* is approximately the equivalent of finding the area under the normal curve between the points $X = a - 1/2$ and $X = b + 1/2$. The operation of *a* is very laborious; hence if *b* can be accomplished with much less work and is satisfactorily accurate, it should be the better method. As a matter of fact, this is the case, as will be shown by some of the subsequent discussion of this chapter.

If the relative frequency of occurrence of a normally distributed measurement, X, is denoted by y_1, we have the following general formula for y_1:

$$y_1 = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(X-\mu)^2/2\sigma^2}. \tag{4.21}$$

Hence, if the μ and σ are known and the measurement X is known to have a normal distribution, we can graph the frequency distribution by the usual methods of algebra. For example, if $\mu = 60$ and $\sigma = 10$, formula 4.21 becomes

$$y_1 = \frac{1}{\sqrt{2\pi}\,(10)} e^{-(X-60)^2/200}. \tag{4.22}$$

Table 4.22 was prepared from this formula, and Figure 4.23 then was constructed from the pairs of values (X, y_1) in that table. It will be left as an exercise for the student to verify the values given for y_1 by using Table VI (end of book) to obtain e^{-w}, where $w = (X - 60)^2/200$. Thereafter, division by 10 gives the numbers in Table 4.22, under the heading y_1.

The following information can be obtained easily from Figure 4.23: (*a*) The normal distribution curve is symmetrical about a vertical line through the point where $X = \mu = 60$; (*b*) the median X, the modal X, and the arithmetic mean of the X's are equal and each is equal to 60; and (*c*) after $(X - 60)$ becomes at least twice the size of the standard deviation, either positive or negative, the corresponding ordinates, y_1, are very small. In fact, when $(X - 60)$ becomes three times the size of the standard deviation, the corresponding y_1 is practically zero. Hence, it follows that for a truly

normal distribution the useful range of the X's is about six times the size of the standard deviation. The reader should recall that this

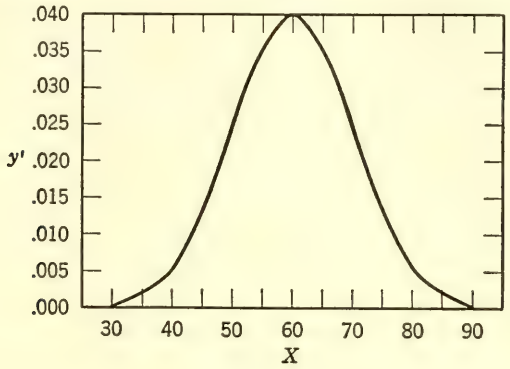


Figure 4.23. The normal frequency distribution curve for a population with $\mu = 60$ and $\sigma = 10$.

approximate relationship between the range and the standard deviation was used in Chapter 2 when the normal frequency distribution was mentioned first.

TABLE 4.22

Coordinates of Points Satisfying Formula 4.22 for a Normal Frequency Distribution with $\mu = 60$ and $\sigma = 10$

X	y_1	X	y_1
30	.000	65	.035
35	.002	70	.024
40	.005	75	.013
45	.013	80	.005
50	.024	85	.002
55	.035	90	.000
60	.040		

Most students have asked (or heard someone else ask) an instructor: "Do you grade on the curve?" *The curve* which the student has in mind is the normal frequency distribution curve; but it appears from the discussion above that there is a different normal curve for each combination of μ and σ. This is correct; but the student who is asking such a question is chiefly interested in his performance relative to the other persons who took the same examination. He hopes that a grade of 40 on one test is as good as a grade of 70 on another test if its relative rank among all grades on that

examination is the same in both instances. In other words, what is of interest is the general form of the frequency distribution of a set of grades and a system for comparing one person's grade with all the other relevant grades. The discussion which follows is intended to show how we can reduce all formulas for particular normal frequency distributions to one general—and simpler—formula which preserves all the information which we usually desire from such a formula.

Multiply through formula 4.21 by σ and then make the following substitutions of variables: let $y = \sigma y_1$ and let $\lambda = (X - \mu)/\sigma$. The result of these substitutions is the following formula for the *standard normal frequency distribution:*

$$y = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2}. \tag{4.23}$$

What has been done by means of these substitutions can be described graphically as follows: (*a*) Both the vertical and the horizontal axes have been marked off in multiples of the standard deviation, σ; and (*b*) the peak of the curve (which is above the point where $X = \mu = md = \text{MO}$) has been placed above the point where $\lambda = 0$. Hence the X's which are less than μ now correspond to negative values of λ, those which are greater than μ now correspond to positive λ's.

The first two columns of Table III give the numbers needed to construct the graph of equation 4.23. Figure 4.24 was constructed by means of this table. Figures 4.23 and 4.24 are essentially the same curve; the only difference lies in the way the vertical and horizontal axes are scaled. In Figure 4.23 λ would be 0 under the point where $X = 60$, would be $+1$ under the point where $X = 70$ because 70 is one times the standard deviation *larger* than 60, the mean. The λ would be -1 under the point where $X = 50$ because 50 is one times the standard deviation *smaller* than the mean, 60. The other corresponding values of λ and X can be determined in the same manner.

An illustration of the application of standard normal frequency distributions to a generally familiar situation can be obtained from the batting averages of baseball players. The conditions which might affect batting averages may change from season to season or from league to league so that such averages for the different situations are not directly comparable. For example, the ball may be livelier one season than during another; or perhaps the pitching may generally

be poorer one season than another. Hence an average of .350 during a season with a lively ball and generally mediocre pitching might not represent a better batting performance than an average of .320 attained with a less lively ball and more effective pitching.

These matters should be reflected in the general level of batting averages and in the consistency with which players' averages grouped about the general average. That is, the mean and the standard deviation of the batting averages should be taken into account. This is precisely what is done when standard normal units are employed.

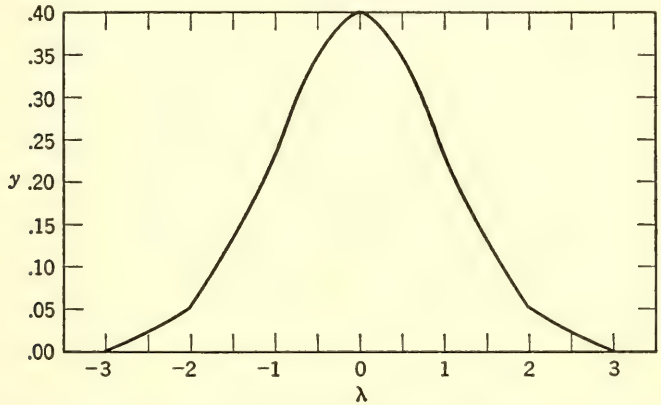


Figure 4.24. Graph of the standard normal frequency distribution whose formula is given by equation (4.23).

There also seems to be evidence indicating that batting averages can be assumed to be reasonably normal in their distribution.

Batting averages for some of the better batters from the National and American leagues (undifferentiated here but kept separate when the standard normal units were computed) are presented in an ordered array in Table 4.23, first as they usually appear and then in terms of standard normal units.

Some interesting conclusions can be drawn from Table 4.23, although they might be disputed upon the basis of other evidence and other points of view. For example: (*a*) The best batter listed for 1940 (Deb Garms) ranks fifteenth in standard normal units, considering all 4 years together. (*b*) The batter with the best average of all (Ty Cobb)—when the level and dispersion of batting averages within a league and year are taken into account—had an actual average of .385, which was bettered by five other batters unless

TABLE 4.23

ORDERED BATTING AVERAGES FOR INDICATED YEARS BEFORE AND AFTER CONVERSION TO STANDARD NORMAL UNITS

(Includes higher ranking batters who were in at least 75 games during the season.)

Actual Average				Standard Normal Units			
1910	1920	1930	1940	1910	1920	1930	1940
.385	.405	.401	.355	3.43	3.17	2.45	1.96
.364	.388	.393	.352	2.85	2.72	2.24	1.88
.340	.382	.386	.348	2.19	2.57	2.06	1.78
.331	.376	.383	.344	1.94	2.41	1.98	1.68
.325	.370	.381	.342	1.77	2.25	1.93	1.63
.322	.369	.379	.340	1.69	2.23	1.88	1.58
.321	.360	.379	.340	1.66	1.99	1.88	1.58
.320	.355	.374	.337	1.63	1.86	1.75	1.51
.312	.351	.373	.326	1.41	1.75	1.73	1.23
.309	.340	.368	.322	1.33	1.47	1.60	1.13
.308	.339	.367	.320	1.30	1.44	1.57	1.08
.306	.338	.366	.319	1.25	1.41	1.55	1.06
.305	.338	.366	.318	1.22	1.41	1.55	1.03
.304	.334	.359	.317	1.19	1.31	1.37	1.01
.302	.333	.359	.317	1.14	1.28	1.37	1.01
.301	.333	.357	.316	1.11	1.28	1.31	0.98
.300	.332	.356	.316	1.08	1.26	1.29	0.98
.300	.328	.355	.316	1.08	1.15	1.26	0.98
.298	.328	.354	.314	1.02	1.15	1.24	0.93
.298	.328	.350	.313	1.02	1.15	1.13	0.90

standard normal units are employed. In these units, his average is a decided stand-out, being 0.26 unit ahead of the runner-up. (*c*) In general, there is reason to believe that the batters in the year 1940 were not up to the standards of the other years shown in Table 4.23, especially those of 1920 and 1930.

PROBLEMS

1. Graph the normal frequency distribution with $\mu = 4$ and $\sigma = 2$ directly from equation 4.21.

2. Graph the normal distribution of problem 1, using formula 4.23.

3. Graph the normal curve which approximates the binomial frequency distribution with $n = 8$ and $p = 1/2$. Do likewise for the binomial with $n = 8$ and $p = 1/4$, and note the decrease in the goodness of the approximation.

4. Graph as in problem 3, with $n = 12$ instead of 8, and comment on the effect of increasing the size of n.

5. Graph the frequency curve for a normal population with $\mu = 10$ and $\sigma = 2$, and estimate roughly from the graph the proportion of all the measurements in this population which are greater than or equal to 12.

6. Make a frequency distribution table for the birth weights of male guinea pigs as recorded in Table 2.61, compute the μ and σ, and then graph the normal curve with the same μ and σ. How does the graph compare with the frequency distribution curve made directly from your distribution table?

7. Perform the operations of problem 6, using the records for the female guinea pigs in Table 2.61.

8. Perform the operations of problem 6, using the 4-day gains of male guinea pigs as listed in Table 2.62.

9. Perform the operations of problem 6, using the 4-day gains of female guinea pigs as given in Table 2.62.

10. Graph the binomial frequency distribution for $n = 16$ and $p = 1/2$ and then plot the corresponding normal distribution on the same axes, adjusting the height to fit the binomial. Also construct for each value of r a rectangle of base $r - 1/2$ to $r + 1/2$ and height $= C_{16,\,r} \cdot p^r q^{16-r}$. Then indicate on your graph the area under the normal curve which is approximately equal to $P(8 \leq r \leq 11)$, the probability that r will have a value from 8 to 11, inclusive.

11. Perform the operations of problem 10, with $p = 3/5$.

12. Choose any available source and compare the batting averages in the National League for 1940 and 1950, using the leading 25 players in each year and converting the batting averages to standard normal units.

4.3 DETERMINATION OF THE PROPORTION OF A NORMAL POPULATION OF MEASUREMENTS INCLUDED BETWEEN ANY SPECIFIED LIMITS

In Chapter 2 the student was given the opportunity to learn how to construct an *r.c.f.* distribution, how to graph it, and how to determine from this graph the limits on X which would include any specified proportion of the data so summarized. Furthermore, the inverse process also was discussed, namely, the determination of the proportion of the data which lies within specified limits. It is desirable to be able to obtain the same sort of information for normally distributed groups of measurements. The basis for such a procedure was given in the preceding section.

There is, however, one major difference between the process taught in Chapter 2 and that which is necessary to handle the standard normal frequency distribution. In the latter situation there is no distribution table with class intervals and cumulative frequencies determined by means of certain arithmetic procedures. Instead the *r.c.f.* distribution must be derived from the formula for the normal

distribution function. The mathematical procedures needed in this process are beyond the level of this course; but the reader can understand that the *r.c.f.* curve of Figure 4.31 plays the same general role in the analysis of normal data that the *r.c.f.* curves did in Chapter 2.

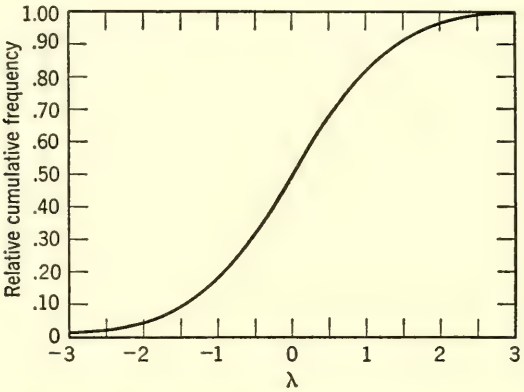


Figure 4.31. Relative cumulative frequency distribution for the standard normal frequency distribution described by formula 4.23.

The following problems will illustrate the uses to which Figure 4.31 can be put.

Problem 4.31. Determine the limits on λ for the third quartile of a standard normal population of measurements.

The limits required are obviously the median and Q_3, respectively. If we read horizontally from .50 on the vertical scale over to the normal *r.c.f.* curve and then downward to the horizontal scale, we find that $\lambda = 0$, as is to be expected. Doing likewise for .75 on the vertical scale, we find that $\lambda = 0.68$; therefore, the limits on the third quartile are $\lambda = 0$ to $\lambda = 0.68$. Since these limits apply to any standard normal distribution, the limits of the third quartile for any particular normal distribution in terms of a measurement, X, can be obtained from the relation: $\lambda = (X - \mu)/\sigma$.

Problem 4.32. What is the probability that a measurement chosen at random from a normal population with $\mu = 50$ and $\sigma = 5$ will be found to lie between 50 and 52? Between 48 and 50? Between 60 and 65?

To reduce this specific normal distribution to the standard normal distribution, substitute $\mu = 50$ and $\sigma = 5$ into $\lambda = (X - \mu)/\sigma$ so

that $\lambda = (X - 50)/5$. If $X = 50$, $\lambda = 0$; and if $X = 52$, $\lambda = 0.40$. To answer the specific question asked regarding probabilities it is necessary now to extend somewhat the concept of probability previously employed herein.

When the possible events correspond to positions along a continuous scale of measurement, the number of possibilities (previously denoted by N) is infinite. Moreover, the likelihood of occurrence changes along the scale. It no longer is useful to ask for the probability that X will have a specific size along this scale on any future trial. Instead, an event E will consist of X lying within certain limits. The probability that a randomly chosen X will fall between the limits $X = a$ to $X = b$ now will be defined to be the proportion of all the X's in the population which are included in that interval. Graphically, this will be the proportion of the whole area under the frequency distribution curve which lies between $X = a$ and $X = b$. Therefore, in problem 4.32, we need to know what proportion of this normal population lies between $\lambda = 0$ and $\lambda = 0.40$. From Figure 4.31 it is learned that 50 per cent of this population has values less than 0 and about 66 per cent has values less than 0.40; therefore, about 16 per cent of the numbers in a normal population have λ's between 0 and 0.40. It follows that $P(50 \leq X \leq 52) = .16$.

It is concluded from the symmetry of the normal curve that $P(48 \leq X \leq 50) = .16$ also. Furthermore, $P(60 \leq X \leq 65) = .025$ because 2.5 per cent of the X's have sizes within the limits 60 to 65.

As a final illustration of the use to which Figure 4.31 can be put consider a problem of grading "on the curve."

Problem 4.33. Given that a large group of grades in psychology conform to a normal distribution with $\mu = 75$ and $\sigma = 7$, suppose it is required to put letter grades on these scores in the proportions: $7A{:}20B{:}46C{:}20D{:}7F$. What are the numerical limits on each letter grade?

It is useful first to translate the proportionality above into a different form as follows. Starting with the lowest grade, F (which will be represented at the left-hand end of the scale of λ), we have the following facts:

.07 of the grades are to be F;
.27 of the grades are to be D or F;
.73 of the grades are to be C, D, or F; and
.93 of the grades are to be B, C, D, or F.

It is learned from Figure 4.31 that, for a normal population,

.07 of the X's correspond to $\lambda \leqq -1.48$;
.27 of the X's correspond to $\lambda \leqq -0.65$;
.73 of the X's correspond to $\lambda \leqq +0.41$; and
.93 of the X's correspond to $\lambda \leqq +1.50$.

In terms of the X's, we have the following facts obtained from the relation: $\lambda = (X - \mu)/\sigma$:

.07 of the X's are $\leqq 65-$;
.27 of the X's are $\leqq 70+$;
.73 of the X's are $\leqq 78-$; and
.93 of the X's are $\leqq 86-$;

therefore, the required numerical limits on the letter grades are as follows:

A = 86 on; B = 78 to 85; C = 71 to 77; D = 65 to 70;
F = below 65.

The preceding applications of Figure 4.31 have given approximate answers to the questions asked, and these answers are as accurate as the graph used and our ability to read values from it will allow. It seems rather obvious that a more accurate and, if possible, more convenient method is desirable. A method of this sort is available through the use of statistical tables. They perform essentially the same service as Figure 4.31. Although their derivation is not appropriate to this book, the reader can simply keep in mind the fact that the information obtained from Table III is the same as that which can be derived directly from Figure 4.31, but is in a more accurate and convenient form.

It will be left as an exercise to rework problems 4.31 to 4.33, inclusive, using Table III in place of Figure 4.31 as was done above.

It is worth while to investigate a set of data from Chapter 2 to see if it seems to be following a normal frequency distribution, at least approximately. Actually it is not feasible at this level of statistics to decide this matter rigorously; but some useful information can be obtained nonetheless.

Consider first the ACE scores of Table 2.01, their frequency distribution in Table 2.42, and the graph of Figure 2.41. Obviously, some approximation is introduced by using such a summary—especially one with only 12 class intervals—but the approximate distribution will serve the purpose here. The graphs of Figure 2.41 would

resemble those of Figures 4.23 and 4.31 rather closely if the former were smoothed curves instead of broken-line graphs. Hence, it appears, superficially, that the population of ACE scores follows a normal distribution fairly well if the more general and important features are the only ones considered. To be more definite, consider the following information:

(*a*) For the ACE scores, $\mu = 96$, approximately, and the median is 97. These averages are equal in a normal distribution but the discrepancy is not at all large.

(*b*) The following table shows the corresponding proportions within stated, and important, intervals on X:

Percentage of the Population Included

Interval on X	ACE	Normal	Difference
$\mu \pm 0.5\sigma$	37.6	38.3	-0.7
$\mu \pm 1.0\sigma$	67.1	68.3	-1.2
$\mu \pm 1.5\sigma$	85.3	86.6	-1.3
$\mu \pm 2.0\sigma$	95.2	95.4	-0.2
$\mu \pm 2.5\sigma$	99.2	98.8	$+0.4$
$\mu \pm 3.0\sigma$	99.8	99.7	$+0.1$

Although the deviations from normal expectancy are somewhat systematic, there being a small deficiency in the middle and a smaller excess in the tails of the distribution, the ACE distribution still seems to be approximated by the normal quite well.

If, then, it is assumed that the ACE scores do essentially conform to a normal distribution, the substitution $\lambda = (X - 96)/26$ would convert the scores of Table 2.01 into standard normal measurements. The graph of their frequency distribution essentially would be Figure 4.23, the *r.c.f.* curve would be given approximately by Figure 4.31, and Table III would present the distribution in tabular form. The statistical analysis of these data then might be more easily and efficiently accomplished than would otherwise be the case, and little or no important information would be lost in the process.

PROBLEMS

1. If a binomial frequency distribution has $p = 1/4$ and $n = 80$, calculate $P(r > 25)$ by means of the normal approximation to this binomial distribution.

2. Suppose that all the residents of a certain city definitely have made up their minds about a particular civic issue, and that 55 per cent favor one specific decision. What is the probability that on a random sample of 100 interviews

less than 50 will favor this decision, that is, it will seem that the residents are against this decision when they actually favor it? *Ans.* .13.

3. Suppose that an event E_1 occurs with a relative frequency $p = 1/2$, and that n random observations are to be made under these conditions. How large must n be before the number of occurrences, r, of E_1 will fall within one per cent of its mathematical expectation with probability equal to .10? That is, you must choose n so that $P(n/2 - n/200 \leq r \leq n/2 + n/200) = .10$.

4. Suppose that a basketball team has established in previous games that it is safe to assume that the probability on each shot by a team member that a goal will be scored is .35. What is the probability that in a game in which they take 60 shots from the field they will hit less than 18 if the idealized assumptions just stated are good? *Ans.* .17.

5. Suppose that a pair of unbiased dice are to be rolled 50 times. What is the probability that a 6 or a 7 or an 8 will appear on 20 to 25, inclusive, of these throws?

6. According to certain records the average length of growing season at Manhattan, Kansas, is 172 days. If the standard deviation about this mean is 13 days, and if lengths of growing seasons in this area are normally distributed, what is the probability that the next growing season will be long enough to mature a crop which requires 190 days to complete its development? *Ans.* .084.

7. Suppose that when wood blocks of a certain type, 2 by 2 by 8 inches, are tested for strength with the proper engineering equipment, their strengths are normally distributed with mean equal 13,000 pounds and standard deviation equal 3600 pounds. How many blocks out of 100 tested would you expect to have strengths below 6000 pounds? Between 10,000 and 15,000 pounds?

8. If you are told that the heights of 10,000 college men closely follow a normal distribution with $\mu = 69$ inches and $\sigma = 2.5$ inches:

(*a*) How many of these men would you expect to be at least 6 feet in height? *Ans.* 1150.

(*b*) What range of heights would you expect to include the middle 50 per cent of the men in the group? *Ans.* 67.3 to 70.7.

9. Assuming that the wages of certain laborers in the building trades are normally distributed about a mean of $1.80 per hour with a standard deviation of 30 cents:

(*a*) What proportion of the laborers receive at least one dollar per hour?

(*b*) What range includes the middle two-thirds of these laborer's wages?

10. Suppose that tests have indicated that certain silk fabrics have breaking strengths which are normally distributed about a mean of 27 pounds, with $\sigma = 8$; whereas, materials with a mixture of silk and rayon have $\mu = 37$ pounds and $\sigma = 9$. How likely is it that a piece of silk selected at random will be at least as strong as the average for the silk and rayon mixture? How likely is it that a randomly chosen piece of the silk-rayon mixture will be no stronger than the average for silk? *Ans.* $P = .11, .13$.

11. Suppose that the persons whose ACE scores are in Table 2.01 are to be given letter grades on the assumption that these scores are normally distributed with $\mu = 95.7$ and $\sigma = 26.1$. If 10 per cent are to get A's and 22 per cent D's, compute the score limits on each letter grade.

12. Suppose that 52 per cent of the voters in a certain city are in favor of a particular one of the possible sites for a new high school. If 100 voters are to be selected at random, what is the probability that less than 50 per cent will vote in favor of this site? If the poll is so taken that 60 per cent of those who favor that site will not participate in the poll, what now is the probability that less than 25 per cent of a sample of 100 will vote for the site in question which 52 per cent of the voters actually favor? *Ans.* .31, .12.

4.4 USE OF THE NORMAL DISTRIBUTION TO APPROXIMATE PROBABILITIES FOR A BINOMIAL FREQUENCY DISTRIBUTION

Another important use to which the normal *r.c.f.* distribution can be put has been suggested previously, namely, the approximation of the summation of $C_{n,r}p^rq^{n-r}$ from $r = a$ to $r = b$, when n is at all large and p is close to 1/2. It has been indicated that this sum is approximately equal to that area under the normal curve between the points $X_1 = a - 1/2$ and $X_2 = b + 1/2$. Moreover, it has been shown that the area under the normal curve between any two points along the X-axis can be obtained quite easily from an *r.c.f.* curve or from Table III.

To illustrate this process and to indicate its accuracy, suppose $n = 20$ and $p = q = 1/2$, and that it is required to determine $P(r \geqq 12)$. For this binomial distribution, $\mu = np = 10$ and $\sigma = \sqrt{npq} = \sqrt{5}$; hence the normal distribution with these parameters will be employed in the approximation. Also, $X_1 = 11.5$, and $X_2 = 20.5$. In terms of standard normal units,

$$\lambda_1 = (11.5 - 10)/2.24 = +0.67, \text{ and}$$

$$\lambda_2 = (20.5 - 10)/2.24 = +4.69.$$

By means of Table III and some interpolation it is found that approximately 25 per cent of a standard normal population has numbers between these λ-limits; hence $P(r \geqq 12) = .25$, approximately.

Using the last column of Table VII from $r = 12$ on down, and using a divisor of $2^{10} = 1{,}048{,}576$, the exact probability—to 4 decimals—that r will have some size from 12 to 20, inclusive, is found to be .2517. Certainly the normal approximation of .25 is excellent for most purposes, and the labor saved is considerable.

PROBLEMS

1. Given a large number of college grades which follow a normal distribution with $\mu = 65$ and $\sigma = 10$, what proportion of the grades would you expect to lie in the interval from 50 to 70, inclusive?

2. Referring to Figure 4.31, how probable is it that 3 random selections from this population will each have λ's $\geqq 2$? *Ans.* $P = .000027$.

3. What proportion of the measurements in a normal population would you expect to lie beyond $X = 1.1$ if $\mu = 0.5$ and $\sigma = 0.25$?

4. What proportion of the data described in problem 3 lies at least 0.15 unit from the arithmetic mean if the numbers are in an array? *Ans.* .55.

5. Certain frost data collected in the neighborhood of Manhattan, Kansas, over a 69-year period indicates that the average date of the last killing frost in the spring is April 24, with a standard deviation of 10 days. Assuming a normal frequency distribution and assuming that the date of the last killing frost cannot be predicted from a current year's weather, what is the probability that the last killing frost next spring will come on or after May 1?

4.5 STUDYING THE NORMALITY OF A FREQUENCY DISTRIBUTION BY RECTIFYING THE *r.c.f.* CURVE

A method was given in a preceding section for detecting gross non-normality by calculating the proportions of a population lying within such intervals as $\mu \pm k\sigma$ and comparing these with those proportions which are typical of a perfectly normal population of measurements. A graphic method will be presented in this section which will make it quite easy to compare the whole of a population with a standard normal population. The graphic method has these advantages: (1) Like other graphs, it utilizes the eye-mindedness of many persons. (2) It compares all the distribution with a standard normal instead of comparing a few segments such as $\mu \pm 0.5\sigma$, $\mu \pm 1\sigma$, etc. However, this graphic procedure has the disadvantage that it may encourage a hasty acceptance of the assumption that the given population is sufficiently near normal for the purposes at hand. More rigorous tests of normality exist in more mathematical textbooks, which can be consulted if the situation demands that additional care. It will be seen when the Central Limit Theorem is discussed in a later chapter that a considerable amount of non-normality can be tolerated in sampling studies; hence a precise—and laborious—test for normality is not often employed. In such situations a graphic test may be sufficiently reliable.

The process of rectifying a curve $y = f(x)$, which is the basic procedure of this section, is one of changing the scale of measure-

ment of either, or both, x and y so that the new graph of $y = f(x)$ becomes a straight line. That is, a curved line is straightened out by a change of scale.

Because the reader is assumed to be familiar with logarithms, the description of a method for rectifying a normal curve will be preceded by a similar discussion regarding logarithmic and exponential curves. If $Y = \log_{10} X$, as in Table 4.51 for selected X's, the pairs of values (X, $Y = \log_{10} X$) plot on the curve of Figure 4.51.

TABLE 4.51

SOME PAIRS OF NUMBERS WHICH SATISFY THE EQUATION $Y = \log_{10} X$

X	Y	X	Y	X	Y
1	0.00	100	2.00	500	2.70
3	0.48	150	2.18	600	2.78
8	0.90	200	2.30	700	2.85
10	1.00	250	2.40	800	2.90
40	1.60	300	2.48	900	2.95
50	1.70	350	2.54	1000	3.00
70	1.85	400	2.60		

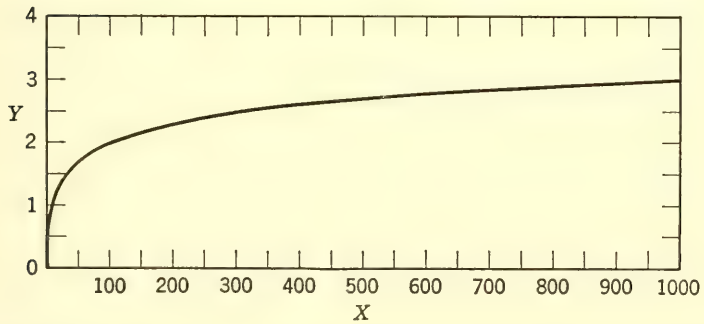


Figure 4.51. Graph of $Y = log_{10} X$ for X in the interval $1 \leq X \leq 1000$.

It is obvious that as the size of X increases, the size of $Y = \log_{10} X$ increases less and less for equal increases in X. For example, when X changes from 500 to 600, log X changes by 0.08; but when X changes from 900 to 1000 (another increase of 100), log X changes by only 0.05. It is typical of straight-line (linear) mathematical relationships that Y changes the same amount for equal increases in X. In other words, Y changes uniformly with increasing X. If log X is put on a uniform scale, as in Figure 4.52, and

the corresponding X's matched with their logarithms, the X-scale is what is called a logarithmic scale. Figure 4.53 shows the effect of graphing $Y = \log X$ against X when X is on a logarithmic scale.

Uniform scale, log X:	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		1.0
Corresponding X:	1			2		3	4	5	6	7	8	9 10

Figure 4.52. Matching of the logarithmic and the arithmetic scales of a measurement, X.

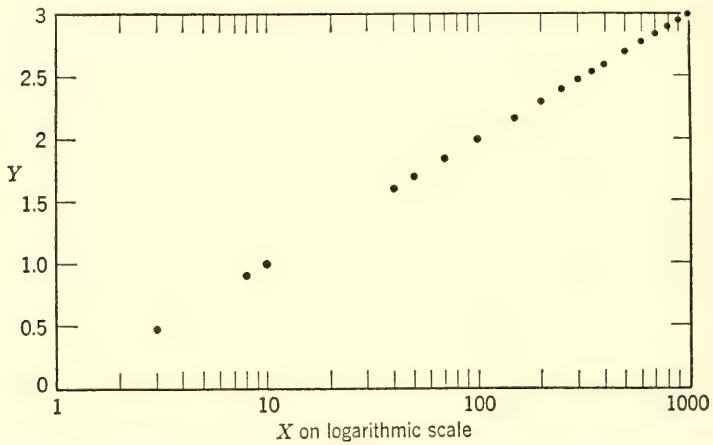


Figure 4.53. Graph of $Y = \log_{10} X$ when X is scaled according to the $\log_{10} X$ as derived from Figure 4.52.

As can be seen, the graph is a straight line, and, for any equal distance along the horizontal axis, the Y changes by the same amount. It is noted that the X-axis falls into parts of equal length: one for numbers from 1 to 10, one for numbers from 10 to 100, and another for X's between 100 and 1000. This corresponds to numbers whose logarithms have characteristics of 0, 1, and 2, respectively. Graph paper with one scale logarithmic and the other arithmetic will be called semi-log paper. When it has three repeated sections along one axis (X-axis in Figure 4.53) it is called three-cycle semi-log paper. The three cycles correspond to any three successive characteristics of logarithms, that is, to numbers which fall between any three successive powers of 10.

Figure 4.54 illustrates the use of semi-log paper to rectify an exponential curve. In this case $Y = 2e^{3X}$, but any base for the power could be used. Clearly $\log_{10} Y = \log_{10} 2 + 3X \log_{10} e$; or

$\log_{10} Y = 1.30X + 0.30$, approximately. This will be a straight line if Y is measured on a logarithmic scale, as in Figure 4.54. Table 4.52 gives the values used in plotting Figure 4.54.

TABLE 4.52

VALUES OF $2e^{3X}$ FOR SELECTED X'S

X	$Y = 2e^{3X}$	X	$Y = 2e^{3X}$
0	2.00	1.15	63.00
0.25	4.23	1.25	85.04
0.50	8.96	1.50	180.03
0.75	18.98	1.75	381.14
0.85	25.61	2.00	806.86
1.00	40.17		

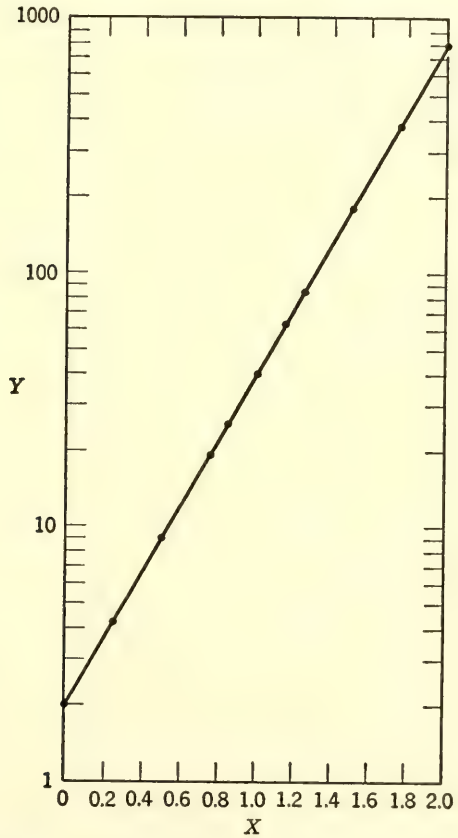


Figure 4.54. Graph of $Y = 2e^{3X}$ on semi-log paper.

With the foregoing introduction to the method of rectifying curves the same general process will be applied to the normal *r.c.f.* curve. As noted earlier, one of the questions which may arise in practice is whether or not a given type of numerical measurement does follow a normal frequency distribution. Although the graphic procedure to be illustrated is definitely not a rigorous test for normality, it may be sufficient for practical purposes.

The vertical scale to be employed will have what is called a normal *r.c.f.* scale marked off in whole percentages. The horizontal scale will be an arithmetic (or uniform) scale for the λ used previously in discussions of the standard normal frequency distribution. Figure 4.55 illustrates the process of obtaining the vertical scale in a manner which is analogous to that illustrated earlier for semi-log paper. It was constructed with the aid of Table III by plotting the normal *r.c.f.* as a percentage (top scale) directly over the corresponding λ, and then interpolating for the "integral *r.c.f.* per cent" found on the middle scale of Figure 4.55. This middle scale is the one to be used here in studying the approximate normality of frequency distributions.

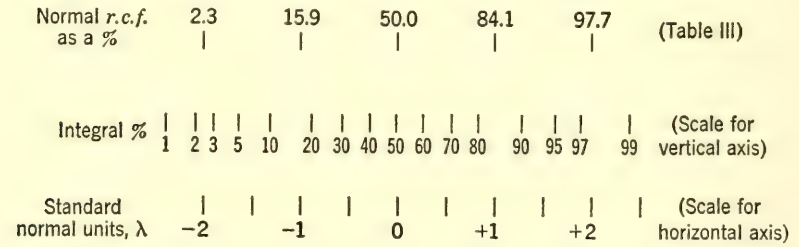


Figure 4.55. Determination of the scales for a normal-arithmetic graph.

As is to be expected after the discussion of semi-log graph paper, it is not necessary to go through the work back of Figure 4.55 because graph paper already exists on which we can do this graphing. Figures 4.56*A* and *B* were constructed on normal-arithmetic paper to illustrate the way the normality or the non-normality of a distribution affects a graph on such paper. Four distributions are employed in these illustrations:

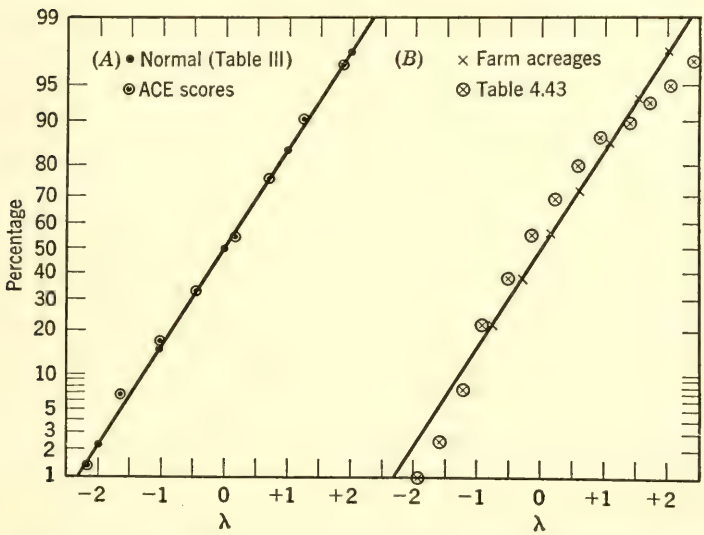
(*a*) Truly normal distribution of Table III;
(*b*) ACE scores of Table 2.01;
(*c*) the data on farm acreages in Table A (below); and
(*d*) the definitely non-normal distribution of Table 4.53.

TABLE 4.53

A Fictitious Non-normal Frequency Distribution

Class Interval	f	*r.c.f.*	λ	Class Interval	f	*r.c.f.*	λ
55–59.9...	1	100.0	3.9	10–14.9...	60	80.1	0.61
50–54.9...	2	99.8	3.5	5– 9.9...	80	69.5	0.25
45–49.9...	5	99.5	3.2	0– 4.9...	100	55.5	−0.11
40–44.9...	8	98.6	2.8	− 5 to − 0.0..1	90	37.9	−0.48
35–39.9...	10	97.2	2.43	−10 to − 5.0..1	80	22.0	−0.84
30–34.9...	12	95.4	2.07	−15 to −10.0..1	30	7.9	−1.21
25–29.9...	15	93.3	1.70	−20 to −15.0..1	10	2.6	−1.57
20–24.9...	20	90.7	1.34	−25 to −20.0..1	5	0.9	−1.94
15–19.9...	40	87.1	0.98		568		

$\mu = 6.6$, approximately; and $\sigma = 13.73$.



Figures 4.56. Some graphs of *r.c.f.* distributions on normal-arithmetic paper.

It should be evident from Figures 4.56 that the following are true:

(*a*) The frequency distribution from Table III yields a perfectly straight line when *r.c.f.* as a percentage is plotted against λ on a normal-arithmetic graph paper.

(*b*) The frequency distribution of the ACE scores apparently is quite near to normal because the points of their *r.c.f.* graph on

normal-arithmetic graph paper appear to deviate only slightly from a straight line.

(*c*) The distribution of the farm acreages in Ness County, Kansas, is essentially normal except that the lower end of the distribution is missing, that is, the distribution is truncated.

(*d*) The fictitious distribution of Table 4.53 clearly is not normal because the points of the *r.c.f.* graph definitely do not follow a straight line on normal-arithmetic paper.

It should be noted with respect to the conclusions above that only gross (and hence certainly serious) non-normality will show up under this sort of scrutiny. A look at the frequency distributions associated with (*b*) and (*c*) above shows that there certainly is some lack of normality. Figures 4.56 show this clearly; but whether or not the relative departure from a straight line is negligible will depend on the particular circumstances. Discussion to be given in Chapter 6 will be helpful in this decision.

PROBLEMS

1. Plot the following pairs of values of X and Y as points on a graph, using semi-log paper with Y measured on the logarithmic scale. Then determine the slope of the straight line through the points and relate it to the way Y changes per unit increase in X.

X:	1,	2,	3,	4,	5,	6.
Y:	2,	6,	18,	54,	162,	486.

2. Plot $Y_1 = \log_{10} X$ and $Y_2 = 5 \log_{10} X$ on the same sheet of arithmetic graph paper and also on the same sheet of semi-log paper. What effect does the coefficient 5 have on these graphs?

3. Plot the *r.c.f.* curve for $y = \dfrac{1}{5\sqrt{2\pi}} e^{\frac{-(X-1)^2}{50}}$ on normal-arithmetic graph paper.

4. Plot the following tabular *r.c.f.* distribution on normal-arithmetic paper and comment on any lack of normality revealed by your graph.

Class Interval	*r.c.f.*	Class Interval	*r.c.f.*	Class Interval	*r.c.f.*
80–82.99...	1.00	65–67.99...	.33	50–52.99...	.09
77–79.99...	.90	62–64.99...	.26	47–49.99...	.07
74–76.99...	.74	59–61.99...	.20	44–46.99...	.05
71–73.99...	.50	56–58.99...	.16	41–43.99...	.02
68–70.99...	.40	53–55.99...	.13		

5. Make an *r.c.f.* distribution for the fly counts of problem 1, section 2.4. Plot this distribution on normal-arithmetic paper, and discuss any apparent non-normality of this distribution.

6. Perform the operations required in problem 5 for all the birth weights of female guinea pigs listed in Table 2.61.

7. Perform the operations required in problem 5 for all the birth weights of male guinea pigs listed in Table 2.61.

8. Perform the operations required in problem 5 for all the 4-day gains of female guinea pigs listed in Table 2.62.

9. Perform the operations required in problem 5 for all the 4-day gains of male guinea pigs listed in Table 2.62.

10. Perform the operations required in problem 2 for logarithms to the natural base e rather than 10, and comment on the effect of this change on the graph.

REVIEW PROBLEMS

1. If you are among 1000 persons, each of whom purchases a one-dollar lottery ticket for a prize of $1000, what is the expected value of your ticket in the mathematical sense?

2. Determine the expected frequencies of sums of 3, 4, 5, and 6, respectively, when three unbiased dice are thrown simultaneously 1000 times.

Ans. 4.6, 13.9, 27.8, 46.3.

3. If 25 pennies and 15 dimes are placed in individual envelopes, thoroughly mixed, and presented to you for the selection of one envelope, what is the probability that you will get a dime? What is your mathematical expectation on such a draw?

4. How many two-digit numbers can you make up by selecting any number from 1 to 9, inclusive, for each digit? How many numbers could you form if none is to contain the same digit twice? *Ans.* 81, 72.

5. Suppose that a turtle is hatched at point A and then wanders over a uniform terrain in search of food. If he never wanders more than 1000 yards radially from spot A, and if he moves over the area in such a way that he is equally likely to be on any preassigned areas of a specified size, what is the probability that he will be within a circular area of 100 square yards whose center is 300 yards from, and northeast of, the spot A?

6. Table A (below) and Figure A present the distributions of the various sizes of farms in Ness County, Kansas. If a stratoliner were to drop a package by parachute so that it will be sure to land on Ness County, but the pilot cannot tell where, what is the probability that it will fall on a farm of more than 1000 acres if 10 per cent of the county is not in farm land and that 10 per cent is uniformly distributed over the county? *Ans.* .30.

7. If 100 farmers are to be selected from Ness County without knowledge of the areas of their farms, and if one supposes one farmer per farm, what is the mathematically expected number of representatives from farms covering less than 500 acres? What fraction of the county's farm acreage do they represent?

8. Determine graphically the lower limit of the sixtieth percentile and of the third decile for the data of Table A. *Ans.* About 520 acres; 260 acres.

9. Table B presents a summary of the years of schooling had by all legal residents of Kansas who were 25 years of age or older on April 1, 1940. Construct what appears to you to be a good graphic presentation of these data.

10. If a roving reporter were to go all over Kansas, impartially asking persons their opinions on a certain educational matter, what proportion of his interviews would you expect to be with persons who have had at least two years of college education if he talked only to persons who were at least 25 on April 1, 1940? What percentage would have no college education?

Ans. 9 per cent, 88 per cent.

11. If an insecticide is known to be 99 per cent lethal to a certain species of insect, what is the probability that less than 5 will survive if 150 selected at random are sprayed with this spray?

12. Suppose that a particular variety of apple grown under specified conditions produces yields (per tree) which are normally distributed with $\mu = 8$ bushels and $\sigma = 2.5$ bushels. What is the probability that a randomly chosen tree will be found to yield less than 5 bushels? That two such trees will each be found to yield less than 5 bushels? *Ans.* .12, .014.

13. Obtain records like those in Table B from the latest census, make an *r.c.f.* graph for those data, and determine the median years of schooling. Compare with the median for 1940 and draw any appropriate conclusions. Use college years as 13, 14, 15, 16, and 17.

14. Plot the *r.c.f.* distribution of Table A on normal-arithmetic paper and comment on the apparent normality, or lack of it, for this distribution of farm sizes.

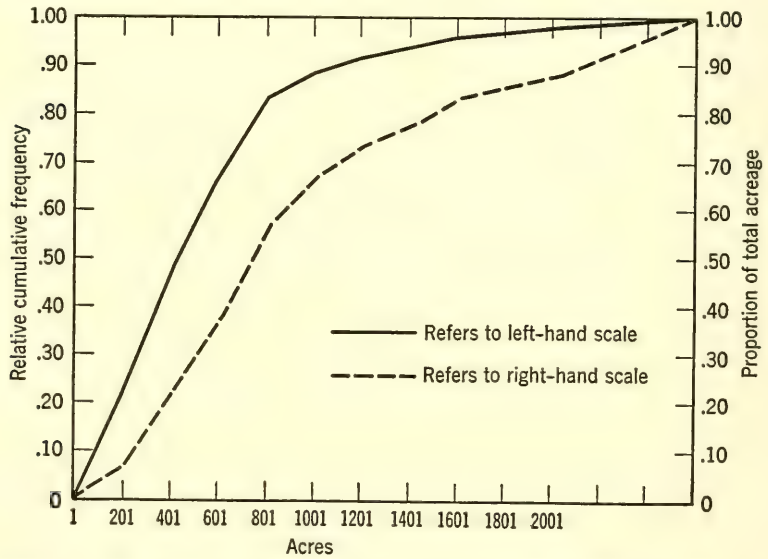


Figure A. Two types of relative cumulative frequency distributions for the sizes of farms in Ness County, Kansas, in 1940.

TABLE A. Distribution of Farm Acreages in the Whole of Ness County, Kansas

(Data furnished by W. H. Pine, Department of Economics and Sociology, Kansas State College.)

Acreage Interval	Frequency, f	Frequency, *r.c.f.*	Percentage of Total Acreage
Over 2000	28	1.00	1.00
1801–2000	7	.98	.88
1601–1800	14	.97	.86
1401–1600	21	.96	.83
1201–1400	28	.94	.78
1001–1200	35	.92	.73
801–1000	77	.89	.67
601– 800	194	.83	.57
401– 600	236	.67	.37
201– 400	320	.48	.21
1– 200	274	.22	.06
Total	1233		

$\mu = 550$ acres $md = 420$ acres

TABLE B. Years of Schooling Completed by Kansas Residents at Least 25 Years Old on April 1, 1940

Years Completed		Number, f
None		11,975
Grade	1	2,136
	2	5,507
	3	14,833
	4	29,745
	5	33,628
	6	45,722
	7	54,326
	8	388,512
High School		
Year	1	62,173
	2	61,935
	3	31,315
	4	173,580
College		
Year	1	29,113
	2	32,374
	3	12,973
	4	35,347
At least	5	12,580
Total		1,037,774

REFERENCES

Dixon, Wilfrid J., and Frank J. Massey, Jr., *Introduction to Statistical Analysis,* McGraw-Hill Book Company, New York, 1951.

Hald, A., *Statistical Theory with Engineering Applications,* John Wiley & Sons, New York, 1952.

Kenney, John F., *Mathematics of Statistics,* Part I, Second Edition, D. Van Nostrand Company, New York, 1947.

Waugh, Albert E., *Elements of Statistical Method,* Second Edition, McGraw-Hill Book Company, New York, 1943.

CHAPTER 5

Sampling from Binomial Populations

When a population of numerical measurements involves so much data that it is either impossible or unwise to attempt to analyze the whole of it, sampling must be relied upon to furnish the desired information. As a matter of fact, most of the statistical analyses now performed involve sampling data. A multitude of examples could be sited to illustrate the need for sampling, but the following will suffice for the purposes of this discussion.

(5.01) *Public opinion polls.* Only a small percentage of the persons eligible for interview actually are questioned about the matter under study. The sole objective of the study is to estimate the proportions of the citizens favoring the various points of view. If the question to be asked has only a yes or a no answer the results of the poll will constitute a sample from a binomial population, and we would be attempting to estimate p.

(5.02) *A study of the toxicities of two insecticides conducted by spraying insects of a certain species with the insecticides and counting the dead insects.* This is another case of sampling a binomial population; but the purposes of the investigation may be different. The following question is to be answered: Is one of the sprays more toxic to these insects than the other? Statistically, the question becomes: Is it reasonable to suppose that the two sets of data obtained with the two sprays are samples from the same binomial population? Of course, such a study also may include the estimation of p as mentioned in (5.01).

(5.03) *Testing the breaking strengths of concrete columns, of wood or of metal beams, and of other engineering materials.* Breaking strengths are measured on a continuous scale of numbers; hence their populations have continuous frequency distributions. Problems

of this sort include the estimation of true average breaking strengths, and comparisons of the strengths of different materials.

(5.04) *Studies involving two variables such as prices of selected stocks and the volume of production of finished steel, ACE score and grade average in college, stand counts of wheat and the yield of a plot, etc.* In such investigations it would be necessary to estimate from sampling data the relationship between the two variables, express it mathematically, and then use it in accordance with the purposes of the investigation.

It can be seen from the examples above that two general types of statistical problems must be considered in sampling studies. One is to derive from the sample observations some numbers which can be used satisfactorily in place of one or more unknown population parameters. These numbers which will be derived from the sample are called *sampling estimates* of the parameters. They are changeable from sample to sample and, being dependent upon chance events, are subject to the laws of probability.

The other general problem is to test hypotheses regarding populations against actual sample evidence. For example, if the populations of the breaking strengths of two types (different shapes, for example) of concrete columns each follow a normal frequency distribution with the same variance, σ^2, these populations can differ only in their means μ_1, and μ_2. That is, it is supposed that the engineers in charge are satisfied that the two types of columns have the same *uniformity* of performance from test to test, but it is yet to be decided whether they have the same average strength. If so (that is, if $\mu_1 = \mu_2$), the populations of breaking strengths are identical normal populations. It then becomes a problem of deciding from samples taken from each population whether or not μ_1 is in fact equal to μ_2. It usually is convenient statistically to assume that μ_1 does equal μ_2, and then to see how reasonable this hypothesis is in the light of sample observations.

It should be clear—intuitively, at least—that decisions based on samples may be in error, and that we do not know in any particular case if our sample is so unusual that it is misleading us. How, then, can sample evidence become a satisfactory basis for making decisions about populations? The answer lies in the fact that, while no one can say whether a *particular* decision is right or wrong, it is possible to determine the relative frequency with which correct decisions will be made over the long-run of much experience if we are following

certain rules for acting upon the basis of sampling evidence. It follows that the probability of making a correct decision from any specified future sample (say the next one we are going to take) also can be stated.

To illustrate some of the preceding discussion, suppose you are about to engage in a coin-tossing game in which "heads" is the event which is of particular interest to you. Assume, also, that you are not satisfied that the coin is unbiased but are not going to worry about bias unless the probability of heads is as low as 1/3. Before playing the game you are going to flip the coin 15 times and then come to a decision regarding the bias of the coin. What rules for action should you adopt and how effective will they be in detecting bias as bad as $p = 1/3$? It is being assumed that you are not entertaining the possibility of bias toward too many heads.

As long as the coin has two sides and one is heads, the other tails, any result from 0 to all 15 heads can occur on 15 flips *regardless of bias in the coin.* However, it should be clear that the relative frequencies of occurrence of the 16 possible results are dependent upon the size of p. For $p = 1/3$, for example, such a result as 15 heads on 15 throws is an extremely rare occurrence. The actual rarity, in terms of probability, can be derived from the binomial series for $(q + p)^n$, with p, q, and n given.

If $p = 1/2$ and $n = 15$, the binomial series is

$$(1/2 + 1/2)^{15} = .000 + .000 + .003 + .014 + .042 + .092 + .153$$
$$r: \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6$$
$$+ .196 + .196 + .153 + .092 + .042 + .014 + .003$$
$$r: \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13$$
$$+ .000 + .000.$$
$$r: \quad 14 \quad 15$$

When p is unknown and a sample has produced $r = 0, 1, 2,$ or 3 heads on 15 random flips, you probably would be very reluctant to accept the hypothesis, $H_0(p = 1/2)$ because the total probability of the occurrence of one of these 4 mutually exclusive events is but .017, or about 1 chance in 59. Although it is true that one of those 4 results can be obtained when the coin is unbiased—and you knew this before you tossed the coin 15 times—you are now faced with the necessity to decide if the coin is biased or not, and you must do

so upon the basis of the sample's evidence. If you decide to reject the hypothesis that $p = 1/2$ whenever the observed number of heads is one of the 4 cases just listed, you will unjustly reject H_0 1.7 per cent of the time because that is how frequently such cases occur by chance when p does equal 1/2. Nevertheless, some rules for action must be adopted or else nothing can be decided from samples. Hence, it will be supposed that the following rules will be followed after 15 sample tosses of the coin in question:

(*a*) If $r = 0$, 1, 2, or 3 heads, you will reject $H_0(p = 1/2)$ and assert that the coin is biased against heads.

(*b*) If $r \geqq 4$, you will accept H_0 and play the game on the assumption that the coin is not biased against heads.

These two rules can lead you to correct conclusions and actions, and they also can cause you to make one of two kinds of errors:

(1) The hypothesis $H_0(p = 1/2)$, which is being tested by sampling, may be rejected when it is true. This will be called an *error of the first kind.* In the above example, the probability that such an error would occur was noted to be .017 under the rules *a* and *b*.

(2) The H_0 may be accepted when it is false. This will be called an *error of the second kind.* It should be clear that the likelihood of committing an error of this kind depends on what possibilities—or alternative hypotheses—there are.

It is customary to set up the hypothesis H_0 in such a way that it is considered more serious to make an error of the first kind than it is to accept a false hypothesis. When this is done the probability of committing an error of the first kind (to be designated by α) is kept low—usually $\alpha < .10$—and the rules adopted for acting upon the basis of sampling evidence are chosen so that for a given α the probability that an error of the second kind will be made (to be designated by β) is as small as possible under the circumstances.

Referring back to the coin-tossing problem, we see that $\alpha = .017$. Also, the person who was trying to decide from 15 throws if the coin was seriously biased would not care if p had some size between 1/2 and 1/3, but did wish to detect a p as low as 1/3. Hence the alternative hypothesis whose truth could lead to *errors of the second kind* includes all p's at or below 1/3. For the sake of simplicity it will be assumed that the only alternative hypothesis to $H_0(p = 1/2)$ is $H_1(p = 1/3)$.

The β can be determined from the following series:

$$(2/3 + 1/3)^{15} = .002 + .017 + .060 + .130 + .195 + .214 + .179$$
$$r: \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6$$
$$+ .115 + .057 + .022 + .007 + .002 + .000 + .000$$
$$r: \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13$$
$$+ .000 + .000.$$
$$r: \quad 14 \quad 15$$

Hence if p actually is $1/3$ so that the hypothesis of no bias should be rejected, the probability is $.002 + .017 + .060 + .130 = .209$ that H_0 will be rejected. Or the probability that H_0 will not be rejected when it should be—an error of the second kind—is $\beta = 1 - .209 = .791$. Obviously, the rules a and b would not be good ones if it is serious to fail to detect the bias indicated by $p = 1/3$. However, if the most serious mistake is to accuse someone of employing a biased coin when he is innocent, rules a and b may be quite satisfactory.

In practice we seldom can compute β as simply as above. Usually the α is set at an appropriate level and then standard tests are employed without actually knowing the β. However, it can be said here that the tests to be discussed in this, and the next, chapter have been chosen with the idea of making the β as small as possible under the circumstances and for the chosen α.

As the heading of this chapter indicates, the subsequent discussion will be confined to samples from binomial populations. Later chapters will take up the normal and the two-variable situations.

5.1 OBTAINING THE SAMPLE

Before a method for obtaining the sample is devised, the population which is to be sampled must be defined clearly. It is recalled from Chapter 4 that a binomial population is possible only if the units in some definable group have attributes which may be described by just two classes. Moreover, the fractional part of the population falling into each class must stay fixed. For example, all the farmers in Finney County, Kansas, on July 1, 1953, could be classified unambiguously into two classes as regards membership in some cooperative association: those who do belong to some cooperative and those who do not belong to any such association. The units would

be the individual farmers' designations as member or as non-member. The fraction of the total number of farmers in that county who were classed as members could be the parameter p of Chapter 4. Then $1 - p$ would be the fraction who were classified as non-members on the stated date.

It is entirely possible for those same farmers to be the basis for other populations. Their answers for or against a proposed new federal farm policy could constitute another binomial population of interest, their per-acre incomes during a specified period could be another (non-binomial) population, and the sizes of their families on July 1, 1953, could be still another (non-binomial) population which might be of interest to some group of persons.

The chief criterion of a good definition of a population which is about to be sampled is that it make entirely clear in all important respects the larger group of units to which the conclusions drawn from the sample will pertain.

Given a well-defined population, the sample obviously must be taken in such a manner that the impression it produces through statistical analyses will have the greatest possible chance to be accurate and dependable. Naturally the facilities and economic resources available for the sampling may be limiting factors; but it will be assumed in the discussion to follow that those resources and facilities are at least good enough to justify undertaking the sampling study at all. For purposes of illustration, suppose that we wish to determine public opinion in a large city regarding a political issue of current interest, and that our resources allow us to interview only one person per each hundred in the population. How should this one per cent sample be taken? If we were to visit the few major business districts we could obtain our allotted number of interviews more quickly and with less cost because more persons are concentrated in these small areas during business hours; but we have no assurance that the opinions of the persons we would meet there are the same as those we would find in the outlying districts, for example. We might consider using the telephone directory until we thought of the fact that some residents do not have telephones. If we are interested only in the opinions of registered voters—as is easily possible—we could use an official listing of those persons. It is possible that for some questions which might be asked we could take a random sample of names from this list and interview them as our sample. Such a random sample could be taken by numbering the entries consecutively from one to the number of voters on the list and then

drawing numbers at random until a sample of the desired size was obtained. Such a procedure would make it true that every possible sample of the size n had been equally likely to be drawn at the outset of the sampling, and this is necessary in random sampling.

In many circumstances the procedure of sampling just outlined would be unsatisfactory. A city might be made up of racial and economic groups of such diverse opinions on the matter being studied that it would not do to leave their representation in the sample to pure chance, as in the random sampling just described. It would be necessary to sample each group in accordance with its proportionate part of the city's total registration of voters.

It is noted that the sampling discussed above has supposed that the sample will be taken by means of personal interviews. Any such systems as calling persons on the telephone or mailing questionnaires, which depend on voluntary and selective responses, or on their being at a certain place at a certain time, are almost certain to produce biased samples. The cause of their not responding, and hence not being in the sample, may be associated with the type of response they would have given.

The theory and techniques of sampling in such a way that the conclusions which can be drawn therefrom will be accurate and reliable are very extensive and cannot be covered here. The remarks above merely point out a few of the more important and general considerations. However, the reader can be warned to be critical of any conclusions drawn from samples until he is satisfied that the samples were taken in such a way that they should be representative of the population about which conclusions have been drawn. If one brand of cigarettes is said to be the favorite of a certain professional group, we should at least wonder if that group was properly sampled. Or, if someone returns from a foreign country and asserts that the residents of that country hold certain points of view regarding a matter of world-wide interest, we should wonder if he did an adequate job of sampling public opinion in that country. Or, as a final example, if someone seeks to obtain a sample by means of a mailed questionnaire, we should wonder if those who do not respond have a different opinion, say, from that generally expressed by those who did return their questionnaires. If so, what population did those who returned their questionnaires represent?

PROBLEMS

1. Suppose that you were sent out to ascertain the public opinion in a certain community regarding the necessity for flood control of a certain type in that area. How would you obtain your sample so that it would be representative of the whole community?

2. Referring to problem 1, would it make any difference in the manner in which you took your sample if it were taken in July of 1936 during a severe drouth or in July of 1951 right after a record flood? Justify your answer.

3. Suppose that a roving reporter goes into a city with the intention of ascertaining public opinion on a matter of foreign policy. He is going to walk about the streets asking persons at random a specific question requiring one of three answers: Yes, No, or No Opinion. Will it make any difference what hours of the day, between 7 A.M. and 6 P.M., he does this? Would the day of the week matter? Would the type of city—industrial, college site, farming community, rich suburb, and the like—have anything to do with the answers to these questions?

4. Suppose that a company which is manufacturing candies develops a new product whose originators believe is especially good. Which of the following possible ways of testing the public's reaction to this new confection would you prefer to use? Why?

(*a*) Sit back and see how the sales go.

(*b*) Have some trained persons take samples out to the public and ask people to taste the candy, to record their reactions, and to give these records to the field representatives directly.

(*c*) In each of the first 10,000 packages manufactured, place a stamped and addressed card requesting that the purchaser record his opinion of the candy and mail the card to the company.

(*d*) Ask a panel of expert candy tasters to decide the matter.

(*e*) Do as in *d*, first, then *a*.

(*f*) Do as in *d*, first, then *b*.

(*g*) Do as in *d*, first, then *c*.

(*h*) Have all the firm's employees record their opinions of the candy and decide from these records if mass production is wise.

(*i*) Combine *h* and *a*.

(*j*) Combine *h* and *b*.

(*k*) Combine *h* and *c*.

(*l*) Combine others above. Specify.

(*m*) Specify another method if you have one you prefer.

5. Suppose that some engineering concern wishes to test the strength of two types of structural beams, each produced and recommended by a different company. Which of the following sampling procedures would you recommend if the engineering group has two laboratories, each with its operating personnel, available for the tests?

(*a*) Ask each company to send a specified number of beams for testing and have each laboratory test half of each company's product.

(*b*) As in *a*, but have one laboratory test all one company's beams, the second laboratory testing all the second company's beams.

(*c*) Go into the public market and purchase the necessary number of beams of each type, and then do as in *a*.

(*d*) As in *c*, but replace *a* by *b*.

(*e*) Specify other ways.

6. An agronomist wishes to run critical yield, protein, and test weight studies on a proposed new variety of wheat before the variety is released to the public. He proposes to use a standard and widely planted variety for comparison with the new one. Plenty of land is available for this study, but it is quite nonuniform in soil qualities, moisture content, and exposure to weather. Which of the following outlines for such a study would you prefer, and why?

(*a*) Plant the new variety on the east half of the available land, the standard variety on the west half (or vice versa, as decided by flipping a coin), harvest and measure wheat from each half, determine test weight and protein content on the yield from each half separately.

(*b*) Divide the available area into 20 equal-sized plots and plant 10 plots to each variety, choosing the variety for a plot by drawing the names from a hat. Then determine yield, protein, and test weight separately from each plot's wheat.

(*c*) Do as in *b*, except that the plots are grouped into 10 *pairs* and each pair has both varieties planted side by side.

(*d*) Save the land for some other purpose, send out samples of each wheat to 10 farmers, and ask them to report the yields and test weights and send in samples for protein analysis.

5.2 CALCULATION OF POINT AND INTERVAL ESTIMATES OF p FOR A BINOMIAL POPULATION

It was indicated in Chapter 4 that a binomial frequency distribution can be defined when individuals are identified only as belonging to one of two possible classes of attributes such as male or female, dead or alive, acceptable product or unacceptable product, and the like. Moreover, the proportions falling into the two classes of attributes are constantly $p:(1-p)$.

If n members of a binomial population are selected at random, the particular individuals drawn are the result of chance occurrences. Hence, we may find that any number from $r = 0$ to $r = n$ of those individuals possess the attribute A, say, even though a fixed proportion, p, have that attribute in the whole population. The possible outcomes of such a sampling vary from $r = 0$ to $r = n$ and form a binomial frequency distribution with mean $\mu = np$ and with standard deviation $\sigma = \sqrt{npq}$, as was shown in Chapter 4. The reader is reminded that the probability that exactly r of the n members of

the sample will possess an attribute with probability of occurrence $= p$ for any specified future trial from the population is given by the formula: $C_{n,r} \cdot p^r(1-p)^{n-r}$.

The point of view in the preceding paragraph is that of Chapter 4 in which the size of p was assumed to be known. More commonly, p is not known and we have only a sample estimate of its size. This estimate is r/n, which varies under repeated sampling from 0 to 1. Even though r/n is a variable quantity, useful and reliable conclusions can be drawn from samples taken from a binomial population, as will be shown shortly. Three types of such conclusions will be considered in this chapter: (*a*) Given a sample, what can we say about the size of p? (*b*) Given a sample from a binomial distribution, how well does it agree with a predetermined hypothesis concerning the magnitude of the p for that population? (*c*) Given two random samples, did they probably come from the same binomial population? The present section is concerned with question *a*.

When the true proportions of the two types of members of a binomial population are not known, they can be estimated by means of a sample, as suggested above. This estimation can take either of two forms: (*a*) a *point*, or specific, *estimate* of p, which would be used in lieu of the p, or (*b*) an *interval estimate* which would have a preassigned probability of bracketing the size of p. This latter process is called placing a confidence interval on p. The confidence we can have that the bracket, or interval, does actually include the unknown parameter is described by the *confidence coefficient*.

Statistical research indicates that the best *point estimate* of p is obtained from $\hat{p} = r/n$, the observed fraction of the sample which possess the particular attribute that is being studied. Some of the reasons for this decision are:

(5.21) The $\hat{p}$ has an expected value $E(\hat{p}) = E(r/n) = E(r)/n = np/n = p$ for any particular sample size, n. That is, the long-run average size of $\hat{p}$ is exactly equal to the true population parameter p. It is customary to call point estimates *unbiased estimates* if their mathematical expectation is the parameter which is being estimated. We generally prefer to employ unbiased estimates, like $\hat{p}$, unless some more important property is missing.

(5.22) The estimate $\hat{p} = r/n$ has a variance $= pq/n$ because the variance of r is npq—as shown in Chapter 4—and the effect of dividing the r by n is to divide the variance by n^2, as was shown in the section of Chapter 2 which dealt with the coefficient of variation. This vari-

ance, pq/n, of the estimate $\hat{p}$ will be quite small for a sample of almost any useful size because the p and the q are each less than unity. This indicates that $\hat{p}$ will not vary greatly from sample to sample, especially if the sample size is fairly large. As a matter of fact, the size of the variance of $\hat{p}$ can be made as small as desired by taking the n sufficiently large. Hence, this estimate, $\hat{p}$, is considered to be a very *efficient* estimate of p.

In view of the fact that $\hat{p}$ is almost always in error to some degree in spite of the fact that it is the best point estimate possible, there are many circumstances in which an interval estimate of p is desirable. The interval estimate also is more difficult to compute and to interpret; hence it will be considered in some detail.

The situation is this: n members of a certain binomial population have been drawn at random so that each member of the population had an equal opportunity to be in the sample, and r of them have been found to have the specified attribute A. Given the proportion r/n observed in the sample, what useful limits can we place on the true proportion, p, of A members in the whole population, and what confidence can we have in those limits? It is customary to call such interval estimates *confidence limits*, or to say that these limits constitute a *confidence interval*. The degree of confidence which we can place in such limits on p is measured by the probability that the sample has given an interval which actually does include p. As might be expected, this probability is the relative frequency with which the sampling process used will produce an interval which does include p. It will be convenient to use the symbol CI_{95}, for example, to designate the confidence interval which has—at the start of the sampling process—95 chances out of 100 of including the parameter which is being estimated.

Suppose that a relatively small manufacturing concern is producing roller bearings which are to be shipped to a larger company manufacturing farm machinery. There will be certain specific standards, such as maximum or minimum limits on diameter, which the bearings must meet before they are considered to be acceptable products. Hence, any large batch of bearings could be grouped into two subgroups marked as "acceptable" and "unacceptable," respectively, if every bearing were to have its diameter measured with perfect accuracy. It will be assumed here for simplicity of discussion that the company which is to receive the bearings requires that each shipment must be 90 per cent "acceptable" or it can be rejected. The concern which is producing the bearings will have to

inspect its products by means of samples because it is inconceivable that every bearing should be carefully measured.

Assume that a sample of 10 bearings has been inspected and that all 10 were found to be acceptable. Is this sufficient evidence that the shipment probably is up to the standard? In this connection, consider a binomial population with p only .80; that is, it is well below the standards set above. The probability that every member of a sample of 10 will be acceptable is $(.80)^{10}$, which is .11; hence there is about 1 chance in 9 that this definitely substandard batch of bearings will show none unacceptable on a sample of but 10. Obviously, if p were less than .80, p^{10} would be less than .11; and, conversely, if p were larger than .80, p^{10} would be greater than .11. Therefore, it should be clear that the result, 10 acceptable bearings out of 10 inspected in a sample, could be obtained from any one of a whole range of possible binomial populations corresponding to values of p ranging from 0 to 1. As a matter of fact, the sample discussed above could be drawn at random from any binomial population with as many as 10 acceptable bearings among the individuals. Of course, with $n = 10$, a sample with r also equal 10 is more likely to come from a population with p near 1 than from a population with p near 0.

The above discussion re-emphasizes the fact that we cannot attain certainty in conclusions drawn from samples: there always must be some risk that the sample has led to a false conclusion. We choose a risk of error which we can afford to take and express it in terms of the confidence coefficient described earlier. If it be supposed that an event which is as unlikely to occur as 1 time in 20 can be ignored, what confidence interval (CI_{95}) can we set on p as a result of the above sample in which $r = 10$ acceptable bearings out of 10 observed in the sample?

We use what will be called a *central 95 per cent* of all possible r's by determining a range on r which is such that not more than 2½ per cent of all samples with the same n and p will fall beyond each end (separately) of the range so determined. For example, in the series below for $(1/4 + 3/4)^{10}$ the first five terms—to the left of the brace—add to .0197, which is less than 2½ per cent, or .0250. If the sixth term from the left is added, the sum exceeds .0250. Therefore, among all possible samples of 10 observations from a binomial distribution with $n = 10$ and $p = 3/4$ the sample number, r, will be below 5 for a bit less than 2½ per cent of all such samples. At the other end of the series for $(1/4 + 3/4)^{10}$ no term is less than or equal to .0250; hence, the "central 95 per cent" will be occupied by samples

for which $r = 5, 6, 7, 8, 9$, or 10. Consequently, if you have drawn a sample with $n = 10$ and p is unknown, it is quite unlikely that p was as large as 3/4 if it was found in the sample that $r = 0, 1, 2, 3$, or 4. As a matter of fact, you could just form the habit of assuming that p never is as large as 3/4 whenever r is 0, 1, 2, 3, or 4 and you would be wrong less than 5 per cent of the time because samples of that sort occur less than 5 per cent of the time when $n = 10$ and p is as large as 3/4.

To answer the question posed earlier, we consider the following reasoning. If $n = 10$, the probability series for p set successively equal to 2/3, 3/4, .69, and .70 (for reasons which will appear soon) are obtained as in Chapter 4 and lead to the following conclusions:

$$(1/3 + 2/3)^{10} = .0000 + .0003 + .0031 + .0163\{+ .0569 + .1366$$
(sum = .0197, is $<.0250$)

r:	0	1	2	3	4	5

$$+ .2276 + .2601 + .1951 + .0867\}+ \quad .0173$$
is $<.0250$

r:	6	7	8	9	10

(Note that the "central 95 per cent" does *not* include the observed number of occurrences, $r = 10$.)

$$(1/4 + 3/4)^{10} = .0000 + .0000 + .0004 + .0031 + .0162\{+ .0584$$
(sum = .0197, is $<.0250$)

r:	0	1	2	3	4	5

$$+ .1460 + .2503 + .2816 + .1877 + .0563\}$$ none excluded

r:	6	7	8	9	10

(The central 95 per cent *does* include the sample result, $r = 10$, but still might do so with a smaller p; hence $p = 3/4$ may be too large to be the lower end of the 95 per cent confidence interval. Therefore, $p = .70$ will be tried.)

$$(.3 + .7)^{10} = .0000 + .0001 + .0014 + .0090\{+ .0368 + .1029$$
(sum = .0105, is $<.0250$)

r:	0	1	2	3	4	5

$$+ .2001 + .2668 + .2335 + .1211 + .0282\}$$ none excluded

r:	6	7	8	9	10

(The central 95 per cent still includes the observed number, $r = 10$, and it again is possible that p could be smaller and still keep $r = 10$ in the central 95 per cent. Hence, try $p = .69$.)

$$(.31 + .69)^{10} = .0000 + .0002 + .0018 + .0108\{+ .0422 + .1128$$

(sum = .0128, is $<.0250$)

r:	0	1	2	3	4	5

$$+ .2093 + .2662 + .2222 + .1100\}+ \quad .0245$$

is $<.0250$

r:	6	7	8	9	10

(The central 95 per cent now just barely excludes the observed number, $r = 10$; therefore, the *smallest* value of p which has been considered here and which still keeps $r = 10$ within the central 95 per cent is .70. However, it is clear that if three decimal places were used, the lower end of the confidence interval would be nearer to .69 than to .70; hence, .69 is taken as the lower end of the 95 per cent confidence interval.)

To determine the upper end of the 95 per cent confidence interval, it is necessary to find out by a similar procedure how *large* p can become and still leave the observation, $r = 10$, in the central 95 per cent of the binomial population with $n = 10$. Obviously, p can go all the way to 1.00, or 100 per cent, and still not exclude the case when $r = 10$; hence, $p = 1.00$ is the upper limit of the 95 per cent confidence interval when r has been found to be 10 when $n = 10$. Therefore, it is concluded that if with $n = 10$, r is observed to be 10 also, the 95 per cent confidence interval on the true percentage in the population is $.69 \leq p \leq 1.00$. At the same time, the person doing the sampling is aware that there are 5, or less, chances in 100 that his sample has been sufficiently "wild," or unusual, that it has produced a confidence interval which fails to include the true proportion, p, of acceptable products in the population which was sampled.

The work done above is illustrative of the principles involved but is too laborious to be repeated each time a confidence interval is needed, especially when $n > 10$. Therefore, advantage is taken of some work done by C. J. Clopper and E. S. Pearson, published in Volume 26 of *Biometrika*. Table 5.21 was obtained by reading from their graphs the 95 and 99 per cent confidence intervals on p for $n = 50$, 100, and 250. If n is smaller than 50, the confidence intervals are so wide that they are of doubtful value in practice. However,

TABLE 5.21

THE 95 AND 99 PER CENT CONFIDENCE LIMITS ON p FOR SAMPLES OF 50, 100, AND 250 TAKEN FROM BINOMIAL POPULATIONS

(L = lower limit)

(Based on graphs by C. J. Clopper and E. S. Pearson, Volume 26 of *Biometrika*.)

	$n = 50$				$n = 100$				$n = 250$			
	95%		99%		95%		99%		95%		99%	
r/n	L	U	L	U	L	U	L	U	L	U	L	U
.000	0	8	0	10	0	4	0	5	0	2	0	2
.025	0	12	0	15	0	8	1	9	1	5	1	6
.050	1	16	0	19	2	11	1	13	3	9	2	10
.075	2	19	1	22	3	14	2	17	4	11	4	13
.100	3	22	2	26	5	18	4	20	6	15	5	16
.125	5	25	3	29	6	21	5	23	8	17	7	19
.150	6	28	4	32	8	24	7	26	11	20	9	22
.175	8	31	6	35	10	26	8	29	13	23	11	25
.200	10	34	7	38	12	29	10	32	15	26	13	27
.225	12	37	9	41	15	32	12	35	18	29	15	30
.250	14	39	11	44	17	35	14	38	20	31	18	33
.275	16	42	12	47	19	38	16	40	22	34	20	35
.300	18	45	15	49	21	40	18	42	24	36	22	38
.325	20	47	17	52	23	42	21	45	27	39	25	40
.350	22	50	18	54	26	45	23	48	29	41	27	43
.375	24	52	21	57	28	48	25	51	31	44	30	46
.400	26	55	23	59	30	50	27	54	34	46	32	49
.425	29	57	25	62	32	53	30	56	36	49	34	51
.450	31	60	27	64	35	55	32	58	39	51	37	54
.475	33	62	29	66	37	58	34	61	41	54	39	56
.500	35	65	31	68	40	60	37	63	44	56	41	58
.525	38	67	34	71	42	62	39	65	46	59	44	60
.550	40	69	36	73	45	65	41	68	48	61	46	63
.575	43	72	39	75	47	67	44	70	51	64	49	65
.600	45	74	41	77	50	69	46	73	54	66	51	68
.625	48	76	43	79	52	72	49	75	56	69	54	70
.650	50	78	46	81	55	74	51	77	59	71	56	73
.675	52	80	48	83	57	77	54	79	61	73	59	75
.700	55	82	51	85	60	79	57	81	64	76	62	77
.725	58	84	54	87	62	81	60	83	66	78	65	80
.750	60	86	56	89	65	83	63	85	69	80	67	82
.775	63	88	59	91	67	85	65	87	71	82	70	84
.800	66	90	62	92	70	87	68	90	74	85	72	86
.825	68	92	65	94	73	90	71	91	77	87	75	89
.850	71	94	68	95	76	91	74	93	80	89	78	91
.875	75	95	71	97	79	93	76	95	82	91	81	93
.900	78	96	74	98	82	95	80	97	85	93	84	95
.925	81	98	77	99	85	97	83	98	89	95	87	96
.950	84	99	81	99	88	98	86	99	91	97	90	98
1.000	92	100	89	100	96	100	95	100	98	100	97	100

Table 5.21*a* has been added to show, through the preceding discussions, just how the numbers in Table 5.21 could be got. Obviously, if n were as large as 50 the work illustrated above would become tremendously laborious.

In Table 5.21 the observed fraction, r/n, was used instead of r because it was convenient to do so.

TABLE 5.21*a*

95 PER CENT CONFIDENCE LIMITS WITH $n = 10$

r	L	U	r	L	U	r	L	U
0	0	32	4	11	75	8	44	97
1	0	46	5	18	82	9	54	100
2	2	57	6	25	89	10	69	100
3	6	66	7	34	94			

The use of Table 5.21 will be illustrated by some examples.

Problem 5.21. Suppose that a random sample of 250 primers for cartridges has been taken from a large batch and has been tested by actual firing. If 6 of the primers fail to fire, place a 99 per cent confidence interval on the true percentage of duds in the whole batch, and interpret these limits.

For this sample $r/n = 6/250 = .024$, which is so near to the value of .025 listed in Table 5.21 that interpolation is unnecessary. Therefore, the required confidence interval is read from the table as 1 to 6 per cent duds in the whole batch. If future action regarding these primers is based on the assumption that at least 1 per cent but not more than 6 per cent of them are duds, a risk of only 1 in 100 is being run that the sample has been misleading. If 6 per cent is more than the allowable proportion of duds, this sample indicates that the batch may be substandard. Whether the primers would be rejected or additional evidence obtained would depend upon the particular circumstances.

Problem 5.22. Suppose that a concern which manufactures roller bearings must meet a standard of 95 per cent acceptable according to certain prescribed measurements. If a sample of 250 yields 3 unacceptable bearings, is the shipment up to the required standard or not?

In this instance $r/n = .012$, so the 95 and 99 per cent confidence intervals are found to be 0 to 3, and 0 to 4, respectively, by interpolation in Table 5.21. Therefore we could conclude that the shipment has less than 5 per cent unacceptable with considerable confidence

because even the upper limit of the 99 per cent confidence interval on the true proportion of duds is below 5 per cent.

The procedures demonstrated above are not suggested as sufficient quality control measures in themselves, but they do illustrate principles which are basic to acceptance sampling.

PROBLEMS

1. Suppose that 100 bolts have been taken at random from a large group and that 2 have been found to be defective. What is the 99 per cent confidence interval on the true proportion of defectives in the group sampled?

2. Suppose that a sample of 250 Germans showed that 101 had type O blood. Place 95 per cent limits (to nearest per cent) on the percentage of such German persons having type O blood. *Ans.* 34 to 46.

3. The little fruit fly, *Drosophila melanogaster,* has been used so extensively in genetic research that a great deal is known about the genes which it carries on its chromosomes. Among these genes are some which produce what are called recessive lethals because they kill the potential offspring at an early stage of development if both chromosomes carry the gene for that particular lethal. Mating studies are able to show if only one chromosome of a fly carries a particular lethal-producing gene. Suppose that a sample of 250 flies is found to include 10 which are carrying one particular lethal. What can you say about the true proportion of lethal-carrying flies in this population?

4. Suppose that two different strains of fruit flies have been developed in a laboratory upon the basis of the numbers of eggs that the females laid per day. Suppose also that a particular recessive lethal, l_1, has been discovered in both strains; and that samples of 250 flies from each strain gave these results: strain A had 18 lethals, strain B had 32 flies carrying lethals among the 250 examined. What can you conclude about the true proportion of lethal-carrying flies in each strain? Are these two proportions probably equal?

5. Suppose that 50 apples have been selected at random from a tree which has a very large number of apples. If 5 apples were found to suffer from a certain blight, what percentage of blight do you estimate for the whole tree if you wish to run a risk of only 1 in 20 that your answer is wrong as a result of an anomalous sample? Do likewise for a risk of only 1 in 100 being in error.

6. Suppose that 100 eggs are selected at random from a large shipment, and that 5 are found to be stale. What would you set as the upper limit on the percentage stale in the whole shipment if you can afford a risk of sampling error of only 1 in 100? *Ans.* 13.

7. If a sample of 250 gun barrels in a large shipment has been examined for defects and none found to be defective, place 99 per cent confidence limits on the true proportion of defective barrels in the whole shipment. Would a sample of 250 be large enough if the shipment must contain one per cent, or less, defective?

8. If 250 one-pound cartons of butter are to be selected from a carload at random and examined for mold particles, what is the maximum number which can be found to contain too much mold before you should conclude that 5 per

cent, or more, of the cartons probably contain too many mold particles? Use a 99 per cent confidence interval as the basis for your answer. *Ans.* 21.

9. Suppose that 100 cattle selected at random from a very large group have been tested for tuberculosis. If 15 were found to be reactors, place an upper limit on the proportion of reactors in the whole group if 95 per cent confidence in the answer is considered adequate in these circumstances.

10. The United States Department of Agriculture publication, *Agricultural Statistics,* 1946, indicates that among United States herds of cattle which are infected with Bang's disease at all, an average of 12 per cent of the cows have the disease. Suppose that a large herd which has some incidence of the disease is to be tested by taking a random sample of 50 cattle. How many out of the 50 must be free of the disease before the owner can be assured (at the 99 per cent level of confidence) that his herd is above average in freedom from Bang's disease? *Ans.* 46.

11. Calculate directly from the binomial series $(q + p)^4$ the 75 per cent confidence interval on p if 3 out of 4 items sampled are found to be acceptable in the sense employed earlier. Obtain the answer to the nearest whole per cent.

12. Verify the entry in Table 5.21*a* for $r = 5$.

13. Suppose that an entomologist wishes to know what percentage of the corn plants in a large field have been infested to *some* degree by the southwest corn borer. He thinks that the percentage is somewhere between 20 and 80, but he wants to reduce that uncertainty to an interval of not over 15 percentage points. If he is willing to accept a risk of 5 in 100 of drawing an erroneous conclusion, how large a sample must he take?

14. Suppose that you are helping to administer a farm management association and wish to learn what percentage of the members use a certain procedure recommended for poultrymen. Suppose, also, that a random sample of 250 interviews reveals that 200 in the sample do use the recommended practices. What can you say about the true percentage using this practice in the whole association? *Ans.* CI_{95}: 74–85% / CI_{99}: 72–86% } use practices.

5.3 TESTING PREDETERMINED HYPOTHESES REGARDING p

In some fields of investigation the probable magnitude of p can be deduced from what appear to be reasonable theoretical considerations, as was illustrated in the discussions of the A-B, and other, blood groups. As another illustration, the theory of sex inheritance might lead geneticists to conclude that male and female offspring of human beings should be produced in equal proportions. If so, $p = 1/2$ when children are classified merely as male or female. Abundant statistical evidence now exists to show that more than one-half the children born in the United States are male; therefore, the original hypothesis that $p = 1/2$ is known to be false. However, mankind cannot afford to wait many years until the collection of a great

volume of data makes it possible to determine, virtually without error, the truth or falsity of many of the hypotheses which play important roles in everyday life and in scientific investigations. In these circumstances samples can be taken and made the basis for satisfactory conclusions.

The statistical methods needed for a test of a predetermined hypothesis regarding some binomial population are intended to decide whether or not it is reasonable (as defined by an accompanying probability statement) to suppose that a given sample actually has been drawn from the binomial population which is specified by the hypothesis being tested. In order that such a decision can be made, a basis must be established for comparing a particular sampling result with results to be expected from sampling *if the hypothesis being tested is strictly correct.* How should this be accomplished? Actually, the problem is a very complex one whose full solution cannot be attempted at the reader's present stage of statistical development; but some useful and informative rationalizations can be presented.

Strange as it may seem, a large part of the complexity of this problem comes from the fact that there are so many possible solutions that the more difficult job is to choose the best one. This was indicated in the introductory part of this chapter. In that introduction a rather simple example was considered and a hypothesis was judged for reasonableness by means of the binomial expansion $(q + p)^n$. It was possible with the aid of that expansion to say that if $p = 1/2$ only 1.7 per cent of a large number of random samples would have r as small as 0, 1, 2, or 3. The rarity of such occurrences was made the basis for rejecting $H_0(p = 1/2)$. Although the risk of falsely rejecting H_0 when p actually is $1/2$ is only .017, the likelihood of falsely accepting H_0 when p actually is $1/3$ was seen to be high; nearly four chances out of five. It was stated in connection with that example that the choice of the best procedure for making decisions from samples depends on this latter probability of an error of the second kind because the probability of an error of the first kind usually is fixed in advance.

In the example just reviewed, a sampling frequency distribution was employed, and events which fell in the lower frequency intervals —that is, the extreme sizes of r—constituted what is called the *region of rejection.* On the scale of measurement of r, the points 0, 1, 2, and 3 comprised the region of rejection. The general problem of choosing best tests of hypotheses regarding population parameters consists of finding functions of the sample observations and of the

population parameters for which the best regions of rejection can be defined. The best region of rejection, among several choices, is the one which for a given α will make β the smallest; that is, for a fixed probability of rejecting a true hypothesis it will give the lowest probability of accepting a false hypothesis in consideration of the other possible hypotheses.

Statistical research has shown that a good function to use in the solution of the problem set up for this section is one which is called chi-square, and is denoted by the symbol χ^2. Its magnitude depends upon the numbers of individuals, or other units, observed in the sample to fall into each of the possible classes of attributes. It also depends upon the numbers which are expected mathematically to fall in those classes, which in turn depends upon the predetermined hypothesis regarding the population parameter p. For example, suppose that we have sufficient reason to believe that one-half the offspring of guinea pigs should be males. The predetermined hypothesis now is that $p = 1/2$. If a sample group of progeny selected at random from a whole population of actual or possible progeny is found to have 38 males and 32 females, is it reasonable to believe that this is a sample from a population for which $p = 1/2$? The mathematically expected number of males out of 70 offspring is $E(r) = (1/2)(70) = 35$; hence the number of males in the sample is 3 greater than expectation. It follows automatically that there are 3 fewer females than expected mathematically.

The function χ^2 will be defined by the following formula:

$$\chi^2 = \sum \frac{(\text{observed number in class} - \text{expected number})^2}{\text{expected number in class}}, \tag{5.31}$$

where the summation includes two terms, one for males and one for females. It is apparent from this formula that if the observed numbers in the two classes agree well with those numbers which are expected mathematically considering the assumed magnitude of p, χ^2 will be relatively small; but if the numbers observed to fall in each class notably disagree with those expected from the predetermined hypothesis, χ^2 will be relatively large. The decision that χ^2 is relatively large or small is based upon the proportion of all such sample values of χ^2 which would be at least that large *if the hypothesis being tested were, in fact, true.*

For the illustration above, $\chi^2 = (38 - 35)^2/35 + (32 - 35)^2/35 = 0.51$. The remaining question is: Is it reasonable to suppose that

χ^2 got so large as this purely as the result of the chance occurrences of sampling? As was done earlier, attention will be called first to some actual sampling experiences, and then a mathematical table will be employed to obtain the required information more quickly and more accurately. Table 5.31 summarizes the results obtained from 652 samples from a population for which p was known to be 1/2. Figure 5.31 is the graph of the *r.c.f.* distribution presented in Table 5.31.

TABLE 5.31

OBSERVED FREQUENCY AND *r.c.f.* DISTRIBUTIONS FOR χ^2 WHEN THE TWO CLASSES, MALE AND FEMALE, DETERMINE THE POPULATION, AND $p = 1/2$

Class Interval	f	*r.c.f.*	Class Interval	f	*r.c.f.*
≥ 5.50	9	1.000	2.00–2.49	50	.885
5.00–5.49	6	.986	1.50–1.99	48	.808
4.50–4.99	4	.977	1.00–1.49	70	.735
4.00–4.49	7	.971	0.50–0.99	101	.627
3.50–3.99	10	.960	0.00–0.49	308	.472
3.00–3.49	12	.945			
2.50–2.99	27	.926	Total =	652	

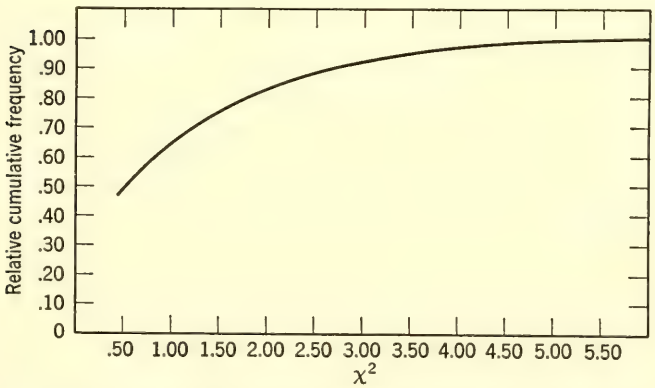


Figure 5.31. Sampling distribution of χ^2 with one degree of freedom, as determined from 652 samples taken from a binomial population with $p = 1/2$.

If we read upward from $\chi^2 = 0.51$ to the graph of Figure 5.31 and then horizontally to the vertical scale, it appears that about 52 or 53 per cent of all such sampling values of χ^2 would exceed 0.51. Obviously, then, 0.51 is not an unusual sampling size for χ^2, provided the hypothesis upon the basis of which the expected numbers were calculated is exactly correct. Hence it is entirely reasonable to suppose that this sample of male and female guinea pigs deviated from

a 50:50 sex ratio purely as a consequence of the chance element in all sampling. On the other hand, if the sampling χ^2 had been 15, say, we see from Figure 5.31 that sampling variation almost never produces such a large value of χ^2. We would then conclude that the sex ratio was not 50:50. It is clear that such conclusions as these are valid only if a representative sample has been drawn. If just a few guinea pigs in one particular laboratory have been the basis for the sample, the conclusions drawn would not apply, without more sampling evidence, to guinea pigs in general.

As a matter of fact, Figure 5.31 can be used as above for samples of various sizes just as long as there are only two classes of attributes involved. Under these circumstances, χ^2 is said to have one degree of freedom. In this connection it should be noted that when the expected number has been calculated for one of the two classes, the other number follows automatically so as to keep the sum of the expected numbers equal to the sum of the observed numbers. Likewise, if there are 3 too many males compared to expectation there must be 3 too few females; that is, there really is but one basic difference between the observed numbers and the expected numbers. Basically, that is the reason there is only one degree of freedom for the χ^2.

Table V makes it possible to determine more easily and accurately the probability that a sample χ^2 will exceed the observed value when the H_0 is correct. Actually this mathematical distribution is not exactly right for the χ^2 as defined in this chapter, but the loss of accuracy is negligible for most sample sizes which would cause people to have faith in the conclusions drawn therefrom.

Problem 5.31. It was stated in Chapter 3 that the A-B blood groups were considered to be inherited in a simple Mendelian manner so that AO $\times$ AO should produce offspring three-fourths of whom test to be type A and one-fourth are type O. Suppose that among a random sample of 400 children from such parents, 312 are A and 88 are O; that is, 78 per cent are A and 22 per cent are O. Does this sampling evidence justify rejection of the hypothesis that 75 per cent should be A, or, in more symbolic terminology, should the hypothesis $H_0(p = 3/4$ for A) be rejected?

The expected numbers are 300 A and 100 O; therefore, $\chi^2 = (12)^2/300 + (12)^2/100 = 1.92$, with 1 D/F. By Table V it is found by interpolation that $P = .17$. By any usual standards $H_0(p = 3/4$ for A) is accepted, especially if the hypothesis seems to be well founded theoretically. In some circumstances we would wish to

place a confidence interval on p. For example, the 95 per cent confidence interval could be obtained (from a larger table than 5.21 on page 127). Such an interval would include $p = 3/4$, because that hypothesis was accepted far above the 5 per cent level, but would also include other possible values of p. If this interval included other defensible hypotheses about p, they also would be acceptable as far as this sample evidence is concerned. Larger samples then could be taken with the hope of so narrowing the confidence interval that only one theoretically defensible hypothesis would be acceptable upon the basis of the sampling evidence.

PROBLEMS

1. According to Table 2.61, 50 female and 59 male guinea pigs were born during the period from January to April, inclusive. If these guinea pigs can be considered as a random sample of all guinea pigs as regards the sex ratio, is the observed difference in numbers of each sex sufficient to cause you to reject the hypothesis that the sex ratio actually is 1:1, if you wish to set the probability of committing an error of the first kind at .05?

2. Use the data for May to August, inclusive, to answer the question of problem 1. *Ans.* No, $P(\chi^2 \geqq 1.12) \cong .30$.

3. Solve as in problem 1 for the data for September to December.

4. Table 2.62 contains data from those guinea pigs which survived long enough to produce 4-day gains. Do the data for January to July, inclusive, indicate that the sex ratio is 1:1 for guinea pigs in that more select population which lives at least 4 days? *Ans.* Yes, $P > .53$.

5. Solve as in problem 4 for the data for August to December.

6. According to genetic theory, if a so-called heterozygous red-eyed fruit fly is mated with a white-eyed fruit fly, one-half the offspring are expected mathematically to be white-eyed. The reasoning is analogous to that given earlier for a mating of O and AB blood types. Suppose that among 500 offspring of such fruit flies, 240 are white-eyed. Does the χ^2-test indicate that such a sample result would occur rarely ($P < .05$) while sampling from a binomial population with $p = 1/2$, or not? *Ans.* No, $P \cong .37$.

7. If you assume (as is reasonable from Figure 5.31) that χ^2 must be at least 3.8 in problem 6 before the hypothesis that $p = 1/2$ should be rejected, how small can the number of white-eyed flies be among 500 offspring before that would occur?

8. Suppose that it were agreed that you should not seriously doubt the hypothesis that $p = 1/2$ unless χ^2 exceeds a value χ_0^2 which is such that $P(\chi^2 \geqq \chi_0^2) \leqq .01$. How small can the number of white-eyed flies among 500 become before you would reject the hypothesis that $p = 1/2$?

Ans. 221 or 222.

9. Suppose that a sample of 100 college students showed that 40 opposed a certain proposal regarding student government. Does that result contradict the hypothesis that 48 per cent of the student body oppose the proposed change?

10. Use a confidence interval approach to answer the question in problem 9, and discuss the difference between the two methods.

Ans. CI_{95}: 30–50 per cent opposed. CI_{99}: 27–54 per cent opposed. Both include 48 per cent.

11. Suppose that a poll of Topekans (Kansas) shows that one candidate received 135 votes to 115 for the other candidate for a certain public office. Use both the χ^2-test and confidence intervals to determine the probable winner, if the election is to be held very soon so that no appreciable change in opinion is expected.

5.4 TESTING THE HYPOTHESIS THAT TWO RANDOM SAMPLES CAME FROM THE SAME BINOMIAL POPULATION

The type of problem to which the tests described in this section apply arises when two groups of observations have been taken under somewhat different circumstances. The question to be answered is: Did the difference in circumstances produce two distinct binomial populations as far as can be told from these samples? For example, consider a simulated test of two house-fly sprays, one made from lethane the other from pyrethrum. Suppose that 500 house flies have been placed in each of two wire cages, identical in all respects. The lethane spray is applied to one cage, the pyrethrum spray to the other, with the following results:

Spray	Dead	Alive	Sums
Lethane	475	25	500
Pyrethrum	450	50	500
Totals	925	75	1000

Actually, the lethane spray killed 95 per cent of the flies in its cage, whereas the pyrethrum killed only 90 per cent. However, if both cages had been sprayed with the same spray, different percentages would have been killed in the two cages in all probability. How rarely would they have been as different as they were found to be in this experiment? The χ^2-test introduced in section 5.3 can be employed successfully in the solution of this problem. However, there is no predetermined hypothesis regarding the magnitude of p like that available before. Hence some other method must be used to calculate the expected numbers needed in the χ^2-test.

There is no theory regarding insecticides which will furnish an expected proportion "dead" in the population; but it was observed

that among the 1000 flies sprayed 92.5 per cent were later classified as dead. If these two sprays are equally toxic to the house flies, they should tend to kill equally many flies per 500 sprayed. Therefore, the probability of death can be taken as .925 on the general hypothesis that the two sprays are equally toxic. This is equivalent to the hypothesis, $H_0(p_1 = p_2)$. Then the expected number dead out of 500 in a cage is $E(r) = .925(500) = 462.5$. That leaves 37.5 as the expected number of survivors for each spray since 500 flies were sprayed with each spray. We then can extend formula 5.31 to obtain the following:

$$\chi^2 = (475 - 462.5)^2/462.5 + (450 - 462.5)^2/462.5 + (25 - 37.5)^2/37.5 + (50 - 37.5)^2/37.5 = 9.01.$$

This χ^2 has only one degree of freedom as before because there is only one chance difference between the observed and expected numbers. Note that only one expected number need be calculated before all the rest follow automatically from the border totals of the table. Figure 5.31 and Table V clearly indicate that χ^2 rarely would attain a size of 9.01, or more, purely from sampling variations; therefore it is concluded that the lethane spray is superior to the pyrethrum spray, that is, the hypothesis that $p_1 = p_2$ is rejected, where p_1 = true proportion which would be killed by lethane and p_2 is the same for pyrethrum over many trials.

The technique just described also can be used to decide if two random samples supposedly drawn from the same binomial population actually are from a common population. For example, suppose that two separate random samples were taken on the toxicity of a lethane spray, with the following results:

	Dead	Alive	Sums
Sample 1	480	20	500
Sample 2	380	20	400
Sums	860	40	900

If p remained constant during this sampling it is best estimated as $\hat{p} = 860/900 = .956$ or 95.6 per cent. In the absence of any logical predetermined hypothesis, the hypothesis $H_0(p_1 = p_2)$ is tested, where p_1 = true probability of death during the taking of the first sample, and similarly for p_2 and the second sample. If the probability of death for any randomly designated fly stays fixed, $p_1 = p_2$.

As usual, the expected number killed during the first sampling is computed to be $E(r) = .956(500) = 478$, which deviates from the observed number by only $480 - 478 = +2$. It follows that the other observed numbers also differ from their expected numbers by 2 in one direction or the other. Hence

$$\chi^2 = \frac{(+2)^2}{478} + \frac{(-2)^2}{382} + \frac{(-2)^2}{22} + \frac{(+2)^2}{18} = 0.422,$$

with 1 D/F. It is learned from Table V that $P = .52$; hence H_0 is accepted readily, and it is considered that the two samples were taken under conditions which kept the probability of death constant. It is not always true that the population can be kept the same under repeated sampling; hence it is well to check this matter before different conditions (such as use of different insecticides) are purposely introduced so that their effects can be studied.

Problem 5.41. Suppose that two sample polls of votes for two candidates for a public office are taken, one from among residents of cities with at least 25,000 population, the other from among residents not in any incorporated town or city. If the results were as given below would you accept the statement that place of residence was unrelated to voting preference in this election? If so, the two samples are from a common binomial population.

	Votes for A	Votes for B	Sums
Rural	620	380	1000
Urban	550	450	1000
Sums	1170	830	2000

Over both the rural and urban samples 58.5 per cent voted for A. If both samples are from the same binomial population, $\hat{p} = .585$ is the best available estimate of p, the true fraction who favor A. Hence the hypothesis $H_0(p_r = p_u)$ will be tested by means of the χ^2 distribution. The expected number of rural residents out of 1000 who favor A is $.585(1000) = 585$. It deviates from the observed number by $620 - 585 = +35$; hence

$$\chi^2 = (35)^2 \left[\frac{1}{585} + \frac{1}{585} + \frac{1}{415} + \frac{1}{415}\right] = 10.09, 1\ D/F.$$

It is apparent from Table V that H_0 should be rejected because $P \cong .002$. It is concluded that p_r actually is $> p_u$; that is, the resi-

dents of rural areas favor candidate A more strongly than do the urban residents because the observed results are very unlikely to be a sampling accident.

PROBLEMS

1. Lerner and Taylor, University of California, published the following data on chick mortality in the *Journal of Agricultural Science,* Volume 60:

	Number Progeny Which	
Sire	Died	Lived
G14	22	65
G36	44	35
G52	17	45
H8	22	39

How would you rate these sires as regards low progeny mortality after taking account of sampling variability?

2. Compute χ^2 for the following practice data and obtain from Table V the probability that sampling variation alone would produce a χ^2 at least this large. Also explain how information of this sort is used to test a hypothesis about a binomial population.

	Answered		
	Yes	No	Sum
First Sample	187	113	300
Second Sample	213	187	400
Sum	400	300	700

Ans. $P(\chi^2 \geq 5.16) \cong .002$.

3. Given the following χ^2's, each with one degree of freedom, classify each as probably due to chance alone, or not, if an event which is as unlikely to occur as 1 time in 20 is considered to be purely a chance occurrence: 3.9, 7.1, 0.95, 2.1, 15.2, 8.7, and 1.2.

4. Within what approximate limits do the lower 75 per cent of all sampling values of χ^2 with one degree of freedom lie when the hypothesis being tested is correct? *Ans.* 0 to 1.31.

5. For a population of χ^2's each with one degree of freedom, the mean $= \mu = 1$, and the standard deviation $= \sigma = \sqrt{2}$. Approximately what proportion of the population of χ^2 with 1 D/F (see Table V) lies in each of the following ranges: $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$? How do these proportions compare with the corresponding ones for a normal distribution, as shown in Table III? What information does this set of comparisons give about the shape of the chi-square frequency distribution curve when χ^2 has one degree of freedom?

6. Samples from records of male and female White Rock chicks up to 8 weeks of age raised at Kansas State College during 1945 gave the following data:

Sex	Died	Lived
Male	46	227
Female	30	290

Use the χ^2-test to decide if there probably is a fundamental difference in chick mortality due to sex. *Ans.* $P(\chi^2 \geqq 7.35) \cong .007$.

7. The following data are derived from a publication by Atkeson et al. (*Journal of Economic Entomology,* 37:428–35) on the effectiveness of 5 sprays in killing flies around dairy barns:

	Number of Flies	
Spray	Killed	Not Killed
A	22	84
B	49	90
C	89	28
D	39	63
E	44	24

Which sprays do you consider as essentially equal in killing power, considering sprays which do not differ beyond reasonable sampling variation as being tied?

5.5 THE χ^2-TEST WHEN MORE THAN ONE DEGREE OF FREEDOM IS REQUIRED

There are many problems in sampling which require the use of the sampling distribution of a χ^2 with more than one degree of freedom, but only a few will be considered in this book. For example, suppose that both parents have the specific blood types AO and Rhrh, in the symbols of Chapter 3. Each parent produces four types of gametes: ARh, Arh, ORh, and Orh, with equal frequencies it is believed. Hence it can be deduced that such parents will produce offspring of four blood types: ARh+, ARh−, ORh+, and ORh−, with associated probabilities 9/16, 3/16, 3/16, and 1/16, respectively. Therefore, if a large number of such parents is obtained for a random sample we can test the hypothesis suggested by the above argument, namely, H_0(9 ARh+ : 3 ARh− : 3 ORh+ : 1 ORh−). To illustrate, suppose that out of 1600 such families in a random sample, the children were classified as follows with respect to the A-B and Rh blood groups: 885 ARh+, 310 ARh−, 292 ORh+, and 113 ORh−. Do these observed numbers deviate enough from the corresponding

theoretical numbers 900, 300, 300, and 100, respectively, to justify the rejection of H_0? In these circumstances

$$\chi^2 = \frac{(885 - 900)^2}{900} + \frac{(310 - 300)^2}{300} + \frac{(292 - 300)^2}{300} + \frac{(113 - 100)^2}{100}$$
$$= 2.49,$$

by analogy with previous problems when χ^2 had only one degree of freedom. How many degrees of freedom does this χ^2 have, that is, how many of the deviations from theoretical expectations can be considered due to chance? The observed numbers and the expected numbers each must add to 1600, hence there cannot be more than 3 degrees of freedom. As this is the only such restriction, there are just 3 degrees of freedom. Naturally a sampling chi-square which results from 3 chance deviations usually will be larger than one based on fewer degrees of freedom because more "room" must be left for sampling fluctuations. It is seen in Table V that a χ^2 with 3 degrees of freedom will exceed the observed value, 2.49, about one-third of the time when H_0 is correct; that is, $P = .33$. Therefore, the hypothesis H_0(9 ARh+ : 3 ARh− : 3 ORh+ : 1 ORh−) is quite acceptable in view of this sample evidence.

Another circumstance which produces a χ^2 with more than one degree of freedom is encountered when the same hypothesis is tested more than once by means of successive but independent random samples believed to have been taken under the same conditions. It may not be reasonably possible to obtain a convincingly large sample during any one experiment or survey so that some means of accumulating statistical evidence from two or more studies is needed. This problem can be solved with the aid of the following theorem.

Theorem. If two or more sample chi-squares are obtained from independent random samples, the sum of these chi-squares follows the chi-square distribution for a number of degrees of freedom equal to the sum of those for the chi-squares so added.

Obviously, the process to which the above theorem refers would make no practical sense unless each χ^2 were obtained while the same hypothesis was being tested. It also is important to be assured that all the samples have been drawn from the same binomial population, regardless of the truth of the hypothesis H_0 because nothing is accomplished by such a study if several different populations are involved. We are trying to test one predetermined hypothesis which supposedly applies to a fixed set of conditions. To illustrate these

ideas, suppose that some respected group of persons has conjectured that 60 per cent of the voters in a certain area will vote yes on a given economic question, and that this conjecture is to be tested by means of three samples taken in the three districts in the area involved. It will be assumed that these districts contain equally many voters.

Before the hypothesis that $p = .60$ for the whole area is tested, it is of interest to determine if the three districts are the same binomial population with respect to yes and no votes on the economic question which is to be asked the voters. Hence it is supposed that a poll of 200 randomly chosen voters in each district gave these results:

	Number Voting		
District	Yes	No	Sum
1	105	95	200
2	100	100	200
3	125	75	200
Sum	330	270	600

If the whole of each district has the same fraction, p, of yes votes, the best estimate of p is $\hat{p} = 330/600 = .55$ or 55 per cent. If this is used as the probability that a randomly chosen voter in a given district will vote yes, the expected number of yes votes in each district is $.55(200) = 110$. That leaves 90 as the expected number of noes; hence

$$\chi^2 = \frac{(105 - 110)^2}{110} + \cdots + \frac{(75 - 90)^2}{90} = 7.07.$$

It is seen that the observed number of yeses in the first district is 5 below expectation; hence the number of noes is 5 above expectation, and only one chance difference between observation and theoretical expectation exists. The same can be said for district 2; but since we know that the yes vote in district 2 was 10 below expectation and that in district 1 is 5 below expectation, it follows that the number of yeses from district 3 must have been 15 above expectation. Hence only two chance deviations are involved basically, and this χ^2 has 2 degrees of freedom.

The specific hypothesis being tested is that the true proportions of yes votes in the three districts are equal. Table V indicates that it is rather uncommon during sampling experience for a χ^2 with $2D/F$ to become as large as the 7.07 observed for these samples if the hy-

pothesis used is correct. In fact $P(\chi^2 \geqq 7.07) = .03$. If we have decided in advance to reject a hypothesis when $P < .05$, $H_0(p_1 = p_2 = p_3)$ would be rejected, and we would say that the true fraction of yes votes is not the same in all three districts. It is clear that after such a decision it would not be valid to conduct a separate survey in each district and then combine the evidence from these samples on the assumption that we have three independent χ^2's testing the same hypothesis, as is supposed in the theorem stated earlier in this section.

It appears from the samples given above that p_1 does equal p_2, but that p_3 is greater than p_1 or p_2. This hypothesis could be tested by the method just illustrated; but for the purposes of this discussion districts 1 and 2 will be used to test the original conjecture that $p = .6$ for the area covered by districts 1 and 2.

For district 1, the expected number of yes votes is $E(r) = .6(200) = 120$ votes. Therefore, $\chi^2 = (15)^2/120 + (15)^2/80 = 4.69$ with 1 D/F so that $P \cong .030$ by Table V. On the basis of this sample evidence $H_0(p = .6)$ is rejected at the 3 per cent level.

For district 2, the same expected numbers are used because 200 votes were recorded in this sample also. Therefore, $\chi^2 = (20)^2/120 + (20)^2/80 = 8.33$ with 1 D/F so that $P \cong .003$. This time $H_0(p = .6)$ is rejected more decisively.

By the theorem of this section, $\chi^2 = 4.69 + 8.33 = 13.02$, with 2 D/F so that $P \cong .002$ by Table V. Therefore $H_0(p = .6)$ is rejected at the 0.2 per cent level upon the basis of the evidence in the two 200-vote samples.

The chi-square distribution with more than one degree of freedom may be useful when the data are classified in a two-way table of r rows and c columns. For example, a random sample of Republicans and Democrats in a certain city might each be grouped on the basis of three income brackets as follows:

	Annual Income			
Party	Under \$5000	\$5000–\$9999	\$10,000 and Over	Sums
Republican	200	50	8	258
Democrat	120	20	3	143
Sums	320	70	11	401

This will be described as a 2 by 3 contingency table. Earlier in this chapter 2 by 2 contingency tables were analyzed by means of the chi-square distribution.

It is likely that a person might wish to answer the following question with the aid of the data in the above table. Is the announced party affiliation of a voter in this city associated with that voter's economic status? Or, in statistical terminology, the question can be rephrased as follows: Let p_1 be the proportion of Republicans in this city with incomes under \$5000, p_2 be the same for Republicans with incomes in the middle income group, p_3 the same for those Republicans in the highest income bracket; and let p_i' = the corresponding proportions in the population for the Democrats in this city. The subscript i takes the values 1, 2, and 3. Then we wish to test the more complex hypothesis: $H_0(p_i = p_i', i = 1, 2, 3)$. As usual, $\Sigma p_i = \Sigma p_i' = 1$ for $i = 1, 2$, and 3.

The p_i and p_i' are unknown and will be estimated from the sample observations on the assumption that H_0 is correct. These estimates of parameters will be obtained as before: $\hat{p}_1 = \hat{p}_1' = 320/401$; $\hat{p}_2 = \hat{p}_2' = 70/401$; and $\hat{p}_3 = \hat{p}_3' = 11/401$. It follows that the expected number of Republicans in the lowest income stratum is (320/401) (258) = 205.9. The other expected numbers are computed in a similar manner and are shown in parentheses in the following table:

	Annual Income			
Party	Under \$5000	\$5000–\$9999	\$10,000 and Over	Sums
Republican	200(205.9)	50(45.0)	8(7.1)	258(258.0)
Democrat	120(114.1)	20(25.0)	3(3.9)	143(143.0)
Sums	320(320.0)	70(70.0)	11(11.0)	401(401.0)

$$\text{Hence } \chi^2 = \frac{(200 - 205.9)^2}{205.9} + \frac{(120 - 114.1)^2}{114.1} + \cdots + \frac{(3 - 3.9)^2}{3.9} = 2.35.$$

How many degrees of freedom does this sampling chi-square have? In the process of estimating the p_i and the p_i' the expected numbers in a column were required to add to the same sum as the observed numbers for the same columns. This causes the deviations from expectation in a column to be the negatives of each other. For example, $200 - 205.9 = -(120 - 114.1)$. Therefore, both these deviations of observation from expectation cannot be chance occurrences. There are, then, at most three chance deviations among the six which go into the computation of the chi-square. Furthermore, the expected numbers of Republicans in the three income classes must add to 258,

the total number of Republicans in the sample. A similar statement holds for the Democrats, but this is not an independent requirement because the six expected numbers have been forced to total 401 by making the column totals add to the observed numbers 320, 70, and 11. Hence the number of chance differences between the observed numbers and those expected mathematically upon the basis of H_0 is reduced to 2. This, then, is the number of degrees of freedom.

In general, the number of degrees of freedom for a chi-square calculated for an $r \times c$ contingency table is $(r - 1)(c - 1)$. In the example above, $r = 2$ and $c = 3$; hence, $(r - 1)(c - 1) = 2$.

Having decided that the χ^2 of 2.35 has 2 D/F, it remains to determine from Table V if this is an unusual size for a sample chi-square. Table V shows that $P(\chi^2 \geqq 2.35, 2\ D/F) = .31$, approximately; therefore it is entirely reasonable that this χ^2 occurred while sampling from a population for which the hypothesis, H_0, is true. With this sampling result at hand, we would accept the proposed hypothesis.

PROBLEMS

1. Suppose that of 300 salmon which went up a fish ladder in a certain river 185 were chinooks, 65 were silver salmon, and 50 were humpbacks. At another ladder farther south suppose that the following numbers were recorded: chinooks, 150; silvers, 80; and humpbacks, 20. Do these samples (if satisfactorily random, and such is assumed) support the belief that the proportions of these three species are the same at the two locations which were sampled?

2. Referring to problem 1, what matters would cause you to consider them as truly random samples? What factors might cause you to think they were not?

3. Suppose that three independent samplings at one fish ladder led to the following records:

	Number Which Were		
Sampling	Chinook	Not Chinook	Sum
First	60	25	85
Second	70	30	100
Third	52	18	70
Sum	182	73	255

Is the hypothesis that the percentage of chinooks stayed the same during the time of the sampling an acceptable one according to these data?

4. Suppose that some entomologists investigated yellow, short-leaved, and spruce pines in a certain forest to see how many were being seriously attacked

by insects. Assume the following data were obtained from random samples of 250 of each species:

Species	Seriously Damaged	Not Damaged	Sum
Yellow	58	192	250
Short-leaved	80	170	250
Spruce	78	172	250

Do the insects studied attack one of these species more than another? Or is the assumption that the percentage of seriously damaged trees is the same for all these acceptable? *Ans.* Chi-square = 11.47, $2D/F$, $P = .03$. Reject the assumption.

5. For each species of pine studied in problem 4, test the hypothesis that one-third of the trees in each species population are seriously damaged. Then combine these tests by adding the chi-squares, and draw appropriate conclusions.

5.6 CONTROL CHARTS

Sampling techniques appropriate to binomial populations have some important applications in industry in addition to those considered previously in this chapter. During a manufacturing process designed to produce marketable goods it is important to check frequently upon the quality of these products. Quality control charts provide a simple but effective means for watching both the general level of quality and the consistency with which this level is being maintained. No attempt will be made herein to discuss all the various methods in use because books devoted solely to industrial statistics or to quality control are available on this subject. However, it can be seen rather easily that some of the topics presented earlier in this book are fundamental to this subject. The subsequent remarks in this section are intended to point out some of these fundamentals.

Consider first a manufactured item which could be classified as either defective or non-defective with respect to predetermined standards of production. Clearly, a binomial frequency distribution must be involved with some unknown proportion, p, of defective products being manufactured. The number of items inspected and classified as defective or non-defective is the n in the previous discussions of sampling from binomial populations. As indicated earlier the standard deviation of a proportion derived from a sample of n observations is $\sqrt{p(1-p)/n}$. If the manufacturing process is running smoothly with $p = .05$, say, and then something occurs to increase the fraction defective to .15, this occurrence will reveal itself in two ways: (*a*) the

observed fraction defective will tend to increase rather soon, although it might not do so for several samples; and (*b*) the variability among samples will increase if p changes toward 1/2, and this is the case cited above. Both these points are illustrated by means of Table 5.61 and Figure 5.61.

TABLE 5.61

SAMPLES WITH $n = 50$, p STARTING AT .05 AND INCREASING .002 PER SAMPLE FROM THE TWENTY-SIXTH TO THE SEVENTY-FIFTH SAMPLES, INCLUSIVE, AFTER WHICH IT REMAINS AT .15

Sample Number	Fraction Defective	Sample Number	Fraction Defective	Sample Number	Fraction Defective
1	.10	35	.10	68	.28
2	.06	36	.10	69	.14
3	.08	37	.02	70	.08
4	.10	38	.06	71	.12
5	.04	39	.10	72	.12
6	.06	40	.04	73	.14
7	.06	41	.06	74	.16
8	.10	42	.14	75	.12
9	.06	43	.16	76	.22
10	.04	44	.12	77	.20
11	.06	45	.06	78	.12
12	.00	46	.08	79	.14
13	.06	47	.10	80	.14
14	.06	48	.06	81	.12
15	.04	49	.04	82	.14
16	.04	50	.04	83	.10
17	.04	51	.08	84	.20
18	.02	52	.14	85	.12
19	.02	53	.12	86	.14
20	.08	54	.04	87	.04
21	.06	55	.12	88	.16
22	.10	56	.08	89	.18
23	.04	57	.16	90	.18
24	.06	58	.10	91	.14
25	.04	59	.18	92	.16
26	.02	60	.14	93	.08
27	.06	61	.04	94	.18
28	.04	62	.08	95	.20
29	.02	63	.10	96	.24
30	.08	64	.10	97	.16
31	.04	65	.15	98	.10
32	.10	66	.12	99	.06
33	.00	67	.12	100	.14
34	.08				

Table 5.61 and Figure 5.61 were obtained in the following manner: (*a*) p was fixed at .05 by combining 950 green and 50 red beads in a receptacle, and considering the red beads as defective manufactured items; (*b*) 25 samples with $n = 50$ were drawn and the fraction defective was plotted over the order number of the sample; (*c*) starting with the twenty-sixth sample and continuing through the seventy-fifth, two green beads in the receptable were replaced by two red beads after each successive sample was drawn and recorded; and (*d*) starting with the seventy-sixth sample, no additional changes were made. In brief, $p = .05$ for the first 25 samples; p increased

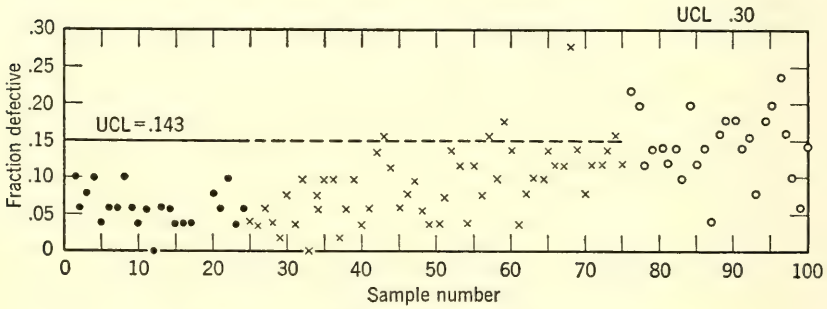


Figure 5.61. Control chart for Table 5.61.

.002 per sample (to simulate slippage or excessive wear) for the next 50 samples so that it finally was .15 for the last 25 samples.

The horizontal lines in Figure 5.61 marked UCL (upper control limit) were obtained from $p \pm 3\sqrt{p(1-p)/50}$ with $p = .05$ until the seventy-sixth sample, and $p = .15$ thereafter. Unless something has occurred unknowingly to change the size of p the fraction defective rarely will go above the UCL; hence, when the observed fraction defective frequently exceeds this limit, it is suspected that the manufacturing process has broken down to some degree. It can be seen in Figure 5.61 that when the "machine" had "slipped" and p began to increase, the fraction defective soon started an upward trend. Shortly, it exceeded the UCL which had been set on the supposition that $p = .05$. Then when p ceased to increase and a new UCL was figured with $p = .15$, the fraction defective again stayed below the UCL. Generally, there also is a lower control limit (LCL), but in this situation it would have been negative and was taken as zero, as is customary.

In practice when the percentage of defectives is unknown the fraction defective observed on at least 25 samples is used in place of p

in the procedure described above. As more samples accumulate a better estimate of p can be computed and used.

It may be that the quality of a manufactured item is judged by means of a measurement such as length, diameter, weight, or a volume which is likely to be a member of a near-normal population rather than a binomial population. The principles involved are the same but now $\bar{x}$ and $s_{\bar{x}}$ (see Chapter 6 for definitions) must be used instead of estimates of p and its standard deviation. In such situations the upper and lower control limits are given, respectively, by:

$$\bar{x} \pm 3\sqrt{(\Sigma x^2)/n(n-1)}\,.$$

Aside from this change, the control charts are constructed and interpreted as before. Of course, sample means, $\bar{x}_i$, are plotted against order of draw.

In view of the fact that the $s_{\bar{x}}$ is somewhat tedious to compute, it has been found to be both satisfactory and economical to use control limits which employ the range as the measure of variation instead of the standard deviation. This procedure and the necessary tables are given and discussed in publications on quality control or on industrial statistics, and will not be given here.

PROBLEMS

1. Another group of samples taken under conditions described for Table 5.61 gave the following results. Make a control chart similar to Figure 5.61 and draw appropriate conclusions.

Sample Number	Fraction Defective	Sample Number	Fraction Defective	Sample Number	Fraction Defective
1	.02	17	.04	33	.06
2	.00	18	.06	34	.08
3	.06	19	.02	35	.10
4	.06	20	.08	36	.04
5	.08	21	.00	37	.04
6	.06	22	.00	38	.02
7	.08	23	.06	39	.08
8	.06	24	.02	40	.08
9	.02	25	.02	41	.00
10	.06	26	.04	42	.10
11	.06	27	.02	43	.10
12	.02	28	.06	44	.08
13	.02	29	.00	45	.16
14	.06	30	.00	46	.14
15	.00	31	.04	47	.10
16	.02	32	.06	48	.16

Sample Number	Fraction Defective	Sample Number	Fraction Defective	Sample Number	Fraction Defective
49	.12	67	.06	85	.12
50	.16	68	.16	86	.14
51	.02	69	.12	87	.12
52	.12	70	.10	88	.22
53	.12	71	.08	89	.08
54	.14	72	.12	90	.18
55	.04	73	.18	91	.10
56	.08	74	.06	92	.12
57	.14	75	.10	93	.18
58	.12	76	.24	94	.06
59	.10	77	.10	95	.10
60	.08	78	.14	96	.06
61	.12	79	.18	97	.06
62	.12	80	.16	98	.10
63	.06	81	.18	99	.22
64	.08	82	.18	100	.06
65	.14	83	.14		
66	.08	84	.08		

2. Calculate an estimate of p from the first 25 samples of problem 1, for the next 25, and for the last 25 samples. Discuss the effect their differences would have on the control chart.

3. Use the estimate of p from the first 25 samples of problem 1 to recompute the UCL for Figure 5.51.

4. Draw 50 successive samples of 50 each from the laboratory population on fraction defective (furnished by the instructor) and construct a control chart from those observations.

5. Perform the operations required in problem 4 for a near-normal population furnished by the instructor.

REVIEW PROBLEMS

1. Which of the following bridge hands are you the more likely to receive on one future random deal?

(*a*) A, K, 10, 4, 3, and 2 of hearts; A, Q, and 10 of diamonds; K, 10, and 3 of spades; and the ace of clubs.

(*b*) No card larger than a 6.

2. Suppose that a coin is so biased that it turns up heads 3 times for each 2 times that it shows tails, on the average. What is the probability that on 8 flips there will be fewer heads than tails? *Ans.* .17.

3. Suppose that you have taken the bid in a bridge game and that you and the dummy have all the trumps except Q, 8, 5, and 2. What is the probability that you would get out all the trumps on successive leads of the A and K of trumps?

4. The prices of barley in the North Central States during 1945 are given below in cents. (*Agricultural Statistics,* 1946, USDA.) Compute two different averages of these prices and discuss their meanings and their limitations.

State:	Ohio	Indiana	Illinois	Michigan	Wisconsin	Iowa	Minnesota
Price:	106	111	111	118	119	103	107

State:	Missouri	N. Dakota	S. Dakota	Nebraska	Kansas
Price:	116	102	103	96	97

5. Determine and interpret the coefficient of variation for the prices of problem 4.

6. Calculate the mean deviation for the prices of problem 4 and discuss its meaning. *Ans.* AD = 6.32.

7. If the egg weights for Rhode Island Red hens are considered to be normally distributed with $\mu = 60.5$ grams and $\sigma = 4.0$ grams, what range of egg weights would you expect to include the middle 90 per cent of all weights?

8. The following table (taken from A. S. Weiner, *Blood Groups and Transfusions,* Thomas, with the consent of the author and the publisher) records the results of a study of the inheritance of the P factor.

Parents' Type	Number of Families	Children's Blood Types P+	P−	Total
P+ × P+	249	677	79	756
P+ × P−	134	286	179	465
P− × P−	34	(4) *	94	98

* Definite doubt established regarding legitimacy.

Recalling that P+ is genetically PP or Pp, and that P− is only pp, test statistically the agreement between the above data on children's blood types and the numbers expected if P+ is assumed to be Pp twice as frequently as it is PP. Consider the (4) * entry as zero.

Ans. $P_1(\chi^2 \geqq 0.33) > .53$; $P_2(\chi^2 \geqq 5.57) = .018$ on P+ × P+ and P+ × P−, respectively.

9. Make up a set of numbers which has an arithmetic mean of 10 and a standard deviation of 2.

10. Is there any evidence in the table of problem 8 for or against the assumption that P+ = Pp twice as frequently as P+ = PP? Explain.

11. Suppose that a large, deep pool in a mountain stream contains a great many trout of just two kinds, rainbow and brook. You wish to learn what percentage are rainbows. Two methods of sampling have been suggested thus far.

(*a*) Fish the pool until 50 trout are caught, and then use this sample evidence as the basis for estimating p.

(*b*) Devise a trap into which the trout will go and be caught, and secure a sample of 50 this way.

Assuming that *b* can be done, which method of sampling, if either, do you recommend as statistically best? Why? Can you suggest a better method than either of these?

12. Referring to the situation of problem 11, suppose that the true proportion of rainbow trout is .40 and that 8 per cent of the rainbows in this area are known to be afflicted with a certain disease. What is the probability that a trout caught at random is a rainbow trout without the disease? *Ans.* $P = .37$.

13. Referring again to problem 11, assume that 60 per cent of the trout in another stream in this same region are brook trout. If on a random sample of 50, 22 are brook trout and the other 28 are rainbows, would you accept or reject the hypothesis that $p = .60$ for brook trout in this stream? Explain.

14. Suppose that there are two mountain streams which run quite close together in a certain area but whose head waters are far apart. You wish to know if the trout populations of these two streams are the same, in a certain well-defined area, as regards proportions of the four species: rainbow, cutthroat, brook, and dolly varden trout. Given the following random sampling data, what would you conclude?

	Number of Trout				
	Rainbow	Brook	Cutthroat	Dolly Varden	Sum
Stream 1	73	68	49	10	200
Stream 2	70	85	80	15	250
Sum	143	153	129	25	450

Ans. Chi-square $= 5.65$ $3D/F$, $P \cong .13$

15. Referring to problem 14, compute and interpret the CI_{95} on the true percentage of brook trout in stream 2.

REFERENCES

Dixon, Wilfrid J., and Frank J. Massey, Jr., *Introduction to Statistical Analysis,* McGraw-Hill Book Company, New York, 1951.

Grant, Eugene L., *Statistical Quality Control,* Second Edition. McGraw-Hill Book Company, New York, 1952.

Neyman, Jerzy, *First Course in Probability and Statistics,* Henry Holt and Company, New York, 1950.

Snedecor, George W., *Statistical Methods Applied to Experiments in Agriculture and Biology,* Fourth Edition, Iowa State College Press, Ames, Iowa, 1946.

Tippett, L. H. C., *Technological Applications of Statistics,* John Wiley and Sons, New York, 1950.

CHAPTER 6

Introductory Sampling Theory for a Normal Population Involving Only One Variable

When the population being sampled has a normal frequency distribution with unknown parameters μ and σ, the problems of estimation and of testing hypotheses by means of samples are fundamentally much the same as those considered in Chapter 5 for binomial populations. Two differences are immediately apparent. (*a*) There now are two unknown parameters instead of one, as for the binomial population, and (*b*) the measurements, X, have a continuous scale of measurement and a continuous frequency distribution. These differences between the normal and the binomial types of populations will appear in the discussions below as the causes of some changes in the mechanics of estimation and of testing hypotheses; but the reader should not lose sight of the fact that the problems and their solutions are much the same as in Chapter 5.

6.1 OBTAINING THE SAMPLE

The process for obtaining good samples from a normal population is similar to that discussed in Chapter 5 for randomization and the avoidance of biases. Here, as there, the population to be sampled must be clearly defined, and the measurement to be taken on the units of this population must be stipulated precisely.

After the population is specified and the units (persons, prices, pigs, plots of land, pots of plants, families, etc.) have been designated unambiguously, it is necessary to devise a method for obtaining the particular units which are to constitute the sample. The sampling situations which come within the scope of this chapter should be handled by completely randomized samples. To illustrate, suppose that a person who is interested in the production of raw rubber wishes

to estimate the percentage rubber in a certain variety of guayule grown under specified environmental conditions. Suppose also that he wishes to select and to analyze 25 plants as a basis for this estimate. The population parameter which is to be estimated is the true average percentage of rubber in plants of the given variety. Assuming that there is a large number of guayule plants from which to select a sample, how should the particular 25 of the sample be chosen? If the 25 tallest, sturdiest, or most thrifty-looking plants were to be chosen they surely would not be representative of the population. If a person were to stroll about among the available plants and choose 25 in what he considered a random manner, he might unconsciously bias the sample. A better way to choose the sample is to assign location numbers (such as row and plant-in-row numbers) to the plants and then effectively "just draw 25 numbers out of a hat." He can use tables of random numbers and similar devices if he chooses. The main point is to see that every plant in the population had at the start an equal and independent chance to be included in the sample.

If two varieties of guayule were compared for percentage of rubber, it might be best to start with a suitable area of land staked off for tree spacings and then assign the varieties at random to the various planting positions. This would make it true that each variety initially had an equal chance for any good, or bad, land among the possible planting positions.

The subject of this section is very broad and complex partly because there are many different sampling situations and a consequent need to devise different sampling procedures to fit these different circumstances. However, as in Chapter 5, only enough is said here to give the reader some general ideas and, perhaps, induce him to do more reading on this subject if he is interested. At the least, the reader can be critical in accepting sampling results presented as information, advertising, or propaganda.

6.2 THE STATISTICAL DISTRIBUTION OF SAMPLE MEANS, $\bar{x}_i$, DRAWN FROM A NORMAL POPULATION

Each sample drawn from a normal population of numerical measurements will nearly always differ from any other sample from the same population in one or more details. Yet certain features of samples from a population, *as a group*, will tend to conform to a predictable pattern. For example, if 10 observations are to be taken

on a normal population with $\mu = 60$ and $\sigma = 10$, no one can say what the arithmetic mean of the sample will be; but a good estimate can be made of its probable size because sample means from such a population will have a frequency distribution over the long run of experience. Therefore, it should be expected that probability statements like those previously discussed herein can be made.

The sample mean, to distinguish it from the unchanging population mean, will be designated by $\bar{x}_i$, where the subscript refers to the ith sample.

The frequency distribution of an approximately normal population with $\mu = 60$ and $\sigma = 10$ is presented in Table 6.21. Six hundred and forty-eight random samples, each containing 10 measurements, were drawn from that population. The arithmetic means of these samples were then computed. The frequency distribution for the 648 sample means also is given in Table 6.21, with the calculated mean ($\bar{\bar{x}}$) and standard deviation ($s_{\bar{x}}'$) of the $\bar{x}_i$ being given at the bottom of the table.

TABLE 6.21

A Frequency Distribution Table for a Near-Normal Population of Measurements X_i, with $\mu = 60$ and $\sigma = 10$; and the Frequency Distribution of 648 Sample Means, $\bar{x}_i$, Taken from That Population with 10 Measurements per Sample

Distribution of Population			Distribution of Sample Means		
Class Interval	f	*r.c.f.*	Class Interval	f	*r.c.f.*
86.1 –90.0	7	1.00	69.1–71.0	0	1.00
82.1 –86.0	13	1.00–	67.1–69.0	7	1.00
78.1 –82.0	31	.99	65.1–67.0	27	.99
74.1 –78.0	64	.97	63.1–65.0	63	.95
70.1 –74.0	113	.92	61.1–63.0	140	.85
66.1 –70.0	169	.85	59.1–61.0	161	.63
62.1 –66.0	218	.74	57.1–59.0	136	.39
58.1 –62.0	241	.59	55.1–57.0	72	.18
54.1 –58.0	222	.43	53.1–55.0	34	.06
50.1 –54.0	176	.28	51.1–53.0	8	.01
46.1 –50.0	120	.16	49.1–51.0	0	.00
42.1 –46.0	69	.08		—	
38.1 –42.0	34	.04	Total	648	
34.1 –38.0	15	.02			
30.0*–34.0	8	.01			
Total	1500				
$\mu = 60$	$\sigma = 10$		$\bar{\bar{x}} = 59.98$	$s_{\bar{x}}' = 3.14$	

* This interval was extended by 0.1 to include the remaining measurement in the population.

The frequency distribution in the right-hand part of Table 6.21 is an approximation to an infinite population of $\bar{x}_i$ which would result if this sampling with $n = 10$ observations were continued indefinitely. Any *one* random sample from the normal population described above necessarily would be a member of the population of $\bar{x}_i$.

As an illustration of the preceding discussion, suppose that an agricultural economist is interested in learning if the per-acre income on a certain type of farm employing good (recommended by an agricultural experiment station, for example) farming practices is greater, on the average, than that for farmers not following those practices. He takes a sample of n farms on which the recommended practices are employed and calculates the mean per-acre income. The same is done for a comparable random sample of farms on which these recommendations are not followed. Some measurement of the consistency of income on each of the two groups of farms also would be needed. If the newer practices are worth recommending to replace those currently in use, they must produce a new population of per-acre incomes with a larger mean, a smaller variance, or both. To obtain information on these points, the economist must have adequate information regarding the manner in which sample means are distributed; that is, where their region of concentration will be, and how they will tend to be dispersed about that region of concentration. Hence, the first objective of this chapter will be to provide that sort of information about $\bar{x}$'s drawn from the same normal population of numerical measurements—such as per-acre incomes.

Figure 6.21 presents the graphs of the frequency distributions shown in Table 6.21. The larger curve is for the near-normal parent population of X's, while the smaller curve is taken as a good approximation to the distribution of the population of $\bar{x}$'s obtained from samples of ten observations taken from the population of X's.

It appears from Figure 6.21 that the two frequency distributions are much alike in general form, and seem to be approximately normal about the same mean. The major difference lies in the fact that the $\bar{x}_i$ exhibit much less variability than the X's of the population from which the samples were taken. This is to be expected because one important reason for combining a number of individual X's into one sample is to achieve a smoothing out of the individual differences among those X's.

It should be noted from the bottom of Table 6.21 that the mean of the $\bar{x}_i$ is 59.98, which is quite near to 60, the size of this population mean, μ. Also, the standard deviation of the 648 sample means is 3.14, which is a bit less than one-third of σ. As a matter of fact, 3.14

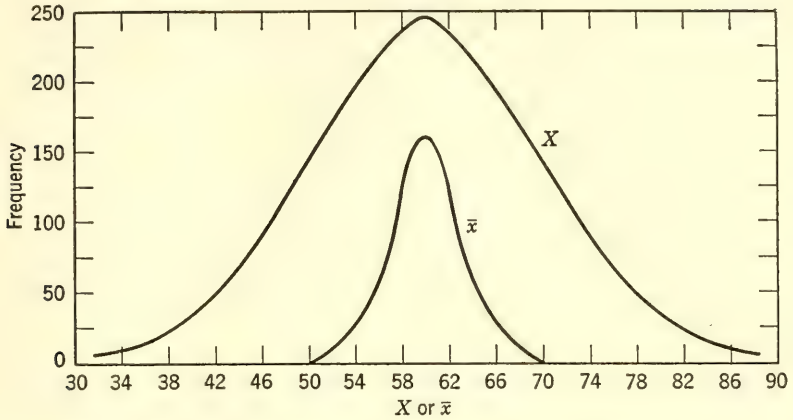


Figure 6.21. Frequency distribution curve for a normal variate, X, and also for the sampling mean, $\bar{x}$, for samples with $n = 10$.

is very nearly equal to $\sigma/\sqrt{n} = 10/\sqrt{10} = 3.16$, to two decimals, where the symbol n is used to denote the number of observations taken in the sample.

The preceding discussion has suggested three features which are exhibited by a large number of sample means obtained from a normal population of numerical measurements. These features are:

(*a*) Although it is impossible to predict the actual content of a particular future sample it may be possible to predict the type of frequency distribution which the sample means will follow, for example, a normal distribution.

(*b*) The average sample mean will be of essentially the same magnitude as the mean of the population sampled.

(*c*) The sample means, $\bar{x}_i$, will display less variability than the X's of the population. It is logical that the variability of the sample means—from sample to sample—should decrease as the size of the samples increases. It was suggested that a factor $1/\sqrt{n}$ is involved here.

The following theorem is given without proof because that proof is inappropriate to this book. The theorem is stated here for the purpose of replacing the indefiniteness of statements (*a*), (*b*), and (*c*) above with precise information which can be used in practice.

Theorem. If a population of numerical measurements, X, conforms to a normal frequency distribution with mean, μ, and stand-

ard deviation, σ, and if a very large number of random samples of n observations each is drawn from that population:

(*a*) The population of $\bar{x}_i$ thus formed will have a normal frequency distribution.

(*b*) The mean of the $\bar{x}_i$ will be μ also.

(*c*) The standard deviation of the $\bar{x}_i$ will be $\sigma/\sqrt{n}$.

Note from this theorem that, if the X's are normally distributed we automatically know the form of the distribution of the sample means and hence can write down specifically the formula for their distribution curve, namely;

$$y_1 = \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{n})} \cdot e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}} \tag{6.21}$$

Formula 6.21 can be transformed into the standard normal formula of Chapter 4 by means of the substitutions

$$y = y_1 \cdot \sigma/\sqrt{n} \quad \text{and} \quad \lambda = (\bar{x} - \mu)/(\sigma/\sqrt{n})$$

whereupon Table III can be employed as shown earlier.

If ten measurements are taken per sample and the parent population has a mean of 60 and a standard deviation of 10, as above, the mean of the $\bar{x}$'s is also 60, and the standard deviation is $10/\sqrt{10} = \sqrt{10} = 3.16$. Table 6.21 and Figure 6.21 furnish approximate verification of these statements from actual experience.

If $n = 15$ and the samples are drawn from a normal population with $\mu = 60$ and $\sigma = 10$, the mean of the resulting population of $\bar{x}_i$ also will be 60, and the standard deviation will be $10/\sqrt{15}$. On 200 such samples, their mean was 60.22 and their standard deviation was 2.53 instead of the expected 60 and 2.58, respectively. Two hundred is a relatively small number of samples from which to seek empirical verification of mathematical theory, but these results do agree quite well with the theorem given above.

Problem 6.21. If chemical determinations of the percentage of protein in samples of a certain variety of wheat are known to have a normal frequency distribution with $\mu = 14$ and $\sigma = 2$, what is the probability that five random samples will have a mean per cent protein above 16?

In the following discussion, $\sigma_{\bar{x}}$ will be used to denote the standard deviation of the population of sampling means. In this problem, $n = 5$, $\mu = 14$, and $\sigma_{\bar{x}} = 2/\sqrt{5} = 0.90$. Therefore, $\lambda = (16 - 14)/0.90 = 2.22$; and $P(\lambda \geq 2.22) = .013$, approximately. In other words, only about 13 times in 1000 sets of 5 observations like these would you

have the mean per cent of protein at or above 16. The rarity of such an occurrence might cause you to doubt the accuracy of the protein analyses and cause you to ask that they be done over.

PROBLEMS

1. Given that a certain population of measurements is normally distributed about a mean of 30 and with a standard deviation of 8. If a sample of 16 members is to be drawn at random, what is the probability that its mean will be below 28?

2. Under the conditions of problem 1, what is the probability that a sample of 9 numbers taken from that population will have an arithmetic mean below 28? *Ans.* .23.

3. Solve problem 2 with the standard deviation changed to 12.

4. If in some particular area the daily wages of coal miners are normally distributed with $\mu = \$15$ and $\sigma = \$1.50$ what is the probability that a representative sample of 25 miners will have an average daily wage below \$14.25? *Ans.* .006.

5. Suppose that a thoroughly tested variety of corn has been found to yield an average of 35 bushels per acre with a standard deviation of 6, and that these yields have a normal frequency distribution. If a random sample of 25 yields for a new variety gives $\bar{x} = 40$, show that there is good reason to believe that the yields of the new variety are from a population with a mean higher than the 35 bushels per acre for the population of the older variety.

6.3 ESTIMATION OF THE UNKNOWN MEAN AND VARIANCE OF A POPULATION FROM THE INFORMATION CONTAINED IN A SAMPLE

If the parameters μ and σ are unknown for a particular normal population which is being sampled (and they usually are or there would be no occasion for sampling) it becomes necessary to estimate them from the X's taken in the sample. How should this be accomplished? Although this really is a mathematical problem whose solution lies beyond the scope of this book, certain desirable requirements for sampling estimates of μ and σ^2 can be considered.

First, it seems logical that an acceptable estimate should have a mean equal to the corresponding population parameter after many samples have been taken. Even though only one sample of n measurements is to be taken, we usually would like to know that the $\bar{x}$ and s^2 we shall obtain as estimates of μ and σ^2 are from populations whose means are μ and σ^2, respectively. Sampling estimates which satisfy this requirement are called *unbiased estimates*, as noted in Chapter 5.

The second—and more important—requirement which we should impose on a sampling estimate is that it be as reliable as possible in

the sense that it have a relatively small variance from sample to sample. For example, suppose two methods of estimating μ each will produce an unbiased estimate but, over many samples, one has a variance of 100 whereas the other has a variance of only 25. The latter estimate obviously is more consistently near μ in size, and hence less allowance need be made for sampling error in this estimate. This second estimate would be considered a *more efficient estimate* than the one whose variance was 100.

The estimate $\bar{x}$ of μ, which already has been mentioned, and whose symbolic definition is

$$\bar{x} = \frac{\Sigma(X)}{n}, \tag{6.31}$$

gives an unbiased and highly efficient estimate of μ. It has been pointed out earlier in a theorem that the variance of $\bar{x}$ under repeated sampling is only one nth of the population variance, when a normal population is being sampled. (As a matter of fact, the variance of $\bar{x}$ is σ^2/n for any population if σ^2 is finite.) Hence the $\bar{x}$ is widely used as an estimate of μ.

The variance σ^2 will be estimated by means of the formula

$$s^2 = \frac{\Sigma(X - \bar{x})^2}{n - 1}. \tag{6.32}$$

This estimate is unbiased and is considered to be about as efficient as any estimate of σ^2 as long as the sample is not extremely small. The usefulness of this estimate in practice will be illustrated repeatedly in subsequent discussions.

By comparison with the methods used in Chapter 2 to compute σ or σ^2, it is seen in formula 6.32 that two changes have been made. The μ is replaced by $\bar{x}$ and the denominator is now $(n - 1)$ instead of n. Logically, the $\bar{x}$ must be used because μ is unknown; but it also must be recognized that the differences, $(X_i - \bar{x})$, are more dependent upon chance events which occur in the process of sampling than were the quantities, $(X_i - \mu)$. The $\bar{x}$ itself is subject to sampling error whereas the μ is a fixed number for a given population. This matter is taken into account in sampling theory. One step in this direction is to associate with each estimated variance a *number of degrees of freedom*. The estimate s^2 of formula 6.32 is said to be based on $n - 1$ degrees of freedom because only $(n - 1)$ of the n differences $(X_i - \bar{x})$ are actually chance differences. This follows from the fact that $\Sigma(X - \bar{x}) = \Sigma X - \Sigma\bar{x} = n\bar{x} - n\bar{x} = 0$. Hence, given *any* $n - 1$ of

the deviations of the sample X's from their mean, $\bar{x}$, the other deviation can be computed without any risk of error. If the true mean, μ, were known, the n $(X_i - \mu)$'s would all be quantities whose specific sizes depended on chance, and σ^2 could be estimated with n degrees of freedom, which is one more than s^2 has. Also, the estimate made with μ known would be more reliable than s^2, a fact which is associated with its greater number of degrees of freedom.

As soon as a satisfactory method is available for the estimation of σ^2, it follows that the standard deviation of $\bar{x}$—which is $\sigma/\sqrt{n}$ and is symbolized as $\sigma_{\bar{x}}$—also can be estimated from the following quantity:

$$s_{\bar{x}} = s/\sqrt{n} = \sqrt{\Sigma(X - \bar{x})^2/n(n-1)}\,, \tag{6.33}$$

which is calculated from the observations taken in the sample. It still is true—as for all sampling estimates—that $s_{\bar{x}}$ is variable from sample-to-sample.

Although $\bar{x}$ is the best specific estimate of the population mean, μ, it is preferable to calculate from the sample an interval in which we can expect the true mean to lie, with a measurable degree of confidence in this expectation. The so-called *point estimate*, $\bar{x}$, is almost never exactly right, but an interval can be defined in such a way that we can attach a measure of confidence to the statement that μ lies in this interval. This problem can be solved by means of a ratio which is analogous to the $(X - \mu)/\sigma$ which was studied in Chapter 4. That ratio involves only one variable, X, and follows the normal distribution. So also does the ratio $(\bar{x} - \mu)/\sigma_{\bar{x}}$. If the standard deviation, σ, is not known—which is the usual sampling situation—the corresponding ratio

$$t = (\bar{x} - \mu)/s_{\bar{x}} \tag{6.34}$$

involves a variable denominator and is not normally distributed. Its degree of departure from normality depends on the size of the sample, n, because the denominator is much less variable for the larger samples.

Mathematicians have derived a formula for the frequency distribution of the ratio, t, for a sample of any size. Although that derivation is not appropriate to this book, sampling experience will provide an approximation to this distribution, and then mathematical tables will be provided which give the same information more accurately and more easily.

Table 6.31 presents the frequency and the *r.c.f.* distributions of 580 sampling t's obtained from random samples drawn from the near-normal population of Table 6.21. All samples contained $n = 10$

observations. The t's were computed from formula 6.34, using $\mu = 60$ and the $\bar{x}$'s obtained from the samples. For example, if the $\bar{x} = 58.2$ and $s_{\bar{x}}$ is calculated from formula 6.33 to be 2.61, $t = (58.2 - 60)/2.61 = -0.69$.

TABLE 6.31

OBSERVED FREQUENCY DISTRIBUTION OF 580 t_i OBTAINED FROM SAMPLES OF 10 MEMBERS EACH DRAWN FROM A NORMAL POPULATION WITH $\mu = 60$ AND $\sigma = 10$

Class Interval	f	$r.c.f.$	Class Interval	f	$r.c.f.$
≥ 3.60	1	1.00	-2.80 to -2.01	17	.04
2.80 to 3.59	4	1.00−	-3.60 to -2.81	3	.01
2.00 to 2.79	16	.99	< -3.60	1	.00
1.20 to 1.99	54	.96			
0.40 to 1.19	127	.87	Total	580	
-0.40 to 0.39	187	.65			
-1.20 to -0.41	115	.33	Arithmetic mean = +0.015		
-2.00 to -1.21	55	.13	Standard deviation = 1.10		

Figure 6.31 presents the frequency and the relative cumulative frequency distribution curves corresponding to Table 6.31. The *r.c.f.*

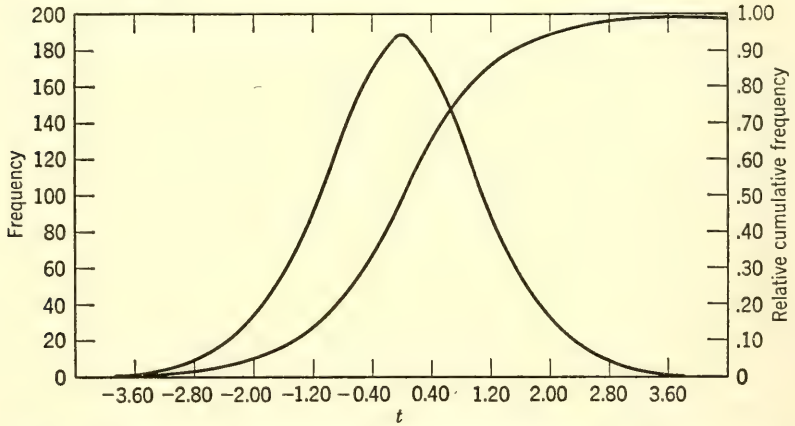


Figure 6.31. Frequency distribution of 580 sample values of t drawn from a normal population with $\mu = 60$, $\sigma = 10$, and $n = 10$.

curve of Figure 6.31 furnishes information concerning the population of t_i for $n = 10$ which is entirely analogous to that to be had from Table III for normal frequency distributions. Figure 6.31 shows that: (*a*) The point where $t = 0$ on the horizontal axis divides the population of t's into two equal portions, each containing 0.50, or 50 per cent, of the whole population (as with the normal distribution

at $\lambda = 0$). (*b*) Approximately 95 per cent (as nearly as can be told from the graph) of the t_i are less than or equal to $+2$ in magnitude. (*c*) The middle 80 per cent of the t's with $n = 10$ fall within the limits -1.5 to $+1.5$, approximately. Such information will be seen to be needed in arriving at the interval estimate for μ described above.

It should be noted that conclusions (*b*) and (*c*) of the preceding paragraph referred only to samples with $n = 10$. The general effect

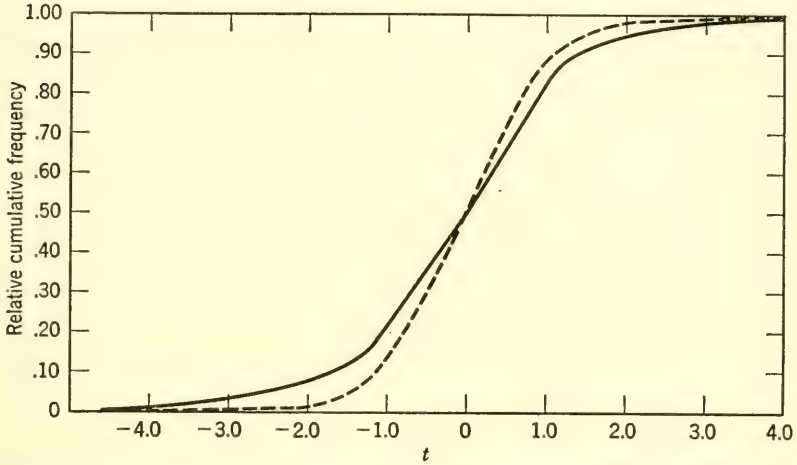


Figure 6.32. Relative cumulative frequency distributions for t when $n = 5$ (solid line) and for $n = 25$ (broken line).

of the magnitude of n on the frequency distribution of the t's is illustrated in Figure 6.32 for $n = 5$ and $n = 25$. The larger the sample size, the less dispersed are the t's. In fact, after n becomes as large as 25 it is difficult to detect much difference between the *r.c.f.* curve for t and that for a normally distributed measurement. Also, if n_1 is smaller than n_2, the ogive for samples with n_1 observations will be above that for samples with n_2 observations for negative t's and below it for positive t's. This is just a graphic verification of the fact that the t's are more dispersed for the smaller-sized samples.

In line with the earlier discussion of degrees of freedom for the estimate of the standard deviation, the t is said to have the same number of degrees of freedom as the standard deviation in the denominator of this ratio. The t's considered so far have one less degree of freedom than the size of the sample, that is, $n - 1$.

Table IV provides an *r.c.f.* distribution of the sampling ratio, t, for most of the commonly used sizes of samples. This table is in the form of those *r.c.f.* distributions discussed in Chapter 2. This form is different from that found in most statistical tables, but the form of Table IV fits the purposes of this book better than the traditional table. However, the values in the more usual table can be derived from Table IV quite easily. For example, by Table IV the probability that a random sampling t will have a size below -2 from a sample of $15(14D/F)$ is seen to be .033. Because the t-distribution is symmetrical about $t = 0$ the probability that a t computed with $14\ D\ F$ will exceed 2 *numerically* is twice .033, or .066. This is the probability given in the usual table for $t = 2$ and 14 degrees of freedom. To obtain such a number as .066 for P in those tables we must interpolate because they give the sampling t's which correspond to specified values of P.

Table IV will be employed in subsequent discussions instead of the *r.c.f.* curve because it is both more accurate and more convenient to do so. However, the reader should remember that the two methods are basically the same. The use of tables for the t-distribution is especially advantageous because there would have to be a different *r.c.f.* graph for each number of degrees of freedom.

Suppose, now, that the true population mean, μ, is not known. In spite of our ignorance of the size of μ it remains true that sampling values of t will conform to the t-distribution. For example, for $n = 10$(9 degrees of freedom), it will be true that 92 per cent of the t's will lie between -2 and $+2$ (see Table IV). Or, put in terms of a mathematical inequality, it remains true that the following statement is correct for 92 per cent of a very large number of samples with $n = 10$:

$$-2 \leq \frac{\bar{x} - \mu}{s_{\bar{x}}} \leq +2. \tag{6.35}$$

Approximate empirical verification of the truth of this inequality is found in Table 6.31 above.

In view of the information just given, the following can be said: If we are about to take a random sample of 10 numerical measurements from a normal population, the probability is .92 that the t for this sample will satisfy inequality (6.35) because 92 per cent of all samples with 9 degrees of freedom do lie within the limits -2 to $+2$. When the μ is not known this statement still is true but we can compute the t only *in terms of the* μ. To illustrate, suppose that a random sample

of 10 observations taken from a normal population has given $\bar{x} = 8$ and $s_{\bar{x}} = 2$. Then $t = (8 - \mu)/2$, a function of μ. Before the sample was taken it could be reasoned that there were 92 chances in 100 that the t to be obtained would have some size between -2 and $+2$. Likewise, after the sample is taken, the assumption that the t does lie within these limits runs a risk of 8 in 100 of being wrong as a result of sampling variation.

What does the assumption that the t obtained from the sample satisfies the inequality (6.35) require of μ now that $t = (8 - \mu)/2$? The quantity $(8 - \mu)/2$ must be at least as large as -2 but no larger than $+2$; therefore, $(8 - \mu)$ must be at least as large as -4 but not larger than $+4$. It follows that μ must be some number from 4 to 12 unless an 8-in-100 event has occurred. We never actually know in practice if such a t has been got; but we do know that the odds against it are 92:8.

The probability, .92, associated with the expression (6.35) is called the *confidence coefficient* for the *confidence interval* 4 to 12 because it measures the confidence we can put in the inference that μ lies within these limits. This usage is identical with that of Chapter 5. That is, a method for basing decisions on sampling evidence has been presented; and, although we know it is not infallible, we know what risk of error we run when we use the method.

Obviously, other confidence coefficients besides .92 could be used. For example, 95 and 99 per cent confidence limits are quite common. They require the use of the following inequalities for 9 degrees of freedom:

$$-2.26 \leq \frac{\bar{x} - \mu}{s_{\bar{x}}} \leq +2.26 \text{ for 95 per cent confidence limits, } \mathrm{CI}_{95}.$$

$$-3.25 \leq \frac{\bar{x} - \mu}{s_{\bar{x}}} \leq +3.25 \text{ for 99 per cent confidence limits, } \mathrm{CI}_{99}.$$

These two inequalities and that of (6.35) can be put into a more convenient form simply by multiplying through each (all three members) by $s_{\bar{x}}$ and then transferring the $\bar{x}$ to the outer members of the inequalities. The final results for 92, 95, and 99 per cent confidence intervals are as follows for 9 degrees of freedom:

$$(\bar{x} - 2s_{\bar{x}}) \leq \mu \leq (\bar{x} + 2s_{\bar{x}}) \text{ for a } \mathrm{CI}_{92}; \tag{6.36}$$

$$(\bar{x} - 2.26s_{\bar{x}}) \leq \mu \leq (\bar{x} + 2.26s_{\bar{x}}) \text{ for a } \mathrm{CI}_{95}; \tag{6.37}$$

$$(\bar{x} - 3.25s_{\bar{x}}) \leq \mu \leq (\bar{x} + 3.25s_{\bar{x}}) \text{ for a } \mathrm{CI}_{99}. \tag{6.38}$$

The latter two are used quite commonly for estimates with 9 degrees of freedom; the first inequality is used here chiefly for convenience of illustration.

As an application of the above-described methods, suppose that you wish to learn what the average life of a certain type of light bulb is. Suppose that ten sample bulbs of this type are left burning until all have burned out, and the time it took each to burn out is recorded. The

TABLE 6.32

Outline of Some Samples from a Near-Normal Population, with $\mu = 60$ (Samples were taken by statistics classes from the population of Table 6.21.)

Sample Number	$\bar{x}$	s	t	Confidence Limits on 80%	90%	95%
1	58.4	6.04	−0.86	55.7–61.0	54.8–61.8	54.0–62.7
2	60.9	12.64	0.24	55.4–66.5	53.6–68.2	51.5–70.4
...	....	.....	....	.....	.....	.....
4	56.4	5.81	−1.96	53.8–59.0*	53.0–59.8*	52.2–60.6
5	59.0	11.09	−0.27	54.2–63.9	52.6–65.4	51.1–67.0
6	53.4	10.33	−2.02	48.9–57.9*	47.4–59.4*	46.0–60.8
7	52.9	6.13	−3.66	50.2–55.6*	49.3–56.5*	48.5–57.3*
8	67.7	10.21	2.38	63.2–72.2*	61.8–73.6*	60.4–75.0*
9	54.0	10.52	−1.80	49.4–58.6*	47.9–60.1	46.5–61.5
10	54.1	8.75	−2.13	50.3–57.9*	49.0–59.2*	47.8–60.4
. . . . .						
301	58.4	7.22	−0.70			
302	61.8	12.67	0.45			
303	63.2	7.69	1.32			
304	61.6	8.04	0.63			
305	58.4	10.69	−0.47			
306	64.6	13.39	1.09			
307	64.5	6.33	2.25	*	*	
308	55.9	7.65	−1.69	*		
309	60.0	15.19	0			
310	55.5	12.95	−1.10			
. . . . .						
578	60.2	7.94	0.08			

Summary of 578 Confidence Intervals

Confidence Coefficient	Limits Did Include μ Number	%	Limits Did Not Include μ Number	%
.80	460	79.6	118	20.4
.90	517	89.4	61	10.6
.95	559	96.7	19	3.3

following results will be assumed to have been obtained: $\bar{x} = 1400$ hours and $s_{\bar{x}} = 70$ hours. The inequalities above become the following after simplification: $1260 \leq \mu \leq 1540$, $1242 \leq \mu \leq 1558$, and $1172 \leq \mu \leq 1628$ hours, respectively, if the computations are rounded to the nearest whole hour. If you act on the assumption that the true average life of this type of bulb is between 1260 hours and 1540 hours, you run a risk of 8 in 100 that the sample has misled you. However, if the widest limits, 1172 to 1628, are used, the risk of an erroneous assumption is only 1 in 100.

Table 6.32 has been included to illustrate further and to clarify the idea of confidence intervals. It contains some sampling results obtained from a normal population with $\mu = 60$, and the $n = 10$. A summary of 578 samples is shown at the bottom of the table. Not all the sampling results are given; just enough to satisfy the purposes of this discussion. The asterisks indicate those intervals which fail to include μ.

Some of the points which are illustrated by Table 6.32 are the following:

(*a*) The confidence coefficients are long-run relative frequencies which are verified only after a large number of samples. If attention were confined to samples 4 to 10, the confidence coefficients would seem to be wrong; but over the set of 578 confidence intervals, they are verified quite satisfactorily.

(*b*) The determination of a confidence interval is doubly dependent on chance: once as regards the mean, and again regarding the magnitude of the standard deviation. For example, samples 306 and 307 had essentially the same mean but the standard deviations were, by chance, so different that even the 80 per cent limits from sample 306 included the true mean, 60. Only the 95 per cent confidence interval from sample 307 includes μ. On the other hand, samples 303 and 308 have practically the same standard deviation, but the sample means are so different that the 80 per cent limits from sample 308 failed to include the true mean.

(*c*) The confidence interval is wider for the larger confidence coefficients, that is, the more certain we choose to be in our conclusions, the more room we must leave for sampling variations.

Problem 6.31. Suppose that a highway commission is interested in the strength of concrete which it wishes to make for highway projects, and that it concludes that the 7-day tensile strengths of standard samples will be the best

criterion of quality. Suppose also that ten of the standard testing models gave these results:

$$\bar{x} = 439.0 \text{ pounds per square inch}, \quad s = 47.0 \text{ pounds per square inch}.$$

What valid and useful conclusions could they draw concerning the true average tensile strength of this concrete?

Although the true average strength, μ, is a hypothetical strength rarely possessed by an actual sample, it does provide a useful description of the tensile strength of a type of concrete. Before a confidence interval can be put on μ a confidence coefficient must be chosen. Such matters as the seriousness of committing an error, and the added cost of demanding narrower limits, are involved in this decision. However, for purposes of illustration it will be assumed that a risk of 1 in 20 of obtaining a confidence interval *not* including μ is appropriate to these circumstances. Then, using inequality (6.37) because $n = 10$ and 95 per cent limits are sought, we obtain the following:

$$439 - 2.26(14.9) \leq \mu \leq 439 + 2.26(14.9)$$

because $s_{\bar{x}} = 47.0/\sqrt{10} = 14.9$ pounds per square inch. When this inequality is simplified it is found that the 95 per cent confidence interval is

$$405 \text{ pounds per square inch} \leq \mu \leq 475 \text{ pounds per square inch},$$

to the nearest 5 pounds. Therefore, the true average tensile strength of this concrete will be considered to be somewhere between 405 and 475 pounds per square inch; but, at the same time, it will be kept in mind that there is 1 chance in 20 that this sample has been "wild" and hence has led to an incorrect conclusion.

If the reader thinks a bit about the material in this section as compared to the corresponding section in the preceding chapter on binomial populations, it should become apparent that these two sections have a great deal in common. In both, a sampling distribution was studied, and we were concerned with the relative frequencies with which certain sampling phenomena would occur. In particular, we were interested in the relative frequencies with which intervals determined from samples would include the unknown population parameter. This probability is the confidence coefficient.

There also are some differences which could be pointed out. A major one is that owing to the discontinuity of the binomial frequency distribution, the confidence coefficient is the lower limit on the rela-

tive frequency with which the confidence interval will include the parameter. Basically, however, the methods of these two chapters involve the same kind of statistical inference.

It may have occurred to the reader to wonder why the confidence interval is taken in the center of the sampling distribution. Although it is true that 92 per cent of all sampling t's with 9 degrees of freedom will have sizes between -2 and $+2$, it is also true that 92 per cent of all sampling t's with 9 degrees of freedom will lie between -5 and $+1.54$ (see Table IV). Therefore, the inequality

$$-5.0 \leq \frac{\bar{x} - \mu}{s_{\bar{x}}} \leq +1.54$$

also will be true for 92 per cent of all samples with 9 degrees of freedom. Why not use this inequality as the basis for computing the 92 per cent confidence interval instead of the one suggested earlier? Suppose the inequality above is used on the example used previously in which $\bar{x} = 8$ and $s_{\bar{x}} = 2$. The 92 per cent confidence interval now is from 5 to 18 instead of the shorter interval, 4 to 12, obtained previously. It always will be longer when a non-centrally located interval on t is used. It should be clear that the shorter the confidence interval for a given confidence coefficient, the better the interval estimate. Why be more indefinite than is necessary?

PROBLEMS

1. Verify the 80, 90, and 95 per cent confidence intervals given in Table 6.32 for samples 1 and 2.

2. Compute 99 per cent confidence limits on μ for samples 8 and 9 of Table 6.32 and interpret them. *Ans.* $57.2 \leq \mu \leq 78.2$. $43.2 \leq \mu \leq 64.8$.

3. Given that $\bar{x} = 35$ and $s = 10$ for a sample of ten observations compute and interpret the 95 per cent confidence interval on μ. Do the same for $n = 15$ and $n = 20$ and compare them. What is the implication regarding the relation between the size of the sample and the width of the confidence interval, everything else being equal?

4. Given that the t was computed to be -2.08 for sample number 525, Table 6.32, determine whether or not the 90 per cent confidence interval includes μ. Do likewise for 95 per cent limits.

Ans. CI_{90} does not include $\mu = 60$; CI_{95} does.

5. Use Figure 6.31 to determine the 86 per cent confidence interval on μ from sample 6 of Table 6.32.

6. Suppose that an improved method of cultivating wheat has produced an average of 5 bushels per acre more yield than an older method on a sample of 21 plots. Also assume that the standard deviation on this sample is $s = 5$ bushels per acre. What are the 95 per cent confidence limits on the true aver-

age additional yield produced by the new method? Suppose that the new method costs \$5 per acre more to use than the older method. What can you say about the probable economic advantage obtained from the new method if wheat is currently bringing \$2.25 per bushel?

Ans. CI_{95}: $2.65 \leq \mu \leq 7.35$ bushels. Gain ≥ 96 cents per acre.

7. Suppose that in problem 6, s had been 10 bushels per acre. Show how this increase in sampling variability among the 21 plots changes the answers to the questions asked in problem 6.

8. Suppose that in problem 6, the sample had involved but 10 plots. Show how this decrease in the size of the sample changes the answers to the questions asked in problem 6.

9. Suppose that chemical analysis shows that the mean per cent protein for 16 wheat samples is 14.28 and that the estimated standard deviation for the population of $\bar{x}$'s being sampled is $s_{\bar{x}} = 2.00$. What conclusions can you draw from the 99 per cent confidence interval on the true mean μ?

10. If basal metabolisms determined for a random sample of 25 sixteen-year-old Kansas girls produced $\bar{x} = 45.80$ calories per square meter per hour, with $s = 0.50$, what are the 80 per cent confidence limits, and what information do they provide in setting up a standard for sixteen-year-old Kansas girls?

Ans. CI_{80}: $45.67 \leq \mu \leq 45.93$ calories per square meter per hour.

11. Suppose that during a recent period of strong prices twenty-five 450-pound choice steer calves were purchased, October 15th, wintered on silage and one pound of cottonseed meal per day, and then sold on April 15th as choice stocker steers. If the average net income per animal was $\bar{x} = \$25$ with $s = \$10$, place a 90 per cent confidence interval on the true average net income per animal for the population so sampled. A similar sampling of choice 600-pound yearling steers produced a 90 per cent confidence interval of \$105 to \$130 net income per steer. What conclusions can you draw regarding the most profitable choice for a cattleman to make between these two systems?

6.4 A STATISTICAL TEST OF A HYPOTHESIS THAT A GIVEN SAMPLE CAME FROM A NORMAL POPULATION WITH A SPECIFIED MEAN

The general problem of deciding whether or not a particular sample came from a normal population whose mean, μ, is specified but whose standard deviation can be estimated only by means of s has received consideration earlier in this chapter. In practice, the specification of μ is based upon a hypothesis about the population under study. For example, if a new method of cultivating wheat does *not* produce higher average yields the population of differences in yield between the new and old methods grown in a series of paired plots of land will have a true mean $\mu = 0$ because, on the average, there is no advantage to the new method. If the hypothesis that $\mu = 0$ is found from statistical analysis to be unreasonable in view of the

sampling results, that hypothesis should be rejected. However, if the evidence in the sample is in reasonable accord with that hypothesis, it should be accepted. This is the idea behind the methods to be presented in this section, and also of all so-called tests of significance.

A generally satisfactory solution to the problem of this section can be obtained from the t-distribution when normal or near-normal populations are being sampled. As the reader already knows, $t = (\bar{x} - \mu)/s_{\bar{x}}$ and has $n - 1$ degrees of freedom if the sample contains n observations. When the μ is specified by the hypothesis to be tested, the t can be calculated. Thereafter, we can determine from Table IV how

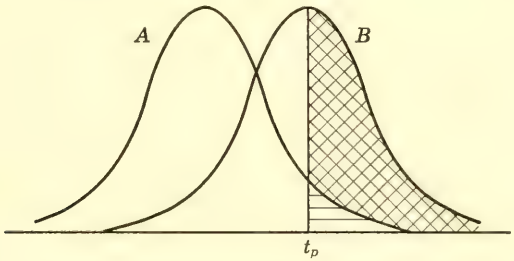


Figure 6.41. Illustration of the effects on the t-distribution of a false hypothesis regarding μ.

uncommon such a t is when the samples are drawn from the supposed population. For example, if $n = 14$ and t turns out to be 0.90 on the assumption that $\mu = 0$, we learn from Table IV that about 38 per cent of all sampling t's with 13 degrees of freedom are numerically larger than 0.90. Therefore, this value of t is not at all unusual and hence we would have no reason to doubt the hypothesis that $\mu = 0$. But, suppose that t had been 3.0. It is seen in Table IV that only about one t in 100 from samples of this size ever gets as large as 3, numerically. Hence, we might reasonably doubt that μ really is zero because t rarely attains such a size when the hypothesis being tested is true.

To illustrate the above discussion graphically, suppose that the true mean, μ, of a normal population of measurements actually is 2 but owing to some error in reasoning μ is considered to be 0. What effect does this have on the frequency distribution of t? For this situation, t really is $(\bar{x} - 2)/s_{\bar{x}}$ but because of the error regarding μ the values of t are calculated from the formula $t = \bar{x}/s_{\bar{x}}$. In view of the fact that $t_1 = (\bar{x} - 0)/s_{\bar{x}}$ is just $2/s_{\bar{x}}$ units larger than $t_2 = (\bar{x} - 2)/s_{\bar{x}}$, we are actually sampling population B of Figure 6.41, but we think that we are sampling from population A. The discrepancy should, and would,

show up through an excess of large t's beyond the proportions predictable from Table IV. To be more specific, suppose that attention is centered upon a particular t such that p per cent of the t-distribution lies to the right of this point. Such a point is indicated in Figure 6.41 as t_p. It is noted from this figure that a much larger fraction of the true t's (figure B) lie to the right of t_p than is true for the population resulting from the calculations with the false value for μ. This discrepancy between the hypothetical and the actual situation will show up in the sampling. Obviously, the greater the discrepancy the more easily it is detected by sampling.

In practice it is not feasible, efficient, or economical to continue to draw samples from a population until the evidence for or against a certain hypothesis is so overwhelming that there is virtually *no* doubt of its truth or of its falsity. Instead, it is common to take what is considered an adequate number of observations on the population, choose the risk we shall take of rejecting a true hypothesis, and then reject the hypothesis being tested if t goes beyond that predetermined limit. To illustrate, suppose that a sample of 15 observations is to be taken under conditions which specify the population being sampled, and that it is decided that it is appropriate to take 1 chance in 20 of rejecting a true hypothesis. For 14 degrees of freedom, a t which is at, or above, 2.15 numerically (see Table IV) will occur about 1 time in 20 when the choice of μ is entirely correct. If we decide to regard all t's which are outside the interval $-2.15 \leqq t \leqq +2.15$ as being the result of a false hypothesis regarding μ, we run a risk of 1 in 20 of rejecting a true hypothesis as a result of sampling variations.

Problem 6.41. Suppose that some educators test two proposed teaching procedures in the following way:

(1) All available records and the opinions of teachers are applied to the selection of 20 students who, as a group, do a good job of representing students who will be studying the materials upon which the test is to be based.

(2) Two equally difficult sections of subject matter are carefully chosen.

(3) The group of 20 students is taught one section by method A, the other by method B.

(4) Two equally difficult examinations, one on each section of the subject matter, are formulated by competent teachers and given to the 20 students.

(5) The average difference, student-by-student, between the two test scores is to be used as the measure of the difference in efficiency between the two methods of instruction.

It will be assumed herein that the following test scores were made under the two methods:

Student Number	Grade on Method A	Grade on Method B	X $(A - B)$
1	90	85	5
2	72	73	−1
3	86	80	6
4	78	75	3
5	97	95	2
6	85	81	4
7	64	50	14
8	69	65	4
9	76	70	6
10	79	70	9
11	81	78	3
12	83	83	0
13	75	71	4
14	85	80	5
15	72	69	3
16	100	90	10
17	88	82	6
18	77	65	12
19	80	70	10
20	73	65	8

What conclusions can we draw validly from these results?

In the usual manner it is found that $\bar{x} = 5.65$ in favor of A, and that $s_{\bar{x}} = 0.87$. Therefore, $t = (5.65 - \mu)/0.87$, with 19 degrees of freedom. What is a reasonable hypothesis regarding the magnitude of μ in the population of $X_i = A_i - B_i$ assumed to follow a normal distribution? The purpose of this study was to determine if one method of instruction is better than the other, and, perhaps, assess the magnitude of the difference if one exists. If one method is superior to the other, μ is not equal to zero; however, there appears to be no logical way to decide ahead of the test just what the size of μ might be. The problem therefore is attack by assuming that $\mu = 0$ and then determining statistically just how satisfactory such a hypothesis is.

If $\mu = 0$, $t = (5.65 - 0)/0.87 = 6.49$, with 19 degrees of freedom. It is clear from Table IV that a t of this size is an extremely rare occurrence, a fact which leads us to reject decisively the hypothesis that $\mu = 0$. In other words, method A most certainly is *some* better than method B. If there is any benefit to be derived from an estimate

of the magnitude of μ, now that the hypothesis $H_0(\mu = 0)$ is rejected, confidence intervals can be obtained at any appropriate level of confidence.

PROBLEMS

1. Suppose that 20 pairs of college students have been so selected that the members of each pair can be considered equal in intelligence, scholastic records, and in other factors associated with the making of good scholastic records in a college. Suppose one member of each pair is enrolled in a class in social science which is to be taught by a discussion method emphasizing analysis of problems and reflective thinking; whereas the other member of each pair is put in a class taught by a more formal lecture-recitation procedure. The subject to be studied is the same in each class, and the teachers are considered to be equal in teaching abilities. At the end of the teaching period all 40 students are given the same examination with the following results:

	Grade			Grade	
Pair	Discussion	Lecture	Pair	Discussion	Lecture
1	120	110	11	115	108
2	79	75	12	103	91
3	65	70	13	75	70
4	67	75	14	92	95
5	80	75	15	105	102
6	85	80	16	82	78
7	98	90	17	78	76
8	110	95	18	87	90
9	108	92	19	131	120
10	86	80	20	50	51

The discussion method appears to be the better for producing good test scores, but there is considerable variation. Test the hypothesis that the two methods actually are equal on the average (that is, $\mu = 0$) and draw appropriate conclusions.

2. A study was made to determine if tomatoes high up on a plant have more ascorbic acid (vitamin C) than those lower down on the same staked plant. To study this matter, 10 pairs of red-ripe tomatoes were taken from 10 plants, with one member of each pair being from the fifth cluster and the other from the seventh cluster from the bottom of each plant. Each of the tomatoes from the seventh cluster had more vitamin C than the corresponding tomato from the fifth cluster by the following respective amounts:

X: 6.6, 11.6, 10.9, 7.4, 8.8, 10.3, 7.4, 7.8, 5.8, and 4.0 milligrams/100 grams.

Given that $\Sigma X = 80.6$ and $\Sigma X^2 = 700.46$, compute a 90 per cent confidence interval on the true average amount by which the ascorbic acid in tomatoes at the seventh cluster exceeds that at the fifth cluster on the same plant, and draw appropriate conclusions regarding μ.

Ans. CI_{90}: $6.7 \leq \mu \leq 9.4$ milligrams/100 grams.

3. Solve as in problem 2, using the following data from the sixth and eighth clusters of 10 plants:

X: 7.0, 13.3, 8.6, 6.4, 8.3, 9.9, 2.6, 9.1, 6.6, and 1.6.

You are given that $\Sigma X = 73.4$ and $\Sigma X^2 = 643.40$.

4. During the winter tomatoes often are shipped green and allowed to ripen in the package. Aside from matters of flavor and appearance, it is of interest to know what effect this practice has on the vitamin C concentration in the fruit. Two tomatoes were picked from each of 18 plants and at the same cluster on the plant. One was red-ripe, the other was green (no red or yellow coloring). The red-ripe member of each pair was analyzed immediately for vitamin C; the other was ripened at room temperatume out of the sun until red-ripe before its vitamin content was determined. Then the differences in vitamin C between members of pairs was determined with the following results:

$\Sigma X = 49.37$ milligrams/100 grams, favoring vine-ripened tomatoes, and

$\Sigma X^2 = 387.5911$.

Determine statistically if there probably is a loss in ascorbic acid which is due to picking tomatoes green and letting them ripen on the way to market or on the shelf.

5. Suppose that a sociologist has conjectured that the average rent for two-room furnished apartments in a certain section of a city is \$90 per month. A sample of 20 apartments had $\bar{x} = \$82.50$, with $s = \$8$. Use the t-test to determine if the hypothesis H_0 ($\mu = 90$) is acceptable when sampling variance is taken into account.

6. Suppose that a timber cruiser has judged that the average breast-high diameter of a certain stand of timber is 2 feet. Is the timber cruiser's estimate reasonable if 31 trees are selected at random with these results: $\bar{x} = 2.3$ feet and $s = 0.8$ feet?

7. Suppose that a store conducts a study of the comparative net profits from roasting ears sold in cellophane packs as compared to the loose ears in the husks. The experiment is conducted for the 26 business days of a month. At the end of each day, the net profit per ear is figured for each way of selling the corn. The average advantage of using the cellophane pack on these 26 daily comparisons was 2 cents, with a standard deviation $s = 0.5$ cent. When sampling variation is considered, was the average advantage of the cellophane pack enough to justify the conclusion that it is really more profitable?

8. If, in a certain investigation, $\bar{x} = 10.5$ and $s = 3$, how large must n be to cause the rejection of the hypothesis: H_0 ($\mu = 0$) at the one per cent level?

9. If in a sampling study to which the t-test is appropriate the n is 28 and $s = 5$, how large must $\bar{x}$ be before the hypothesis H_0 ($\mu = 0$) will be rejected at the 5 per cent level?

10. Suppose that 27 pairs of plants of a certain species have been selected for close similarity and are planted in pairs as close together as is appropriate for this species. One member of each pair has had some boron added to the fertilizer; otherwise the plants are treated identically. If the 30 plants having boron outgrew their partners by an average of 3.6 centimeters, with the standard deviation of the difference being $s = 1.2$ centimeters, is this sufficient evidence for the statement that the addition of the boron produces some additional growth?

6.5 A STATISTICAL TEST OF THE HYPOTHESIS THAT TWO SAMPLES OF OBSERVATIONS HAVE BEEN DRAWN FROM THE SAME NORMAL POPULATION OF NUMERICAL MEASUREMENTS

If two samples have been taken under the same conditions but with some one important feature changed, we usually wish to learn if this change has produced a new population of measurements. For example, if two groups of Duroc-Jersey pigs have been fed two different rations, the experimenter wants to know if the difference in ration has produced an important difference in average daily gains. That is, has the difference in ration created different populations of average daily gains? Fundamentally, the method to be employed in the solution of this problem is the same as that described in the preceding section, but the mechanics of the procedure need to be altered to fit the new sampling situation.

The following symbols will be employed:

$\bar{d}_i = \bar{x}_{1i} - \bar{x}_{2i}$ = the difference between the ith sample mean from samples from group 1 and the corresponding sample from group 2, and

$s_{\bar{d}}$ = the standard deviation of the $\bar{d}_i$.

Before the general method for attacking the problem just posed is described, some actual sampling experiences will be presented in tabular form, and discussed. Table 6.51 shows a summary of 403 $\bar{d}_i$ obtained from pairs of samples, each with $n = 10$ drawn from the near-normal population of Table 6.21. It is recalled that the standard deviation of that population is $\sigma = 10$.

TABLE 6.51

FREQUENCY AND *r.c.f.* DISTRIBUTIONS FOR 403 SAMPLE VALUES OF $\bar{d}_i$ WITH $n = 10$ DRAWN FROM A NEAR-NORMAL POPULATION WITH $\mu = 60$ AND $\sigma = 10$

Class Interval	f	*r.c.f.*	Class Interval	f	*r.c.f.*
16.5–19.4	1	1.000	− 4.5 to − 1.6	88	.342
13.5–16.4	2	.998	− 7.5 to − 4.6	33	.124
10.5–13.4	3	.993	−10.5 to − 7.6	15	.042
7.5–10.4	14	.985	−13.5 to −10.6	1	.005
4.5– 7.4	36	.950	−16.5 to −13.6	1	.002
1.5– 4.4	80	.861		—	
−1.5– 1.4	129	.663	Total	403	

Arithmetic mean of $\bar{d}_i = +0.06$; standard deviation = 4.43.

The frequency distribution in Table 6.51 displays one notable contrast to that of the $\bar{x}$'s of Table 6.21, namely, the $\bar{d}_i$ are more variable. As a matter of fact, the standard deviation of the $\bar{d}_i$ is greater than that of the $\bar{x}_i$ by a factor of about 1.4 in this instance in which $n = 10$. It can be shown mathematically that the factor theoretically is $\sqrt{2}$, which = 1.414, approximately; hence the empirical results of Table 6.51 agree quite well with the theory.

The following theorem summarizes some of the above information and makes it more precise:

Theorem. If a very large number of pairs of independently drawn samples of n observations is taken from a normal population with standard deviation = σ, then:

(*a*) The population of differences $\bar{d}_i = \bar{x}_{1i} - \bar{x}_{2i}$ will conform to the normal distribution.

(*b*) The arithmetic mean of the population of $\bar{d}_i$ is 0.

(*c*) The standard deviation of the population of $\bar{d}_i$ is

$$\sigma_{\bar{d}} = \sigma\sqrt{2/n}\,.$$

For the situation summarized in Table 6.51, $\sigma_{\bar{d}} = 10\sqrt{2/10} = 4.47$, an amount which agrees quite well with the 4.43 shown in that table as the observed standard deviation for 403 $\bar{d}$'s.

In practice, the standard deviation, σ, nearly always is unknown so that an estimate must be made from the sample. When a pair of samples has been taken it has been determined by mathematical analysis that the best procedure to follow is this: Lump together, or pool, the sums of the squares of the x_i in each sample taken separately and divide that sum by $2(n - 1)$ before taking the square root. In symbols, the following is the recommended estimate of σ:

$$s = \sqrt{\frac{\Sigma(x_1^2) + \Sigma(x_2^2)}{2(n-1)}}, \tag{6.51}$$

where $\Sigma(x_1^2)$ = the sum of squares of the deviations of the X's of the first sample from their mean; and likewise for $\Sigma(x_2^2)$.

When the theorem above is applied, we obtain the following formula for the sampling estimate of $\sigma_{\bar{d}}$:

$$s_{\bar{d}} = s\sqrt{2/n} = \sqrt{\frac{\Sigma(x_1^2) + \Sigma(x_2^2)}{n(n-1)}}. \tag{6.52}$$

It turns out from mathematical analysis that the sampling ratios $t_i = (\bar{d}_i - \mu)/s_{\bar{d}_i}$ follow the same sampling distribution as the t previously discussed, if μ = the true average $\bar{d}_i$; hence Table IV can be used here provided we employ $2(n - 1)$ degrees of freedom for t instead of the $(n - 1)$ used before.

The way is now open to solve the type of problem proposed at the beginning of this section. To illustrate, suppose that 20 steers of the same breed, weight, and previous history are divided into two equal lots by some impartial means such as drawing numbers from a hat. Thereafter, one group is fed a ration 50 per cent of which is peanut meal and 50 per cent is a standard ration. The other group of steers is fed only 20 per cent peanut meal, the remainder being the same standard ration. After an adequate period of time, the average daily gains of the steers were obtained as follows, with A standing for the group of steers whose diet contained 50 per cent of peanut meal:

Group	
A	B
1.55 lb	1.66 lb
1.68	1.82
1.42	1.71
1.45	1.78
1.52	1.69
1.58	1.73
1.56	1.75
1.61	1.61
1.54	1.90
1.48	1.72
15.39	17.37

$\bar{x} = 1.54$ $\bar{x} = 1.74$

$$\Sigma(x_1^2) = 0.0531, \quad \Sigma(x_2^2) = 0.0608.$$

$$s_{\bar{d}} = \sqrt{\frac{0.0531 + 0.0608}{10(10 - 1)}}$$

$$= \sqrt{0.001266}$$

$$= 0.036, \text{ approximately.}$$

$$t = (0.20 - \mu)/0.036 = 5.56 \text{ if } \mu = 0.$$

t has 18 degrees of freedom.

We learn from Table IV that less than one-half of one per cent of the sample t's with 18 degrees of freedom are numerically as large as 5.56; therefore, the hypothesis that $\mu = 0$ is rejected and the two samples are regarded as having been drawn from different normal

populations of average daily gains. It is concluded that the steers on a diet containing 50 per cent peanut meal will, on the average, produce lower gains than those on only 20 per cent peanut meal.

Ordinarily the experimenter would wish to carry the statistical analysis farther than this by means of confidence intervals. If the steers on 20 per cent peanut meal do not gain enough more to pay for the added expense of using more of the standard ration which costs more than the peanut meal, it still may not pay to use the diet B. If 95 per cent confidence limits are chosen here, they are determined by the usual methods from

$$-2.10 \leq (0.20 - \mu)/0.036 \leq +2.10; \text{ or}$$

$$0.12 \leq \mu \leq 0.28.$$

Therefore, it can be concluded with considerable confidence (associated with odds of 19 to 1) that the average advantage due to feeding 20 per cent peanut meal instead of 50 per cent is at least 0.12 pound of gain per day but not over 0.28 pound. Given the current price of steers of the sort under study, we can decide which ration is economically preferable. Obviously, other factors would be considered in practice, but they are separate considerations.

Although it seems preferable in studies such as those illustrated in this section to have equally many observations in each group, this is not always an attainable goal. If the sample sizes are unequal, say n_1 and n_2 instead of n each, the above methods are applicable but the formulas are changed to fit these new circumstances. Formula 6.51 is replaced by

$$s = \sqrt{\frac{\Sigma(x_1^2) + \Sigma(x_2^2)}{n_1 + n_2 - 2}}; \text{ and} \tag{6.53}$$

formula 6.52 is replaced by

$$s_{\bar{d}} = s\sqrt{1/n_1 + 1/n_2} = \sqrt{\frac{\Sigma(x_1^2) + \Sigma(x_2^2)}{n_1 + n_2 - 2}\,(1/n_1 + 1/n_2)}. \tag{6.54}$$

Formulas 6.51 and 6.53 are fundamentally the same in all important respects; each is an estimate based on the deviations $(X_{1i} - \bar{x}_1)$ and $(X_{2i} - \bar{x}_2)$ in both samples. Likewise, formulas 6.52 and 6.54 are fundamentally alike; each comes from the theorem of mathematical statistics that the variance of the difference between the means of pairs of random samples is the sum of the variances of the two means con-

sidered separately. The reader should verify the fact that if in formulas 6.53 and 6.54 $n_1 = n_2 = n$, these formulas become 6.51 and 6.52, respectively.

Many other applications of the t-distribution, and accompanying statistical techniques, could be cited; but the fundamental principles are essentially the same as those already explained.

PROBLEMS

1. Suppose that 5 experimental concrete cylinders of each of two types of concrete have been tested for breaking strength, with the following results in hundreds of pounds per square inch:

Type 1: 40, 50, 48, 46, and 41; and
Type 2: 65, 57, 60, 70, and 55.

Use the t-distribution to determine if the difference in average breaking strength between the two types of concrete can be assigned reasonably to mere sampling variation.

2. Suppose that two groups of 10 steers have been fed two different rations (one to each group) and that the steers are of the same age, breed, and initial weight. Given the following computations determine the 99 per cent confidence interval on the true difference between the means of the average daily gains under the two rations:

Ration A	Ration B
$n_1 = 10$	$n_2 = 10$
$\bar{x}_1 = 1.90$ lb/day	$\bar{x}_2 = 1.55$ lb/day

$$s = 0.20,\ 18\ D/F;\quad t = 3.92$$

Ans. CI_{99}: $0.1 \leq |\mu_1 - \mu_2| \leq 0.6$ lb/day.

3. Suppose that an experiment has been set up at an engineering laboratory to determine the difference in average breaking load between oak and fir beams of the dimensions: 2 inches x 2 inches x 28 feet. The data from tests on 10 beams of each wood are as follows, in pounds:

Oak:	725,	1015,	1750,	1210,	1435,	1175,	1320,	1385,
Fir:	1205,	810,	1110,	530,	765,	1075,	1475,	950,

Oak: 1505, and 1340. Sum = 12,860: $\Sigma X^2 = 17{,}243{,}550$.
Fir: 1020, and 1070. Sum = 10,010: $\Sigma X^2 = 10{,}625{,}400$.

If you can afford a risk of an error of only 1 in 100 what confidence limits do you set on the true difference in average breaking load for these two materials?

4. Draw 5 pairs of samples, each with $n = 10$, from the laboratory population furnished you by the instructor, and compute $t = \bar{d}/s_{\bar{d}}$ for each pair of samples. Then obtain from Table IV the probability that a numerical value of t that size or larger would be obtained while pairs of samples are drawn from the same normal population.

5. An experiment designed to find out if supplemental lighting with incandescent lights will increase the vitamin C content of greenhouse tomatoes produced the following results in milligrams per 100 grams for tomatoes on the bottom two clusters of the plants:

No extra light: 25.92, 28.08, 21.27, 22.53, 26.27, 22.57, 22.57, 30.19, and 20.35.

$\Sigma X = 219.75$, $\Sigma X^2 = 5454.8279$.

Incandescent: 20.30, 29.21, 20.50, 21.50, 23.71, 29.34, 26.32, 15.55, and 29.56.

$\Sigma X = 215.98$, $\Sigma X^2 = 5378.5612$.

Use the t-test to decide whether or not the incandescent lights changed the average ascorbic acid concentration in the greenhouse tomatoes.

6. Given the following two sets of simulated data, assume first that the observations are paired (vertically) and compute and interpret the t. Then assume that the observations are not paired and again do a t-test. Compare these results and the hypotheses tested.

A: 85 72 28 59 75 46 39 68 53; $\Sigma X = 525$. $\Sigma X^2 = 33{,}369$.

B: 80 65 24 58 65 40 38 60 42; $\Sigma X = 472$. $\Sigma X^2 = 27{,}198$.

Ans. (*a*) $t = 4.97$, 8 D/F, reject $H_0(\mu = 0)$.
(*b*) $t = 0.78$, 18 D/F, accept $H_0(\mu_A = \mu_B)$.

7. The antibiotic, aureomycin, has been found to be a growth stimulant for certain animals. The discovery is illustrated by the following two sets of data obtained at Kansas State College under the direction of Dr. E. E. Bartley of the Department of Dairy Husbandry. The measurement of growth used is the total gain during a 12-week period, expressed as a percentage of birth weight.

No Aureomycin	Had Aureomycin
77.6	125.6
81.3	135.5
109.2	122.9
124.1	144.8
101.4	103.3
106.0	142.9
81.7	
70.6	$\Sigma X_2 =$ 775.0
54.8	$\bar{x}_2 =$ 129.17
43.3	$\Sigma(X_2^2) =$ 101,298.36
119.2	
100.0	
$\Sigma X_1 = 1069.2$	
$\bar{x}_1 = 89.10$	
$\Sigma(x_1^2) = 6954.96$	

Obtain the 95 per cent confidence interval on the true difference between the two means μ_1 and μ_2 and tell what information this interval makes available.

8. Suppose that 31 rainbow trout and 31 brook trout are taken at random

from a mountain stream and are measured for length. The rainbows averaged 9.2 inches, with $s = 2$ inches; the brook trout averaged 8.7 inches, with standard deviation = 2.1 inches. Test the hypothesis $H_0(\mu_1 = \mu_2)$ and draw appropriate conclusions.

9. If from a certain study, $\bar{x}_1 = 32.7$ and $\bar{x}_2 = 35.9$, and the pooled estimate of σ is $s = 7.5$. Both samples contained 12 observations. Test $H_0 \mid \mu_1 - \mu_2 \mid = 1$.

10. Suppose that 15 samples of each of two varieties of tomatoes have been analyzed for vitamin C, with these results:

Variety 1	Variety 2
$\bar{x}_1 = 28.5$	$\bar{x}_2 = 30.4$
$\Sigma(x_1^2) = 50$	$\Sigma(x_2^2) = 60$

Test the hypothesis that the true average ascorbic acid concentration in these two varieties is the same.

6.6 USE OF THE SAMPLE RANGE INSTEAD OF THE STANDARD DEVIATION IN CERTAIN TESTS OF STATISTICAL HYPOTHESES

The most difficult computational part of the t-test is the determination of either $s_{\bar{x}}$ or $s_{\bar{d}}$, as the case may be. Another method of testing hypotheses can be used in some situations without the need to compute these standard deviations at all. It uses the sample range as its measure of variation. The loss of precision is not serious for small samples, becomes greater as the size of the sample is increased, and renders the method useless for large samples. The trouble is that the sampling variability of the range is almost as low as that of the standard deviation for small samples but increases quite rapidly with n.

The ratio

$$G = (\bar{x} - \mu)/R, \tag{6.61}$$

where R = sample range can be used in a manner analogous to the t-test procedure. When G has been calculated, Table IX gives the probability that such a sampling $\mid G \mid$ will occur by chance for samples of size n if the hypothesis regarding μ is exactly right. Thereafter the reasoning is just as it was in section 6.4.

When two random samples, each of size n, have been drawn from what is assumed to be the same normal population, the ratio

$$G = \frac{x_1 - x_2}{\text{mean range}}, \tag{6.62}$$

where mean range = arithmetic mean of the ranges of the two samples can be used on problems like those in section 6.5. Table X now is used

instead of Table IX. Again the tables give $P(|G| > G_0)$, where G_0 is the observed numerical size of G.

To illustrate the application of formulas 6.61 and 6.62 reference is made to the problems solved in sections 6.4 and 6.5. First consider problem 6.41. The sample mean is $\bar{x} = 5.65$ and the range is 15; therefore, for $H_0(\mu = 0)$: $G = (5.65 - 0)/15 = 0.377$, with the sample size $= n = 20$. By Table IX, the probability that G would be so large if μ actually were zero is much less than .001; hence the hypothesis that $\mu = 0$ is rejected decisively, as it was from the t-test.

The next example is from section 6.5 and involves two diets fed to steers. In fact, $\bar{x}_1 = 1.54$ and $\bar{x}_2 = 1.74$, $R_1 = 0.26$, $R_2 = 0.29$, and hence the average sample range $= 0.275$. Then

$$G = 0.20/0.275 = 0.727, \text{ with each } n = 10.$$

By Table X a G larger than 0.727 would occur by chance less than 0.1 per cent of the time if both samples were from the same normal population. The hypothesis is rejected; that is, the second diet, which produced the higher average gain in the sampling is considered to produce higher average gains than the first diet.

Given the tables and formulas above, we can derive confidence intervals on μ as before when n is small. This interval would not be expected to be identical with one obtained from the t-distribution for the same confidence coefficient; but it has been shown that, on the average, the two intervals are very close to the same length as long as n is small. (Specifically, K. S. C. Pillai has shown in the September, 1951, number of the *Annals of Mathematical Statistics* that the ratio of the average lengths of the CI_{95}'s by the two methods still is 0.97 when $n = 20$.) To illustrate, consider again the problem of section 6.4 just used above to illustrate the G-test when there is one set of n observations. In this problem the two confidence intervals are obtained as follows:

$$-2.1 \leq \frac{5.65 - \mu}{0.87} \leq +2.1$$

and the 95 per cent confidence interval is

$$3.82 \leq \mu \leq 7.48.$$

Using the ratio G, we have

$$-0.126 \leq \frac{5.65 - \mu}{15} \leq +0.126$$

and hence the 95 per cent confidence interval is

$$3.76 \leq \mu \leq 7.54,$$

which is very much like that derived from the t-distribution.

PROBLEMS

1. Solve problem 1, section 6.4, with the G-test instead of the t-test.

2. Solve problem 3, section 6.5, by means of the G-test.

Ans. $G = 0.289$; $P \cong .064$; accept H_0 tentatively.

3. Draw 25 samples, each with $n = 10$ from a near-normal population, and compute the G for each sample. How many of these G's fall beyond 0.186 in numerical size? How do your results check with Table IX?

4. Suppose that a college is attempting to learn if instruction of a certain type improved in one year a student's ability to think analytically. Also assume that tests exist which reliably measure such ability, and that these tests are given at the beginning and at the end of the school year. If the following *differences* between the last and the first score of each student were obtained, would the G-test cause you to accept or to reject the hypothesis that the teaching procedures employed *failed* to improve analytical thinking?

X: 5, 0, 10, −4, −6, 8, 1, 7, −10, 0, 3, 5, −1, 8, 4, 0, −3, 7, 7, and 9.

Ans. $G = 0.125$; $P = .05$; reject $H_0(\mu = 0)$.

5. Make up, and solve, a problem like problem 4, which has the same $\bar{x}$ but for which G is twice as large. Half as large.

6. Suppose that information is sought analogous to that in problem 4, but there are two separate classes of 15 students being taught by each method. The two classes are supposed to be equal at the start of the teaching period. Given the following gains (+) or losses (−) in score during the year, draw appropriate conclusions by means of the G-test and Table X:

Method I: 10, 3, −2, 5, 0, −8, 14, 1, −12, 5, 5, 9, 7, −1, and 9.
Method II: −2, 5, 5, 4, 0, 7, 6, −1, 4, 10, 8, 11, 10, 0, and 13.

Ans. $G = 0.114$; $P > .10$; accept $H_0(\mu_1 = \mu_2)$.

7. An experiment intended to discover if blue fluorescent lights will increase the vitamin C concentration in tomatoes on the seventh and eighth clusters from the bottom of the plant gave these results, in milligrams per 100 grams:

No extra light: 38.57, 39.39, 33.44, 34.32, and 38.01.
Blue fluorescent: 33.72, 37.85, 39.07, 31.16, and 35.69.

Test the hypothesis that the blue light does not change the vitamin C concentration, and draw valid conclusions.

8. Suppose that two methods of computing basal metabolism for the same 11 subjects produced the following pairs of records, in calories per square meter per hour.

I: 31.42, 30.90, 34.92, 30.59, 30.53, 33.08, 32.61, 30.46,
II: 30.73, 31.44, 32.82, 31.80, 29.16, 32.96, 32.32, 30.76,
I: 30.55, 33.19, and 29.22. $\Sigma X_1 = 347.47$.
II: 27.65, 32.54, and 29.30. $\Sigma X_2 = 341.48$.

Use the G-test to decide if one method tends to produce higher metabolism records than the other, and explain your decision in terms of sampling phenomena. *Ans.* $G = 0.100$; $P > .10$; accept $H_0(\mu_1 = \mu_2)$.

9. Some varieties of wheat produce flour which typically takes longer to mix into proper doughs than others. Decide by the G-test if Kharkof actually has (as appears from the samples) a longer mixing time than Blackhull:

Kharkof: 3.00, 1.88, 1.62, 1.50, 1.75, 1.38, 1.12, 1.88, 2.50, 1.62, 2.88, 2.50, 3.88, and 2.75. Mean = 2.16.

Blackhull: 1.25, 2.38, 1.62, 1.50, 1.25, 1.38, 2.25, 2.12, 1.84, 2.38, 2.25, 1.50, 2.00, and 1.62. Mean = 1.84.

10. Compute the 90 per cent confidence intervals for the two varieties of problem 9 and compare them. Draw appropriate conclusions.

Ans. CI_{90}: $-0.04 \leq |\mu_1 - \mu_2| \leq +0.69$.

6.7 THE CENTRAL LIMIT THEOREM AND NON-NORMAL POPULATIONS

The statistical methods which have been discussed in this chapter are based on the assumption that the populations involved are normal. In practice this requirement rarely is met rigorously; hence we may wonder if the subject matter of this chapter is chiefly of academic interest because it does not fit actual conditions. This is not the situation because of the truth of the *central limit theorem.*

This theorem states essentially that if any population of numerical measurements has a finite mean and variance, μ and σ^2, respectively, the frequency distribution of the sampling mean, $\bar{x}$, will be essentially a normal distribution with mean = μ and variance = σ^2/n *if the* n *is very large.* As a matter of fact, the necessary size of n depends on the degree of non-normality of the original population. Tables 6.71*A*, *B*, *C*, and *D* summarize a decidedly non-normal population of counts of flies on dairy cattle, and show some observed distributions of $\bar{x}$'s for samples with $n = 9$, 16, and 25. Figure 6.71 displays these same distributions visually. It is rather obvious that none of these sample sizes is very large, and therefore the distributions of $\bar{x}$ are still noticeably non-normal. However, the meaning of the central limit theorem is illustrated.

It can be seen from Tables 6.71 and from Figures 6.71:

(*a*) That the parent population is extremely non-normal.

(*b*) That even with only nine observations per sample, the distribution of $\bar{x}$ has gone a long step towards fulfilling the ideal expressed by the Central Limit Theorem.

(*c*) As n was increased, the distribution of $\bar{x}$, and its mean and variance, approached more and more closely to those features which the Central Limit Theorem assures us will be attained if n is sufficiently large.

The foregoing discussions are not intended to make us ignore the non-normality of distributions met in practice, but they do indicate that a great many moderately non-normal distributions can be studied by means of the techniques explained in this chapter.

In this chapter the ratio $(\bar{x} - \mu)/s_{\bar{x}}$ was said to follow the t-distribution with the same number of degrees of freedom that $s_{\bar{x}}$ has as a sampling estimate of $\sigma_{\bar{x}}$. Actually any ratio $(w - \mu)/s$ will follow

TABLE 6.71*A*

Summary of Counts of Flies on Dairy Cattle Tethered in a Field at Kansas State College After They Were Sprayed with an Effective Fly Repellent

Class Interval	f	*r.c.f.*	Normal *r.c.f.*, Same μ and σ	Difference
168–175	1	1.000	1.000	0
160–167	0	1.000	1.000	0
152–159	1	1.000	1.000	0
144–151	0	.999	1.000	−.001
136–143	0	.999	1.000	−.001
128–135	0	.999	1.000	−.001
120–127	1	.999	1.000	−.001
112–119	3	.999	1.000	−.001
104–111	6	.998	1.000	−.002
96–103	2	.995	1.000	−.005
88– 95	7	.994	1.000	−.006
80– 87	8	.992	1.000	−.008
72– 79	10	.988	1.000	−.012
64– 71	11	.984	.999	−.015
56– 63	31	.980	.999	−.019
48– 55	26	.968	.994	−.026
40– 47	59	.957	.976	−.019
32– 39	102	.934	.932	+.002
24– 31	206	.893	.837	+.056
16– 23	392	.812	.687	+.125
8– 15	771	.656	.500	+.156
0– 7	869	.348	.309	+.039
	$\Sigma(f) = 2506$			

$\mu = 15.37, \quad \sigma^2 = 257.28$

TABLE 6.71*B*

DISTRIBUTION OF MEANS OF RANDOM SAMPLES WITH $n = 9$ DRAWN FROM THE POPULATION OF TABLE 6.71*A*

$\bar{x}$ Interval	f	*r.c.f.*	Normal *r.c.f.*, Same μ and σ	Difference
36–38.99...	3	1.000	1.000	0
33–35.99...	8	.997	1.000	−.003
30–32.99...	4	.990	1.000	−.010
27–29.99...	11	.986	.998	−.012
24–26.99...	30	.975	.988	−.013
21–23.99...	73	.947	.954	−.007
18–20.99...	139	.877	.866	+.011
15–17.99...	221	.744	.701	+.043
12–14.99...	268	.533	.480	+.053
9–11.99...	211	.278	.266	+.012
6– 8.99...	74	.076	.113	−.037
3– 5.99...	6	.006	.037	−.031
	$\Sigma(f) = 1048$			

$\mu_{\bar{x}} = 15.27, \quad \sigma_{\bar{x}}^2 = 26.72$

By Central Limit Theorem (if n is large enough) $\mu_{\bar{x}} = 15.37$
$\sigma_{\bar{x}}^2 = 28.59$

TABLE 6.71*C*

DISTRIBUTION OF MEANS OF RANDOM SAMPLES WITH $n = 16$ DRAWN FROM THE POPULATION OF TABLE 6.71*A*

$\bar{x}$ Interval	f	*r.c.f.*	Normal *r.c.f.*, Same μ and σ	Difference
32–33.99...	2	1.000	1.000	0
30–31.99...	1	.998	1.000	−.002
28–29.99...	6	.997	1.000	−.003
26–27.99...	13	.991	.999	−.008
24–25.99...	13	.979	.994	−.015
22–23.99...	40	.966	.978	−.012
20–21.99...	74	.928	.938	−.010
18–19.99...	106	.856	.855	+.001
16–17.99...	162	.754	.719	+.035
14–15.99...	209	.597	.540	+.057
12–13.99...	206	.395	.352	+.043
10–11.99...	145	.196	.195	+.001
8– 9.99...	51	.056	.091	−.035
6– 7.99...	7	.007	.035	−.028
	$\Sigma(f) = 1035$			

$\mu_{\bar{x}} = 15.56, \quad \sigma_{\bar{x}}^2 = 17.32$

By Central Limit Theorem (if n is large enough) $\mu_{\bar{x}} = 15.37$
$\sigma_{\bar{x}}^2 = 16.08$

TABLE 6.71*D*

DISTRIBUTION OF MEANS OF RANDOM SAMPLES WITH $n = 25$ DRAWN FROM THE POPULATION OF TABLE 6.71*A*

$\bar{x}$ Interval	f	*r.c.f.*	Normal *r.c.f.*, Same μ and σ	Difference
26.00–27.49	3	1.000	1.000	0
24.50–25.99	3	.997	1.000	−.003
23.00–24.49	16	.994	.997	−.003
21.50–22.99	22	.978	.990	−.012
20.00–21.49	43	.956	.968	−.012
18.50–19.99	85	.914	.917	−.003
17.00–18.49	121	.830	.821	+.009
15.50–16.99	180	.710	.673	+.037
14.00–15.49	188	.532	.492	+.040
12.50–13.99	168	.346	.313	+.033
11.00–12.49	133	.179	.169	+.010
9.50–10.99	41	.048	.077	−.029
8.00– 9.49	5	.007	.029	−.022
6.50– 7.99	2	.002	.009	−.007
	$\Sigma(f) = 1010$			

$$\mu_{\bar{x}} = 15.51, \quad \sigma_{\bar{x}}^2 = 10.07$$

By Central Limit Theorem (if n is large enough) $\mu_{\bar{x}} = 15.37$
$\sigma_{\bar{x}}^2 = 10.29$

the t-distribution as long as w is normally distributed with mean μ, and s is calculated as described earlier. Hence if w is a sample mean drawn from a non-normal population which satisfies the few requirements of the Central Limit Theorem, and if n is large enough, the ratio $(w - \mu)/s$ can be considered quite accurately to follow a t-distribution. Thereafter the methods introduced in this chapter for estimating parameters and for testing hypotheses regarding parameters become applicable.

One word of warning is in order, however, before this subject is left. In any particular sampling situation, the standard deviation, $\sigma_{\bar{x}}$, needs to be estimated from the sample. This is done by means of $s_{\bar{x}}$. What happens to the quality of this estimate when the parent population is radically non-normal? Under such circumstances the beginner is advised to seek the advice of a statistician.

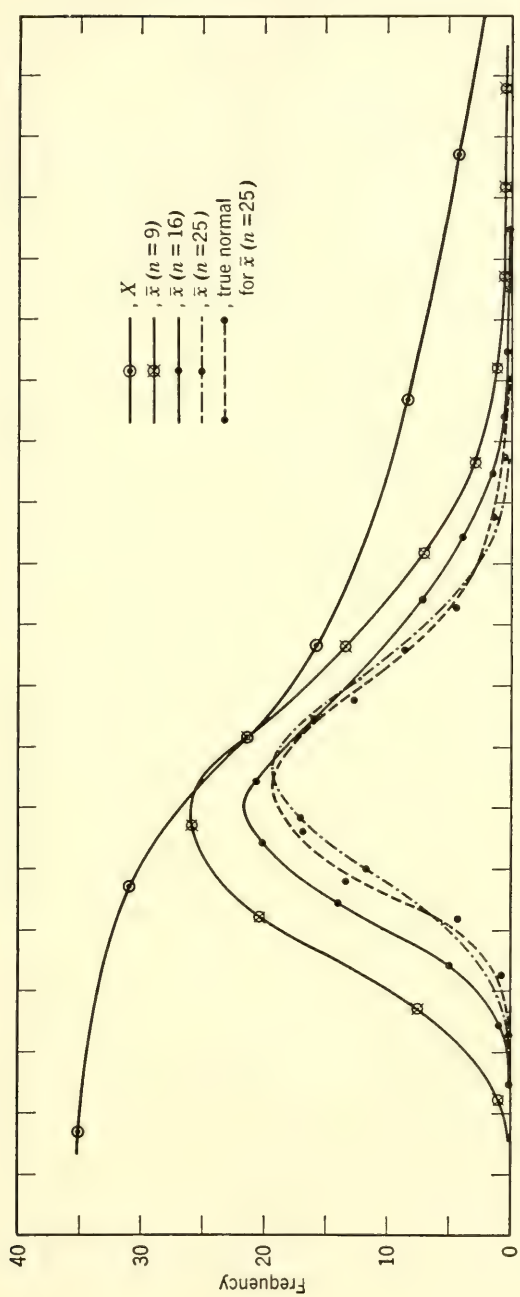


Figure 6.71. Frequency distribution of $\bar{x}$ ($n = 9$, 16, and 25), of the non-normal population of X's from which samples were drawn, and of the normal population of $\bar{x}$ specified by the Central Limit Theorem when $n = 25$. (See Tables 6.71*A*, *B*, *C*, and *D*.)

REVIEW PROBLEMS

1. Who was Student, and how was his work connected with the development of present-day methods of statistical analysis?

2. Calculate the arithmetic mean and the standard deviation for a set of numbers, Y_i, given that $\Sigma Y = 900$ grams, and $\Sigma Y^2 = 55{,}465$ grams^2, where the Y's are the weights of female rats at 28 days of age. There are 20 rats in the sample.

Ans. $\bar{y} = 45$ grams. $s = 28.1$ grams.

3. Compute the 80 per cent confidence interval for problem 2 on the true mean 28-day weight of such rats, and draw appropriate conclusions.

4. What would be the general change in the confidence interval of problem 3 if 95 per cent limits instead of 80 per cent limits had been computed? What would be the effect if the ΣY^2 had been smaller, the remainder of the numbers staying the same?

5. Graph the binomial frequency distribution of the numbers of sums of 6 thrown with two unbiased dice on sets of 8 throws.

6. Compute for problem 5 the probability that on any particular future set of 8 throws at least 3 sums of 6 will be thrown. *Ans.* .087.

7. Take any newspaper which lists prices of bonds and determine the median price and also the range.

8. Calculate the coefficient of variation for problem 2, using $\bar{y}$ in place of μ and s in place of σ, and tell what sort of information it provides about the weights of the rats in the sample. *Ans.* CV = 62.4 per cent.

9. Draw 10 samples of 12 members each from the laboratory population and compute t and G for each sample, using the correct hypotheses regarding μ.

10. Determine the upper limits of the 20th and 85th percentiles for the frequency distribution of Table 6.31 and state what information they give.

Ans. Upper limit of 20th percentile = 0.90 by interpolation, = 0.86 by Figure 6.31. Upper limit of 85th percentile = 1.12 by interpolation, = 1.10 by Figure 6.31.

11. If 100 t_i were to be drawn at random from among those summarized in Table 6.31, what is the expected number of them falling between $t = 0$ and $t = 1.50$?

12. Following are some experimental results from tests of the breaking strengths of the wet warp of rayon and wool fabrics in pounds:

Rayon: 29.5, 31.0, 28.7, 29.1, 28.4, 28.9, 30.9, and 29.0.
Wool: 25.3, 28.9, 19.2, 25.1, 21.1, 31.4, 25.6, and 19.0.

Does the difference in average breaking strength lie beyond the bounds of reasonable sampling variation according to the t-test? Solve problem also by the G-test, and compare the conclusions. $\Sigma X_R = 235.5$, $\Sigma X^2{}_R = 6939.33$. $\Sigma X_W = 195.6$, $\Sigma X^2{}_W = 4921.48$.

Ans. $t = 3.10$, $14\,D/F$, $P = .008$; reject $H_0(\mu_1 = \mu_2)$. $G = 0.665$, $n = 8$, $P = .002$; reject $H_0(\mu_1 = \mu_2)$.

13. Suppose that twelve 2 inch x 12 inch x 8 inch wood blocks were tested for strength with the following results in thousands of pounds: 6.5, 17.0, 10.0, 15.1, 13.5, 16.4, 19.8, 7.7, 11.5, 14.5, 12.7, and 12.9. Place 95 per cent confidence limits on the true average strength of such blocks, and interpret these limits.

14. You are given the following hypothetical data from an experimental study of the average daily gains (in pounds) of two groups of 10 steers each:

$$\text{For group A: } \bar{x}_A = 2.35,\ s^2_A = 12.$$
$$\text{For group B: } \bar{x}_B = 1.75,\ \Sigma x^2_B = 180.$$

Is the difference in mean average daily gain, $\bar{d} = 0.60$ pound, beyond the bounds of reasonable sampling variation; that is, is it statistically significant?

Ans. $t = 0.34$; 18 D/F, $P \cong .63$; accept H_0 ($\mu_1 = \mu_2$).

15. Suppose that you have taken the bid in a bridge game and that you and your partner have all the trumps but J, 10, 7, 4, and 3. Before you have led at all, what do you compute as the probability that you would get all the trumps out within 3 leads?

16. Suppose that you have been told that when 6 unbiased coins were tossed at least 3 of them showed heads. What is the probability that exactly 4 of the coins turned up heads. *Ans.* $P(r = 4 \text{ heads}) = 15/42$.

17. Suppose that a large jug contains the following numbers of each denomination of paper currency, and that you are to withdraw a bill without looking and keep it: 50 one-dollar bills, 25 five-dollar bills, 10 ten-dollar bills, 5 twenty-dollar bills, 2 fifty-dollar bills, and 1 one-hundred-dollar bill. What is your mathematical expectation on such a game?

18. If 2 cards are drawn simultaneously from a bridge deck, what is the probability that one will be a spade, the other a heart? *Ans.* 13/102.

REFERENCES

Dixon, Wilfrid J., and Frank J. Massey, Jr., *Introduction to Statistical Analysis,* McGraw-Hill Book Company, New York, 1951.

Hald, A., *Statistical Theory with Engineering Applications,* John Wiley and Sons, New York, 1952.

Neyman, Jerzy, *First Course in Probability and Statistics,* Henry Holt and Company, New York, 1950.

Snedecor, George W., *Statistical Methods Applied to Experiments in Agriculture and Biology,* Fourth Edition, Iowa State College Press, Ames, Iowa, 1946.

CHAPTER 7

Linear Regression and Correlation

It often is advantageous to consider two types of numerical measurements simultaneously because they are related to each other. For example, the following table records the mean monthly temperatures from January to July at Topeka, Kansas, along with the month of the year:

Month of the Year:	Jan.	Feb.	Mar.	Apr.	May	June	July
Mean Temperatures (degrees Fahrenheit)	38.0	41.7	54.0	66.0	74.4	83.8	88.7

If the month to which each temperature applies were to be ignored, these temperatures simply would be seven numbers which might fall in the following random order (obtained by drawing them at random): 88.7, 54.0, 66.0, 38.0, 83.8, 74.4, and 41.7. In this form the numbers seem to be quite variable about their arithmetic mean, 63.8°F. However, when considered in conjunction with the month as the second variable, these temperatures follow an orderly pattern. This point is illustrated graphically in Figures 7.01*A* and *B*, in which temperatures are first plotted against the random order in which they were drawn, and then against the month to which they apply.

Figure 7.01*A* merely re-emphasizes the remarks made above about the excessive variability about the mean, 63.8, and suggests that such an average would be of doubtful utility because the temperatures are too inconsistent. But it appears from Figure 7.01*B* that the mean temperatures for the first six months of the year increased in quite an orderly manner from month to month, with little deviation from a linear upward trend. Hence, a better analysis of these data can be obtained by taking proper account of the second variable, time.

A straight line is drawn into Figure 7.01*A*, 63.8 units above the horizontal axis, to represent the arithmetic mean of the temperatures whose individual magnitudes are indicated by the ordinates of

the points on the graph. The amounts by which the monthly mean temperatures are above or below the mean of all the temperatures are shown as vertical distances above or below the horizontal line.

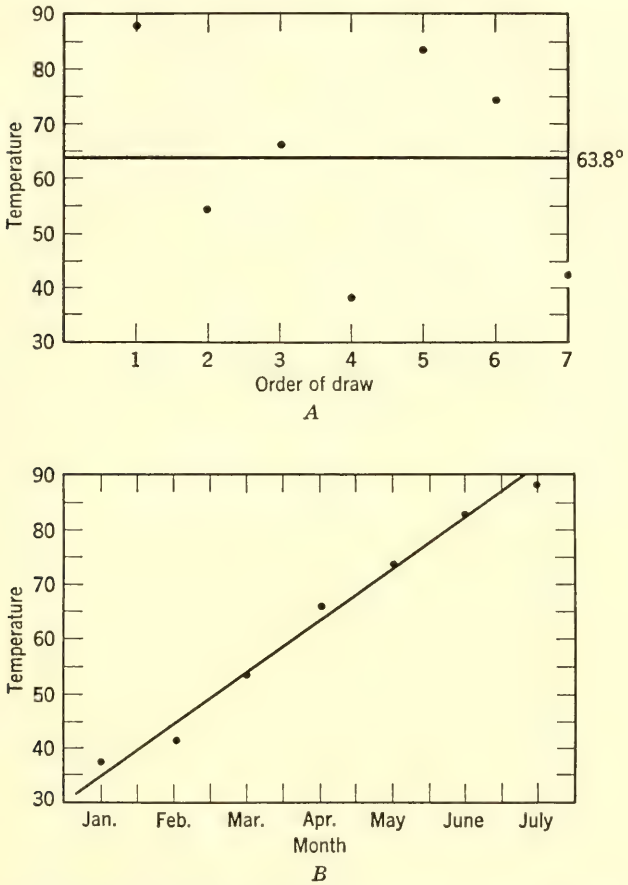


Figure 7.01. Mean monthly temperatures at Topeka, Kansas, plotted first at random and then according to the corresponding month.

As far as Figure 7.01*A* is concerned these deviations are simply the consequences of unexplained variations in temperature. However, when each temperature is associated with the month to which it belongs (as in Figure 7.01*B*) it is apparent that all but a small amount of the variability among these temperatures is associated with a definite tendency to increase rather uniformly with succeeding months of the season.

The trend line drawn into Figure 7.01*B* was determined just "by eye"; but it usually is preferable to have a standard method of determining where the line should be drawn. This matter will be discussed in the following four sections.

7.1 SCATTER DIAGRAMS AND TYPES OF TREND LINES

A number of the statistical methods with which the reader is already familiar can be employed in the analysis of data involving two variables. One additional matter must be studied, however, namely, the relationship between the two variables. A little graphic analysis usually is worth while before the numerical analyses are undertaken.

There are many ways in which one variable, Y, can change with respect to another variable, X, as successive pairs of observations are taken with the X, say, increasing in magnitude. The size of Y may tend to increase as X increases; Y may tend to decrease as X increases; or some of both may occur over the range of values studied. In addition there are numerous ways in which Y can increase as X increases; and similarly for the other possibilities just mentioned. To illustrate, consider the following tables of pairs of values for X and Y:

(*A*)		(*B*)		(*C*)		(*D*)		(*E*)		(*F*)		(*G*)	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
1	7	1	50	1	25	0.0	0.75	−3	8.5	1	2.5	1.0	40
3	12	2	39	2	40	0.5	0.80	−2	5.2	2	8.0	1.5	25
5	14	3	24	3	28	1.0	1.20	−1	0.5	3	32.2	2.0	44
7	20	4	21	4	42	1.5	2.60	0	0.5	4	60.0	2.5	20
9	22	5	9	5	22	2.0	3.80	1	0.7	5	127.2	3.0	35
11	30					2.5	5.75	2	4.8			3.5	28
						3.0	10.40	3	8.9			4.0	47
												4.5	26
												5.0	34

It is helpful to a mathematical study of the relationship between two variables if the pairs of corresponding numerical measurements, X and Y, are represented by points on a graph, as they were in elementary algebra. This has been done in Figures 7.11*A*, *B*, ..., *G* for the data immediately above.

Such graphs are called *scatter diagrams*. It is noted from these figures that the points may not exactly fit any simple curve, but they sometimes do exhibit a general pattern which may make it possible to study the relationship between Y and X. It is necessary here to

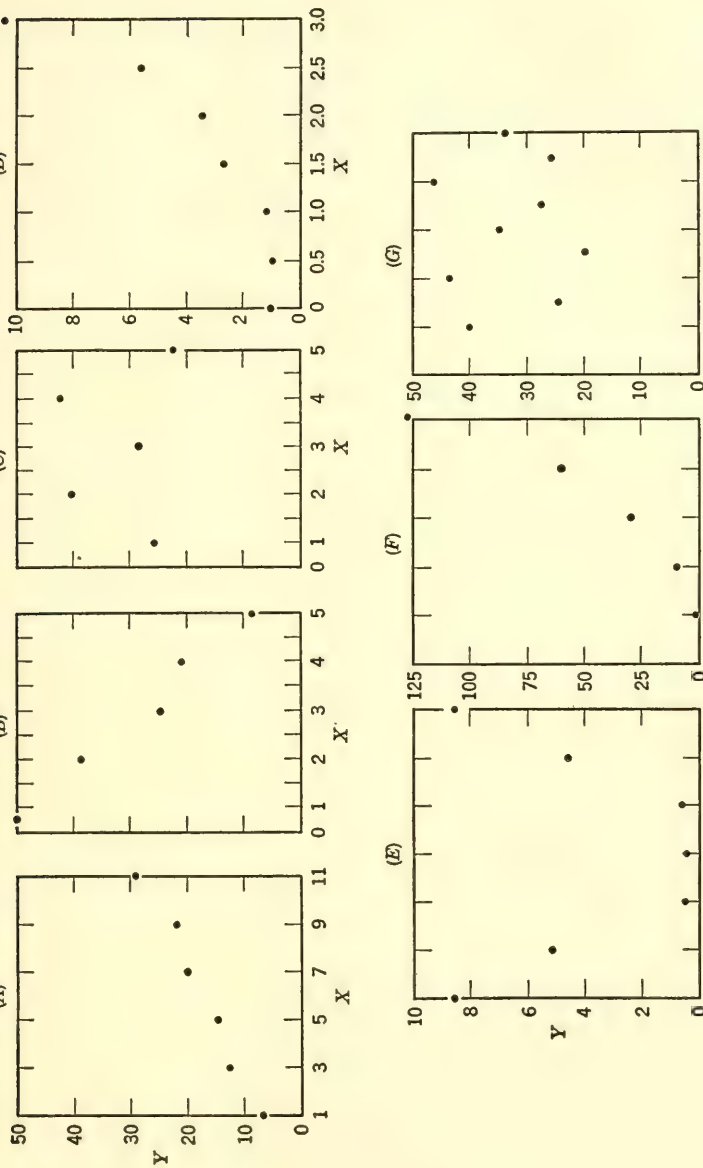


Figure 7.11. Illustrative scatter diagrams showing some of the various types of relationships between two variables, such as X and Y.

think in terms of general rather than precise curves as was done in algebra, where all points which belonged with a certain graph fell exactly on that graph. Data to which statistical analysis is applied are not so well behaved as that. It will be necessary later to learn how to decide which curve to choose as best describing the relationship between X and Y suggested by a scatter diagram; and it will not be expected that all the points will fall perfectly on the line finally chosen.

The following information can be derived from a careful inspection of Figures 7.11:

From (A)*:* Y definitely tends to increase uniformly (linearly) as X increases. On the average, Y increases about 13/6 units for each unit increase in X.

From (B)*:* Y decreases in proportion to the increase in X. Again the relationship can be briefly described as linear. More specifically, Y tends to decrease about 10 units for each unit increase in X. As a result the slope of the straight line which indicates the linear trend is said to be -10.

From (C)*:* Y has no apparent relation to X; hence the X measurement may as well be ignored in the statistical analysis of the measurements, Y.

From (D)*:* Y increases with X, but the increase is not uniform. In fact, Y increases more rapidly for large X's than for the smaller X's. This relationship between Y and X is called *curvilinear*. In this instance, it follows the non-linear mathematical law: $Y = 0.5e^X$, where e is the base for natural logarithms.

From (E)*:* As the measurement represented by X increases from -3 toward 0, the corresponding measurement, Y, tends to decrease in a non-linear manner. Thereafter, Y increases non-uniformly. As a matter of fact, the points on this scatter diagram tend to follow the curve, $Y = X^2$.

From (F)*:* Y tends to increase non-uniformly with X, as in (D), but the curve rises more sharply here.

From (G)*:* There is no apparent relationship between X and Y, as in (C).

Another point should be noted regarding the scatter diagrams of Figures 7.11. If the concomitant measurement, X, were to be ignored during an analysis of the data of Y corresponding to any of the situations except (C) and (G), a considerable portion of the variability of the Y's about their means would represent unnecessary variation in this sense. We know from (B), say, that if $X = 1$, the corre-

sponding Y is necessarily about 40 units larger than if $X = 5$ because there is definitely a downward trend of Y as X increases. If the X were ignored, *all* that observed difference of 40 units must be assigned to errors of measurement and/or to sampling accidents when, in fact, only about one unit should be so assigned. Methods will be presented later on in this chapter by which the apparent variation among the Y's can be reduced by taking account of the statistical relation between X and Y. However, nothing extensive will be done with non-linear trends.

PROBLEMS

1. Construct a scatter diagram for the following pairs of measurements, and draw in by eye a straight line which seems to you to best depict the trend of Y with X. Is the assumption that Y and X are linearly related a good one in your opinion?

X:	2	4	6	8	10	12	14	16	18	20
Y:	100	140	200	235	280	325	370	415	450	500

2. From the following sampling data estimate how much Y changes, on the average, for each unit increase in X.

X:	36	43	50	40	42	45	40	45	39	48
Y:	1.35	1.70	1.90	1.55	1.65	1.80	1.63	1.75	1.60	1.93

Ans. About 0.04.

3. Make a scatter diagram of the following pairs of observations and draw in a freehand line to summarize the way Y changes with X:

X, Y: 1,1; 2,2.5; 3,4.5; 4,6.0; 5,10.0; 6,14.5; 7,23.0; 8,35.0.

4. Would you approve of the assumption that the carotene and the nitrogen-free extract contents of pasture grasses are linearly related if assured that the following pairs of such values form a representative sample for pasture grasses of a given sort? Justify your decision.

X(NFE):	50	48	53	51	49	53	51	48
Y(Carotene):	.44	.26	.20	.24	.44	.23	.26	.34

5. The following are means and corresponding standard deviations obtained from samples of 10 observations, each drawn from an approximately normal population. Construct a scatter diagram and decide what, if any, relationship exists between the sampling mean and standard deviation from a normal population if these samples are representative of such populations. Plot $\bar{x}$ on the horizontal axis.

$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
58.4	6.04	53.4	10.33	63.0	9.33	60.4	11.40
55.7	9.61	60.9	12.64	52.9	6.13	61.2	6.45
62.0	6.70	56.7	8.03	58.1	8.15	67.7	10.21
61.7	11.40	55.0	11.51	53.6	9.83	56.4	5.81
54.0	10.52	59.3	9.05	57.8	10.92	54.4	10.78
59.0	11.09	54.1	8.75	60.0	8.17	62.8	8.43

6. Grades in elementary statistics and in mathematics of finance for the same students are given below. What do you conclude is the relation between a student's grades in these two subjects? Give evidence upon which your conclusion is based.

X(statistics):	94	83	91	98	80	82	61	81	58	90	85	75	75	70	92	62
Y(finance):	89	90	91	97	85	87	41	88	60	85	86	83	87	72	97	64

7. The following are weights of the larvae of honey bees at different ages:

Set A.	X(days):	1	2	3	4	5	6
	Y(milligrams):	2.0	4.3	23.1	93.1	148.7	295.5
Set B.	X(days):	1	2	3	4	5	6
	Y(log milligrams):	0.30	0.63	1.36	1.96	2.17	2.47

Construct scatter diagrams for each set separately and decide for which, if either, the assumption of a linear relation between X and Y appears to be justified. If either set produces a satisfactorily linear trend, estimate the slope of the best-fitting freehand line and state what information it provides. Should you make some allowance for the fact that you used a freehand line in a position with which others might disagree?

8. Make a scatter diagram for the following pairs of observations. Draw in a freehand line which appears to you to be the best-fitting straight trend line, and derive from this line an estimate of the Y which should correspond, on the average, to $X = 0$, 4.5, and 7.5:

X:	1	2	3	4	5	6	7	8	9	10
Y:	21	20	17	15	14	14	12	9	6	5

Ans. About 23.3, 15.1, and 9.7, respectively.

7.2 A METHOD FOR DETERMINING LINEAR REGRESSION (OR TREND) LINES

It is quite customary to use the term *regression line* to describe the line chosen to represent the relationship between two variables when this decision is based on sample points, as in a scatter diagram. The origin of the term *regression* probably lies in genetic studies of the tendency for offspring of parents who are well above, or below, the group average to go back, or "regress," toward that group average. The term *trend line* will be used interchangeably with regression line, even though the former is frequently associated with discussions of time series.

Previously, in this chapter, freehand lines have been used to depict the average change of one variable with respect to another. Such

a procedure, however, obviously is somewhat subjective because it depends quite a bit upon personal opinion. One of the chief purposes of numerical measurement and statistical analysis of such measurements is to free decisions based on relatively precise numbers from distortions which might result from the exercise of personal tastes and opinions. It is for this reason that it is desirable to be able to describe such a line by a method which will produce the same result no matter who uses it.

The types of relationships between two kinds of numerical measurements which were discussed in the preceding section are illustrative of sampling experiences involving errors of observation and measurement. The dots of the scatter diagram usually fail to fall exactly on any simple curve for one of two reasons: (*a*) Sampling errors or chance variations cause the values of Y, say, to be partially inaccurate. (*b*) There are real variations from the general trend of Y and X which, however, are of minor importance compared to the general trend and should be smoothed out in order that the general trend may be studied more effectively. The data of Table 7.21 and the corresponding scatter diagram of Figure 7.21 help to illustrate these points. The data in the table are considered to be a population of pairs of observations, that is, a bivariate population. For convenience these data have been grouped by 16-week weights (X) to the nearest pound in the scatter diagram of Figure 7.21.

The bivariate population of Table 7.21 possesses several characteristics which are of statistical interest and importance. These features are exhibited by Figure 7.21, from which it is learned: (*a*) There is a general upward trend of the 28-week weight, with increasing 16-week weight of the same bird. (*b*) Within each 16-week-weight class there is a frequency distribution of 28-week weights, and this distribution is relatively symmetrical about the mean 28-week weight for the class. (*c*) The means of the six 16-week-weight classes lie perfectly on a straight line with a slope of 1/2. Thus the true linear regression line passes through the points representing the true average Y's for the given X's. The slope of this true trend line is denoted by the Greek letter β (beta).

In a study based on samples the β is unknown, as is the exact location of the true linear trend line, and only the n pairs of sample measurements are available as a basis for making decisions about the linear trend line. For example, a random sample of 30 pairs (X, Y) was drawn from the bivariate population of Table 7.21, with the

TABLE 7.21

PAIRS OF OBSERVATIONS OF THE 16-WEEK WEIGHTS AND CORRESPONDING 28-WEEK WEIGHTS OF TURKEYS RAISED ON THE KANSAS STATE COLLEGE POULTRY FARM

(X is the 16-week weight in pounds; Y is the 28-week weight.)

X	Y	X	Y	X	Y	X	Y	X	Y
4.9	13.3	5.4	14.6	5.2	13.4	5.4	14.0	4.6	13.0
4.7	12.7	4.8	12.9	4.9	12.4	5.0	12.6	5.1	12.5
5.2	13.1	5.3	13.5	5.4	13.6	4.7	14.3	4.9	13.8
5.1	13.7	5.2	13.2	4.6	14.0	4.7	14.3	4.8	14.1
4.9	13.6	5.0	14.0	5.1	13.6	5.2	13.5	5.3	13.5
6.2	13.5	6.5	14.8	6.4	14.3	6.5	13.5	6.4	14.4
6.5	13.0	5.5	13.5	6.1	15.3	5.5	13.0	6.0	14.3
6.4	14.6	6.5	13.8	6.5	13.3	5.5	14.3	6.2	14.4
6.1	13.4	6.3	13.9	6.5	13.7	6.4	14.0	6.4	14.4
6.5	15.4	5.9	14.4	6.5	14.9	5.9	14.8	5.5	14.6
5.5	13.1	5.6	13.2	5.7	12.8	5.8	13.5	5.9	13.5
6.0	13.6	6.1	13.7	6.2	14.0	6.3	14.0	6.4	14.0
6.5	14.1	5.5	12.5	5.6	13.3	5.7	13.2	5.8	13.9
5.9	13.8	6.0	13.8	6.1	15.0	6.2	14.7	6.3	15.2
6.4	15.1	6.6	13.6	6.7	13.8	6.8	13.7	6.9	13.7
7.0	14.5	7.1	14.2	7.2	14.2	7.3	14.3	7.4	14.3
6.6	14.3	6.7	14.4	6.8	14.4	6.9	14.4	7.0	14.5
7.1	14.5	7.2	14.5	7.3	14.5	7.1	14.7	7.2	13.3
6.7	15.5	7.0	15.9	7.4	14.7	7.3	14.7	6.6	14.1
6.6	14.4	7.1	15.3	7.1	14.9	7.0	15.6	7.4	14.0
7.0	15.0	7.2	15.5	7.2	15.2	6.8	15.3	7.3	15.4
7.1	13.9	7.0	14.7	6.9	14.8	6.6	15.5	7.0	14.7
6.8	15.1	7.1	13.2	7.3	14.0	7.2	14.9	6.7	14.6
6.7	14.8	6.8	15.0	7.3	12.8	7.3	14.3	7.3	13.6
7.0	13.5	7.0	15.0	6.8	13.9	7.4	14.1	6.9	15.0
7.2	13.3	7.3	15.3	6.6	14.0	6.6	14.8	7.4	14.1
7.3	14.0	6.9	14.0	7.0	14.1	6.8	14.2	7.4	15.6
7.3	16.3	7.0	15.8	6.6	14.7	6.8	12.9	7.0	13.4
7.2	14.8	7.4	14.9	7.5	14.7	7.6	14.7	7.7	13.7
7.8	14.7	7.9	14.8	8.0	15.7	8.1	15.1	8.2	15.1
8.3	15.7	8.4	15.7	8.5	15.8	7.5	14.5	8.2	15.1
8.3	15.5	7.8	15.3	7.9	16.6	8.3	17.1	8.3	15.3
7.5	15.4	7.6	14.7	8.5	14.5	8.5	15.0	7.5	14.8
8.4	15.6	7.5	15.6	7.8	16.1	7.8	15.5	7.7	15.9
7.8	16.0	8.3	13.4	8.2	14.9	8.1	14.4	8.2	14.0
8.3	14.5	8.2	13.5	7.5	16.0	8.0	15.0	7.5	15.5
7.7	15.8	7.5	14.4	7.6	15.4	8.0	16.7	7.5	14.3
8.0	16.4	8.1	16.7	8.5	15.1	7.6	14.6	8.5	15.2
7.8	14.3	7.8	15.2	7.6	17.0	7.9	15.2	8.1	14.2

TABLE 7.21 (*Continued*)

PAIRS OF OBSERVATIONS OF THE 16-WEEK WEIGHTS AND CORRESPONDING 28-WEEK WEIGHTS OF TURKEYS RAISED ON THE KANSAS STATE COLLEGE POULTRY FARM

(X is the 16-week weight in pounds; Y is the 28-week weight.)

X	Y	X	Y	X	Y	X	Y	X	Y
7.5	13.8	8.5	14.5	7.9	15.9	8.1	14.4	8.3	15.0
8.2	15.6	8.4	14.8	7.8	13.7	8.2	15.0	7.6	13.3
7.6	14.2	8.1	14.1	8.0	13.5	8.1	15.2	8.2	15.3
7.5	15.5	7.9	15.8	7.7	16.3	7.8	17.0	7.8	13.5
8.0	15.6	8.3	16.2	8.4	14.8	8.4	14.4	7.8	13.2
8.5	16.9	8.0	13.7	8.2	15.5	8.3	15.0	8.2	15.7
8.0	13.6	8.0	14.3	8.4	15.0	8.5	15.1	8.2	15.4
8.0	14.9	8.1	15.0	8.5	15.2	7.8	14.0	7.6	15.2
7.6	14.1	8.5	15.5	7.7	14.3	7.5	14.6	8.1	14.3
8.1	14.4	8.0	13.8	8.2	13.8	7.6	14.9	8.2	15.3
8.1	15.3	7.5	14.1	8.5	15.3	8.1	15.4	7.5	15.0
8.0	15.0	7.8	14.0	8.5	16.0	7.9	15.5	8.1	16.5
8.5	15.6	7.5	13.9	7.5	14.3	7.5	14.0	7.6	14.0
7.8	14.0	8.8	14.2	9.0	15.5	8.9	15.0	8.8	14.4
9.0	15.5	9.5	17.2	9.4	16.0	9.0	15.8	9.1	16.2
8.7	16.3	9.4	16.9	9.4	16.0	9.0	15.1	8.7	16.3
9.2	15.7	9.1	15.2	9.2	16.4	9.4	16.5	8.6	15.2
8.7	15.7	9.0	15.9	8.9	14.6	9.0	16.3	9.0	16.1
8.6	16.4	9.2	14.3	8.6	14.1	9.2	15.1	8.6	13.9
8.6	15.0	8.7	16.2	9.0	16.1	9.1	15.2	9.2	15.1
9.2	14.0	9.0	14.4	9.0	15.5	8.6	14.4	8.6	15.0
8.7	15.2	9.1	16.5	8.8	15.5	8.8	14.5	9.1	15.3
8.6	15.9	8.6	15.1	9.2	15.3	9.2	15.5	8.6	14.6
8.6	15.3	9.2	15.3	9.0	15.6	9.1	15.4	8.7	17.0
8.8	14.8	9.0	15.5	9.2	15.8	9.0	15.5	9.0	15.5
8.6	14.1	8.7	15.6	9.1	16.2	8.8	15.1	8.6	14.9
8.7	14.9	8.7	14.7	8.8	14.8	8.9	14.8	9.0	14.9
9.1	14.6	9.2	15.8	9.3	15.8	9.4	15.9	8.6	16.0
8.7	16.0	8.8	16.1	8.9	16.6	9.0	16.6	9.1	16.7
9.2	16.8	9.3	16.9	9.4	17.0	8.8	17.2	9.3	17.6
8.6	14.5	8.7	14.5	8.8	14.2	8.9	15.0	9.0	15.0
9.1	14.8	9.2	14.4	8.9	15.1	8.6	17.4	10.1	16.7
9.8	15.4	9.5	14.8	9.6	15.0	9.8	15.2	9.8	16.0
9.6	15.5	9.8	15.2	9.5	17.0	9.5	17.8	10.4	16.7
9.5	16.7	10.5	17.1	9.8	16.4	9.5	15.5	9.6	16.0
9.7	16.2	9.8	16.2	9.9	15.8	10.0	16.4	10.1	16.5
9.5	15.7	10.2	16.8	10.3	16.9	10.4	15.9	9.5	14.8
9.7	15.6	9.9	14.5	10.1	15.8	10.3	14.9	10.3	15.3
9.6	14.4	10.5	17.5						

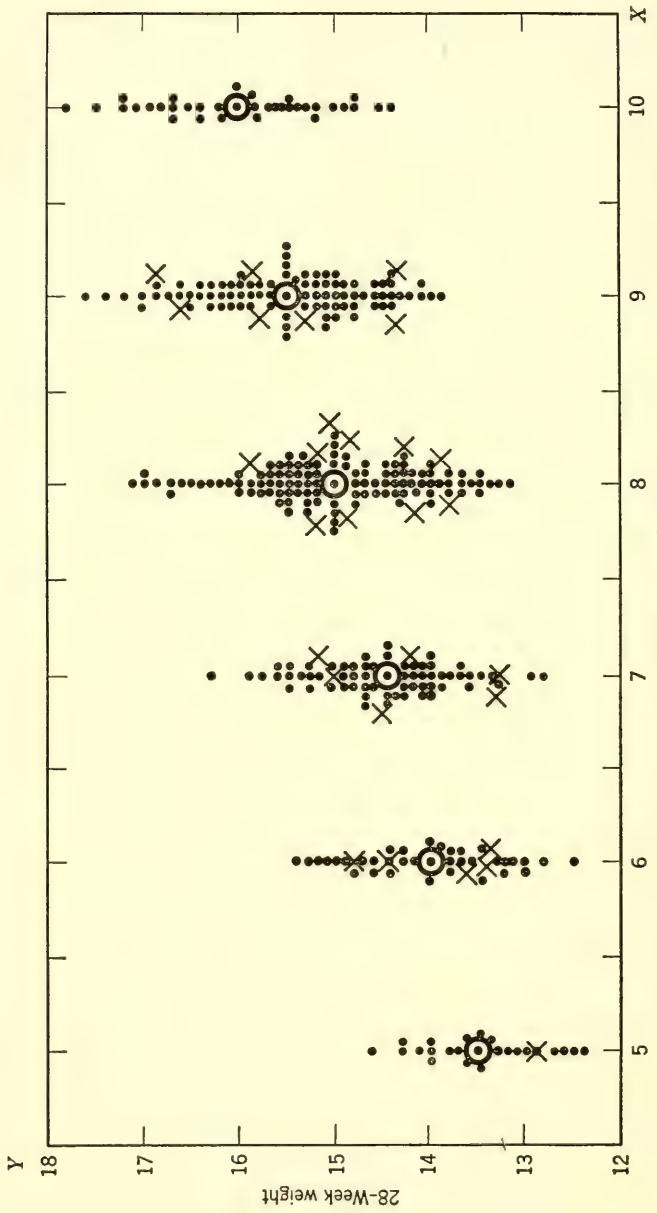


Figure 7.21. Bivariate distribution of 16- and 28-week weights of turkeys (see Table 7.21). The symbol ⊙ signifies the mean 28-week weight for the indicated X; the symbol $\times$ = a sample observation (see Figure 7.22).

results shown specifically in Table 7.22 and graphically as $\times$'s in Figure 7.21. The decision regarding linearity of trend and the estimation of any desired features of the true trend line (such as slope) must be accomplished from the information contained in the sample.

If the trend of one variable with another is linear, the relationship between the two kinds of measurements, X and Y, is of the form $Y = A + BX$, in which, for illustration, Y stands for the 28-week weight of a certain breed of turkey and X = the 16-week weight of

TABLE 7.22

A Random Sample of $n = 30$ Pairs (X, Y) from Table 7.21

X	Y	X	Y	X	Y	X	Y
4.8	12.9	7.2	13.3	7.9	15.9	8.8	14.4
6.5	14.8	7.0	15.0	8.0	14.9	9.2	15.7
6.4	14.4	6.9	14.7	8.1	15.0	9.0	15.9
5.5	13.5	6.8	15.1	8.0	13.8	8.6	14.4
6.1	13.4	7.2	13.3	7.6	14.9	9.1	15.3
6.0	13.6	7.5	14.3	7.5	14.1	9.0	16.6
7.4	14.3	8.5	15.1	7.5	13.9	9.2	16.8
7.0	14.5	8.5	15.2				

the same turkey. If all the observed pairs of measurements (X, Y) in Table 7.21 satisfy a linear equation perfectly, all the points of Figure 7.21 will lie exactly on the same straight line; and the relationship between the two variables will be perfectly linear. Moreover, the equation of the line can be determined from the coordinates of any two distinct points. Such obviously is not the case in Figure 7.21 because errors of measurement and uncontrollable fluctuations in the 28-week weights of turkeys which weighed the same at 16 weeks of age must be averaged out before the trend appears to be linear. By contrast with situations met in elementary algebra, where the equation is given and all appropriate points lie on the line, the present situation starts with the points given from sample observations, and the problem is to determine which straight line best fits these observations, and, it is hoped, best estimates the true linear regression line.

Assuming that a set of observations really does follow a linear trend quite well, how can a specific equation of the form $Y = A + BX$ be determined and also defended as the best straight line to be employed under the circumstances? The answer to this question depends upon the interpretation of the word "best." One interpretation, and the one most frequently accepted, can be illustrated by

means of the line drawn into Figure 7.22. Some points lie above this line, some lie below, at distances whose magnitudes can be measured by the lengths of the vertical lines which could be drawn connecting the points of the scatter diagram to the regression line. In a useful sense, the goodness of fit achieved by any line drawn among the points to depict their trend should be measured somehow in terms of the amounts by which the proposed line misses the points of the scatter diagram.

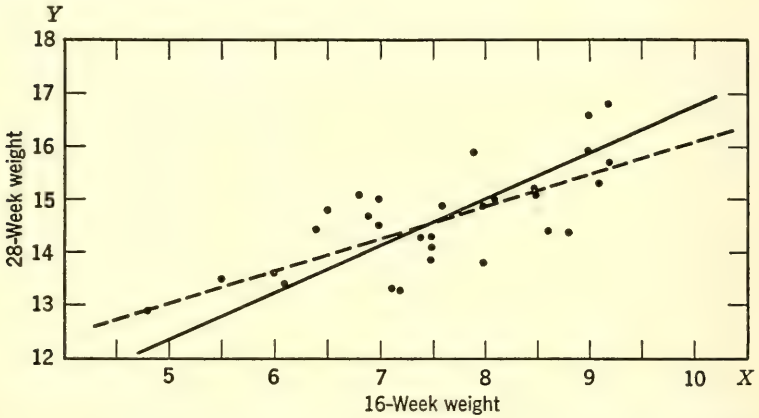


Figure 7.22. A random sample of 30 pairs of observations from the population of Figure 7.21 and Table 7.21. Free-hand line to indicate the trend as it might appear to the eye (———). Line determined by the method of least squares (------).

It will facilitate the discussion to introduce some symbolism before presenting the specific methods to be used in the determination of the equation of the regression line. For a given value X_i of the measurement X, let the corresponding value of Y be called Y_i if it was observed with X_i when the sample was taken. It will be designated as $\hat{Y}_i$ if it is calculated from the equation of the regression line. Also, let the general linear equation relating $\hat{Y}_i$ and X_i be written in the form

$$\hat{Y}_i = a + b(X_i - \bar{x}), \quad (7.21)$$ *

where a and b are the two constants which must be determined in order to have a specific trend line for a particular scatter diagram.

* This form and the notation do not agree entirely with some other textbooks, but they are used here for convenience. The $(X_i - \bar{x})$ is x_i so that the subsequent formulas and discussions come quite simply from this form of the equation for $\hat{Y}$. Some authors use other letters than a and b; and several others use b as herein, but their a = (above a) $- b\bar{x}$.

As stated above, the a and b will be calculated in terms of the collective amount by which a line misses the points of the scatter diagram. The a and the b also are considered estimates of the population parameters α and β in the true linear regression equation,

$$Y = \alpha + \beta(X - \mu), \tag{7.22}$$

where μ = the true mean of the X's.

For reasons given earlier for using $\Sigma(X - \bar{x})^2$ to measure variation about the mean instead of either $\Sigma(X - \bar{x})$ or $\Sigma|X - \bar{x}|$, it is found advisable to use $\Sigma(Y - \hat{Y})^2$ to measure the scatter of the Y's about the trend line. Therefore, the best-fitting straight line has been chosen as that one for which the $\Sigma(Y - \hat{Y})^2$ has the least possible size. This action makes the *standard deviation about the trend line* as small as possible. The mathematical process of achieving this goal produces formulas from which the a and the b can be computed. When these values are substituted into formula 7.21 a specific equation of a regression line is obtained. This line will have the property that the standard deviation about it is as small as it is possible to make it for any straight line. In other words, the variability of the Y's has been reduced as much as it can be in consideration of their linear trend with X.

The formulas for a and b are as follows:

$$a = \bar{y}, \text{ and } b = \frac{\Sigma[(X_i - \bar{x})(Y_i - \bar{y})]}{\Sigma(X_i - \bar{x})^2} = \frac{\Sigma(xy)}{\Sigma(x^2)}, \tag{7.23}$$

where $\bar{y}$ = mean of the Y's in the sample and y = the deviation of a Y from the mean, $\bar{y}$. The $\Sigma(xy)$—which the student has not met before in this book—is $(X_1 - \bar{x})(Y_1 - \bar{y}) + (X_2 - \bar{x})(Y_2 - \bar{y}) + \cdots + (X_n - \bar{x})(Y_n - \bar{y}) = x_1y_1 + x_2y_2 + \cdots + x_ny_n$.*

For the data of Table 7.22, $a = \Sigma(Y)/n = 439.0/30 = 14.63$, $b = \Sigma(xy)/\Sigma(x^2) = 23.0200/37.912 = 0.6072$, and $\bar{x} = \Sigma(X)/n = 226.8/30 = 7.56$. Therefore, since $\bar{y} + b(X - \bar{x}) = bX + (\bar{y} - b\bar{x})$,

$$\hat{Y} = 0.6072X + 10.04. \tag{7.24}$$

Students in a statistics course are in an unusually fortunate position because when they take samples from laboratory populations they can see readily how well, or poorly, certain features of their samples agree with the corresponding features of the populations being sampled.

* Experience shows that beginners in this field tend to think that $\Sigma(xy) = \Sigma(x) \cdot \Sigma(y)$. If the reader will recall that $\Sigma(x) = \Sigma(X - \bar{x}) = 0$ for any set of data —and likewise for $\Sigma(y)$—it becomes apparent that $\Sigma(xy)$ is not the same as $\Sigma(x) \cdot \Sigma(y)$ or it always would be zero. This obviously is untrue.

For example, the slope of the above sample estimate of the linear regression line is calculated to be $b = 0.6072$, whereas the true slope is known to be $\beta = 0.5000$. In actual practice, only the b is known, and it is necessary to measure its reliability as an estimate of β. This will be done later when the necessary techniques have been discussed; but it can be stated here that if the sample has been taken with the X's fixed—as suggested in Figure 7.21—so that there is no sampling error in X or in $\bar{x}$, the b as defined is an unbiased estimate of the parameter β.

The value $\hat{Y}$ which is obtained from formula 7.24 by substituting a particular value for X is described as *the estimated average Y for that X*. For example, if X is taken as 5, $\hat{Y} = 0.6072(5) + 10.04 = 13.1$, approximately. By reference to Figure 7.21 we learn that this estimate is somewhat low because the true average Y for turkeys weighing 5 pounds at 16 weeks of age is 13.5 pounds. If X is taken = 8, $\hat{Y} =$ 14.9 pounds, which is nearer to the true average Y of 15 pounds than was obtained when $X = 5$ and the true Y was 13.5 pounds. It will be seen in a later discussion that greater accuracy in estimating the true average Y is to be expected for X's near the mean X. There often are more sample data near the mean; but also errors in estimating the β will cause the ends of the trend line to be swung farther from the true position than is the middle of the line. In the above example the slope was $b = 0.6072$ instead of $\beta = 0.5000$; hence the line determined from the sample is too steep and therefore too low at the left-hand end. This appears to be the major reason why the estimate of the true average Y for $X = 5$ was too small. Of course, the general height of the sample line must be in error to some extent, and this also contributes to the inaccuracy of any estimate made from the sample trend line.

The method described for obtaining the straight line which fits a linear trend best is called *the method of least squares* because it makes the sum of squares of the vertical deviations of the points of the scatter diagram from the regression line the least it can be made for any straight line. Table 7.23 has been prepared to illustrate specifically the meaning of this minimization. Columns 6, 7, and 8 were obtained from the equation given over the right-hand side of the table. This equation represents a straight line which appears to the eye to fit the trend of the scatter diagram about as well as the line obtained by the method of least squares, as can be verified from Figure 7.22, which shows both lines.

It should be noted that the total of the fifth column of Table 7.23 is less than that of the eighth column. This will always be true no matter which straight line is used to obtain $\hat{Y}_j$ as long as the equation is

TABLE 7.23

ILLUSTRATION OF SOME FEATURES OF THE METHOD OF LEAST SQUARES USING DATA OF TABLE 7.22

	Method of Least Squares $\hat{Y} = 0.6072X + 10.04$				Freehand Straight Line $\hat{Y}_j = 0.88X + 7.92$		
X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	$\hat{Y}_j$	$Y - \hat{Y}_j$	$(Y - \hat{Y}_j)^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
4.8	12.9	12.95	−0.05	0.0025	12.14	+0.76	0.5776
6.5	14.8	13.99	+0.81	0.6561	13.64	+1.16	1.3456
6.4	14.4	13.93	+0.47	0.2209	13.55	+0.85	0.7225
5.5	13.5	13.38	+0.12	0.0144	12.76	+0.74	0.5476
6.1	13.4	13.74	−0.34	0.1156	13.29	+0.11	0.0121
6.0	13.6	13.68	−0.08	0.0064	13.20	+0.40	0.1600
7.4	14.3	14.53	−0.23	0.0529	14.43	−0.13	0.0169
7.0	14.5	14.29	+0.21	0.0441	14.08	+0.42	0.1764
7.2	13.3	14.41	−1.11	1.2321	14.26	−0.96	0.9216
7.0	15.0	14.29	+0.71	0.5041	14.08	+0.92	0.8464
6.9	14.7	14.23	+0.47	0.2209	13.99	+0.71	0.5041
6.8	15.1	14.17	+0.93	0.8649	13.90	+1.20	1.4400
7.2	13.3	14.41	−1.11	1.2321	14.26	−0.96	0.9216
7.5	14.3	14.59	−0.29	0.0841	14.52	−0.22	0.0484
8.5	15.1	15.20	−0.19	0.0361	15.40	−0.30	0.0900
8.5	15.2	15.20	0	0	15.40	−0.20	0.0400
7.9	15.9	14.84	+1.06	1.1236	14.87	+1.03	1.0609
8.0	14.9	14.90	0	0	14.96	−0.06	0.0036
8.1	15.0	14.92	+0.08	0.0064	15.05	−0.05	0.0025
8.0	13.8	14.90	−1.10	1.2100	14.96	−1.16	1.3456
7.6	14.9	14.65	+0.25	0.0625	14.61	+0.29	0.0841
7.5	14.1	14.59	−0.49	0.2401	14.52	−0.42	0.1764
7.5	13.9	14.59	−0.69	0.4761	14.52	−0.62	0.3844
8.8	14.4	15.38	−0.98	0.9604	15.66	−1.26	1.5876
9.2	15.7	15.63	+0.07	0.0049	16.02	−0.32	0.1024
9.0	15.9	15.50	+0.40	0.1600	15.84	+0.06	0.0036
8.6	14.4	15.26	−0.86	0.7396	15.49	−1.09	1.1881
9.1	15.3	15.57	−0.27	0.0729	15.93	−0.63	0.3969
9.0	16.6	15.50	+1.10	1.2100	15.84	+0.76	0.5776
9.2	16.8	15.63	+1.17	1.3689	16.02	+0.78	0.6084
			Sums	12.9226 = $\Sigma(Y - \hat{Y})^2$			15.8926

not obtained by the method of least squares, and as long as sufficient accuracy is kept in the calculations to pick up small differences.

This is the basis for the statement that the method of least squares makes the standard deviation of the Y's from the trend line as small as possible for any straight line, which is a strong argument for the use of this line in practice.

PROBLEMS

1. Obtain the linear equation whose graph fits the points of Figure 7.11*B* best in the sense of the method of least squares. Graph the line on the scatter diagram and indicate graphically those deviations whose sum of squares is the least possible for any straight line.

2. Do as in problem 1, for Figure 7.11*A*. Also compute $\Sigma(Y - \hat{Y})^2$.

Ans. $\hat{Y} = 2.16X + 4.54$; $\Sigma(Y - \hat{Y})^2 = 9.77$.

3. By what average amount would you expect Y to increase for a unit increase in X if the data corresponding to Figure 7.11*A* constitute a representative sample of some two-variable population?

4. Compare the $\Sigma(Y - \hat{Y})^2$ and $\Sigma(Y - \bar{y})^2$ for the data of Figures 7.11*B* and *G*. What conclusions can you draw?

5. Use the method of least squares to estimate for the data of Figure 7.11*B* the average value of Y for $X = 1.5$, 2.5, 3.5, and 4.5, respectively.

6. Make up two sets of 10 pairs of observations each and such that b is about 2 in one set and about -3 in the other.

7. Write down the equation of a trend line with slope = 5 and for which $\hat{Y} = 10$ when $X = 4$. Graph this line, and then construct a scatter diagram which fits the trend and has $\Sigma(Y - \hat{Y})^2 = 50$.

8. Do as in problem 7, with slope $= -3$ and everything else the same.

9. Assign row and column numbers to the data in Table 7.21. Then draw two random samples of 30 pairs each—as in Table 7.22—and obtain the least-squares regression line from each sample. Plot these lines on their corresponding scatter diagrams and discuss their differences. (Round off each X to the nearest pound before doing your computations.)

10. "Cull" the flock of Table 7.21 at 16 weeks of age by eliminating all turkeys which weighed under 6 pounds at that time, then do as in problem 9. Would β still be 0.5 for this population?

7.3 MEASUREMENT OF THE VARIATION ABOUT A LINEAR TREND LINE DETERMINED BY THE METHOD OF LEAST SQUARES

If measurements are taken on but one normally distributed variable, Y, the variability, or dispersion, of the Y_i should be measured by means of the standard deviation about the mean, and estimated from $s_Y = \sqrt{\Sigma(Y_i - \bar{y})^2/(n-1)}$ because $s_Y{}^2$ is an unbiased and highly efficient sampling estimate of $\sigma_Y{}^2$. The variation measured by s_Y is then considered to be sampling variation. However, if for each Y_i there is an associated measurement, X_i, and if the X's and Y's tend to be linearly related, not all the apparent variability among the Y_i should be assigned to mere sampling errors. Part of it can be accounted for in terms of the varying X_i associated with the Y_i. For example, if Y tends to increase about 5 units for each unit increase in the magnitude of X the Y associated with $X = 10$ is expected to be

about 15 units greater than the Y for $X = 7$; hence some of any observed difference between the Y for $X = 7$ and the Y for $X = 10$ can be accounted for and need not be considered as sampling error.

Graphically the preceding remarks mean that if Y and X can be considered to be linearly related the standard deviation of the Y_i should be calculated from the trend line rather than from the horizontal line: $\hat{Y} = \bar{y}$. This means that the quantity $\Sigma(Y - \hat{Y})^2$ should be employed in this calculation instead of $\Sigma(Y - \bar{y})^2$. However, the divisor in this calculation will not be $(n - 1)$ as it is for s_Y, above.

The divisor needed in the computation of the standard deviation about the trend line is $(n - 2)$. The reason for this cannot be given conclusively without mathematical analysis which is not appropriate to this book; but it can be rationalized in the following manner. Suppose that a sample of 5 observations on X and Y simultaneously were as follows:

X:	1	2	3	4	5	$\bar{x} = 3$
Y:	5	4	6	8	7	$\bar{y} = 6$

It is readily determined that $\hat{Y} = 0.80X + 3.6$; hence the following table can be set up for purposes of illustration:

X:	1	2	3	4	5
Y:	5	4	6	8	7
$\hat{Y}$:	4.4	5.2	6.0	Y_4	Y_5
$(Y_i - \hat{Y}_i)$:	0.6	−1.2	0	?	?

What are the deviations from the trend line for $X = 4$ and $X = 5$, respectively? The fact that $\Sigma(Y_i - \hat{Y}_i) = 0$ will be found to account for one of these deviations. The fact that $b = 0.80$ will allow the determination of the second unknown deviation.

Let the unknown deviations $Y_4 - \hat{Y}_4$ and $Y_5 - \hat{Y}_5$ corresponding to $X = 4$ and $X = 5$ be denoted by v and w, respectively. It follows that

$$\Sigma(Y_i - \hat{Y}_i) = 0.6 + (-1.2) + 0 + v + w = 0,$$

which reduces easily to

$$v + w = 0.6. \tag{7.31}$$

The slope of a straight line can be computed by determining the amount by which Y changes for any chosen change in X, and by taking the ratio of the former to the latter. For example, if it is determined by measurement on the graph or by substitution into a mathematical formula that for the interval from $X = 1$ to $X = 5$ the height of the

straight line above the horizontal axis increases from 10 to 30, the slope of this line is measured by $20/4 = 5$. Hence, in the situation of the preceding paragraph the slope is given by $(\hat{Y}_5 - \hat{Y}_4)/(X_5 - X_4)$. But $(X_5 - X_4) = 1$, and the slope is known to be 0.80; hence $(\hat{Y}_5 - \hat{Y}_4) \div 1 = 0.80$. In order to transform this equation into one involving v and w consider the following two equations:

$$Y_5 - Y_4 = Y_5 - Y_4$$

$$\hat{Y}_5 - \hat{Y}_4 = 0.80.$$

When the left and right members of the second equation are subtracted from the corresponding members of the first equation, it is found that

$$(Y_5 - \hat{Y}_5) - (Y_4 - \hat{Y}_4) = Y_5 - Y_4 - 0.80;$$

but $Y_5 - \hat{Y}_5 = w$, $Y_4 - \hat{Y}_4 = v$, and $Y_5 - Y_4 = -1$; therefore,

$$v - w = 1.8. \tag{7.32}$$

When equations 7.31 and 7.32 are solved simultaneously it is found that $v = 1.2$ and $w = -0.6$. Hence, two of the deviations from the trend line can be calculated from the size of b and from the fact that $\Sigma(Y - \hat{Y}) = 0$. Although there are five actual deviations from the linear trend line, only three (any three) of them actually should be considered chance deviations from the regression line determined from the sample. Hence, in the present problem, $n - 2 = 5 - 2 = 3$ will be used as the divisor of $\Sigma(Y_i - \hat{Y}_i)^2$ in the computation of the standard deviation about the linear trend line.

This divisor, $n - 2$, is generally called the *number of degrees of freedom* for the estimated standard deviation about the linear trend line, just as the number, $n - 1$, is the number of degrees of freedom for the estimated standard deviation, s_Y, about the mean.

With the above discussion as a background, the formula for the estimated standard deviation of the Y_i about the trend line becomes

$$s_{y \cdot x} = \sqrt{\Sigma(Y_i - \hat{Y}_i)^2/(n - 2)}\,, \tag{7.33}$$

wherein the symbol, $s_{y \cdot x}$, is read "s sub y dot x."

For the data used as illustration in this section, $\Sigma(Y_i - \hat{Y}_i)^2 = 3.60$, $n - 2 = 3$; hence $s_{y \cdot x} = \sqrt{3.60/3} = 1.10$. This is a measure of the variation among the Y-measurements which remains unexplained even after the linear trend with X has been taken into account. When the trend with X is ignored, $s_Y = \sqrt{10/4} = 1.58$; so $s_{y \cdot x}$ is 0.48 of a unit smaller than s_Y. In other words, the variability of the Y_i (as measured

by the standard deviation) has been reduced $100(0.48)/1.58 = 30.4$ per cent by taking the linear relation between the two measurements into account statistically. Such success in accounting for part of the variation among the measurements, Y_i, clearly is important in statistical analyses because the only occasion for such analyses arises as a result of variability among numerical measurements.

The standard deviation about the trend line, $s_{y \cdot x}$, also is specifically useful in certain applications of linear trend analysis, two of which will be considered. The regression coefficient, b, estimates the average change in the Y-measurement for each unit increase in the X-measurement. Its accuracy as such a measure is of interest, and its accuracy is measured by its standard deviation. The standard deviation of b is shown in more advanced statistics courses to be

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\Sigma x^2}}. \tag{7.34}$$

For the data of Table 7.22: $\Sigma(Y - \hat{Y})^2 = 12.9226$, $\Sigma x^2 = 37.9120$, $n = 30$, and hence $s_{y \cdot x} = \sqrt{12.9226/28} = 0.679$. Therefore, $s_b = 0.679/\sqrt{37.9120} = 0.110$, approximately.

It can be shown that the ratio,

$$t = (b - \beta)/s_b, \tag{7.35}$$

where β = the true regression coefficient which is estimated by b, follows the same t-distribution as that summarized in Table IV with $n - 2$ degrees of freedom. Therefore, a confidence interval can be computed for β, and it can be interpreted in the manner previously shown. For Table 7.22, the 95 per cent confidence interval is obtained as follows:

$$-2.05 \leq (0.6072 - \beta)/0.110 \leq +2.05$$

will be a true inequality for 95 per cent of all samples with 28 degrees of freedom. Hence the 95 per cent confidence interval is found to be as follows after some simplification of the preceding inequality:

$$0.38 \leq \beta \leq 0.83. \tag{7.36}$$

It would be concluded in practice that the slope of the true linear regression line is some value between 0.38 and 0.83, but it is recognized that there are 5 chances in 100 that the sample has led us to a false statement. A more useful statement might be that it is estimated from (7.36) that a turkey which is one pound heavier than another at 16 weeks of age will, on the average, be 0.38 to 0.83 pound

heavier at 28 weeks of age. That is, the lighter turkeys at 16 weeks tend to catch up some, but they usually remain 0.38 to 0.83 pound lighter at 28 weeks for each pound that they were lighter at 16 weeks of age.

Another application of linear trend analysis which makes use of $s_{y\cdot x}$ is one in which Y is to be estimated for some unobserved value of X; for instance, for $X = 9.5$ pounds at 16 weeks. If X is set equal to 9.5 in formula 7.24, $\hat{Y} = 0.6072(9.5) + 10.04 = 15.8$ pounds at 28 weeks of age. How reliable is this estimate? A look at the scatter diagram leaves only the impression that this estimate should be fairly reliable; hence a more specific measure of its accuracy is needed. The standard deviation of $\hat{Y}$ is given by the following formula:

$$s_{\hat{Y}} = s_{y\cdot x}\sqrt{1/n + (X - \bar{x})^2/\Sigma x^2}\,, \tag{7.37}$$

where X is the value used to calculate the $\hat{Y}$. This estimate of the standard deviation of $\hat{Y}$ is based on $n - 2$ degrees of freedom, as explained earlier. It will be convenient in subsequent discussions to add "with $n - 2$ D/F" after an estimate of this sort to indicate the number of chance deviations upon which the estimate is based. In the example considered in this paragraph,

$$s_{\hat{Y}} = 0.679\sqrt{1/30 + (9.5 - 7.56)^2/37.912} = 0.679(.364)$$
$$= 0.247, \text{ with } 28\ D/F.$$

This standard deviation applies when the X's have been chosen in advance and are not subject to sampling error. As noted earlier, the b is then an unbiased sample estimate of the population parameter, β. Under these circumstances the formula 7.37 can be partially rationalized in lieu of a more rigorous demonstration of its validity. The $\hat{Y}$ for a particular X, say X_i, is obtained from $\hat{Y} = \bar{y} + (X_i - \bar{x})b = \bar{y} + x_i b$. Hence the variance of $\hat{Y}_i$ is obtained from the variance of a sum, $\bar{y} + x_i b$, in which the x_i is a fixed number. The variance of $\bar{y}$ for this particular X will be one-nth of the variance about the trend line, or $s_{y\cdot x}{}^2/n$. The variance of b is $s_{y\cdot x}{}^2/\Sigma x^2$, as noted earlier, and x_i is a constant; hence the variance of $x_i b = x_i{}^2 \cdot s_{y\cdot x}{}^2/\Sigma(x^2)$. *If* the variance of the sum, $\bar{y} + x_i b$, is just the sum of the variances of those two terms, it follows that

$$s_{\hat{Y}}{}^2 = \frac{s_{y\cdot x}{}^2}{n} + \frac{x_i{}^2 \cdot s_{y\cdot x}}{\Sigma(x^2)} = s_{y\cdot x}{}^2(1/n + x_i{}^2/\Sigma(x^2)$$

so that $s_{\hat{Y}} = s_{y\cdot x}\sqrt{1/n + x_i{}^2/\Sigma(x^2)}$, as in formula 7.37 for a particu-

lar $X = X_i$. It is true that the variance of $\bar{y} + x_i b$ is the sum of the variances of the two terms, but this will not be proved here.

It can be shown that the ratio

$$t = \frac{\hat{Y} - \mu_{y \cdot x}}{s_{\hat{Y}}}, \tag{7.38}$$

where $\mu_{y \cdot x}$ is the true average Y for the given X, follows the t-distribution with $n - 2$ degrees of freedom. This fact makes it possible to place a confidence interval on $\mu_{y \cdot x}$ with any appropriate confidence coefficient.

The meaning of the $\mu_{y \cdot x}$ can be made clearer by reference to Figure 7.21. For each particular X there is a frequency distribution of the corresponding Y's. This distribution of Y's has a true arithmetic mean, which is the $\mu_{y \cdot x}$ for that X.

If we wish to make an interval estimate which applies to an individual rather than to a group mean, we must take account of the greater variation exhibited by such individuals as compared to the group. For example, suppose that a study has been made of the relationship between the ages of Kansas females and their basal metabolism rates as expressed in calories per square meter of surface area per hour. It is supposed that the age interval chosen is such that a linear relationship exists between these two variables, and that the least-squares equation for $\hat{Y}$ has been obtained from a sample. Suppose, furthermore, that the equipment needed to determine the basal rate is not available in a certain area, and a Kansas woman 25 years of age wishes an estimate of *her* basal metabolism rate as a matter of interest. The best point estimate is the $\hat{Y}$ calculated for $X = 25$; but when an interval estimate is needed—and it is more useful in the present problem—account must be taken of the fact that this woman is not supposed to be an average person representing all those who are Kansas females 25 years of age. She is regarded as one particular person who wishes an estimate of her own basal rate. In this circumstance the variance of $\hat{Y}$ used earlier in this section is not correct because it includes only two sources of variation: one from the mean and one from the sampling regression coefficient. In the present problem a third source must be included, namely, individual variation about the mean. When the particular X has been taken into account, this additional variance is just $s_{y \cdot x}{}^2$; hence—again it turns out that this can be added to the other two components—we obtain the following formula for the variance of the $\hat{Y}$ for an individual:

$$s_{\hat{Y}}{}^2 = s_{y \cdot x}{}^2 \left[1 + \frac{1}{n} + \frac{(X - \bar{x})^2}{\Sigma(x^2)} \right].$$

When this change in $s_{\hat{Y}}$ is made in formula 7.37 we obtain the formula for $s_{\hat{Y}}$ which is employed in the following t-ratio:

$$t = \frac{\hat{Y} - \mu_{y \cdot x_i}}{s_{\hat{Y}}}, \tag{7.38a}$$

where $\mu_{y \cdot x_i}$ = the true Y-value for the ith individual for whom $X = X_i$.

Formula 7.38a and the usual methods make it possible to obtain a confidence interval on $\mu_{y \cdot x_i}$ with any specified confidence coefficient. It should be both clear and reasonable that such a confidence interval will be longer than a corresponding one from formula 7.38 because the standard deviation is larger.

Problem 7.31. You are about to buy one turkey which weighs 6.5 pounds at 16 weeks of age, and you are going to keep it until it is at least 28 weeks of age. What is the 94 per cent confidence interval on its 28-week weight, assuming it comes from the population sampled in Table 7.22?

It was seen in the discussion of Table 7.22 that $\hat{Y} = 0.6072X + 10.04$, which = 13.99 pounds for $X = 6.5$ pounds. Also, $s_{y \cdot x} = 0.679$ pound, and $\Sigma(x^2) = 37.9120$; hence by formula 7.39, after taking the square root,

$$s_{\hat{Y}} = 0.679\sqrt{1 + 1/30 + (6.5\text{–}7.56)^2/37.9120} = 0.700$$

is the standard deviation of $\hat{Y}$ for $X = 6.5$ pounds at 16 weeks of age. Therefore, $t = (13.99 - \mu_{y \cdot x_i})/0.700$, and the 94 per cent confidence interval is derived from

$$-2.19(0.700) \leq (13.99 - \mu_{y \cdot x_i}) \leq 2.19(0.700).$$

It is found that

$$\text{CI}_{94} \text{ is } 12.5 \leq \mu_{y \cdot x_i} \leq 15.5,$$

to the nearest one-half pound.

Notice that the situation just considered clearly is one in which the confidence interval for a particular turkey is required. You do not have a group of turkeys so that the high 28-week weights of some of them can be expected to offset the low 28-week weights of others. Therefore, you must face the fact that this particular turkey's weight at 28 weeks of age may be quite low, as well as quite high, for turkeys weighing 6.5 pounds at 16 weeks of age.

Problem 7.32. Suppose that problem 7.31 is changed to state that you have bought a rather large flock of 6.5-pound turkeys each 16 weeks of age. Compute the CI_{94} appropriate to this new situation.

The only change in the computations is that the standard deviation of $\hat{Y}$ now is

$$s_{\hat{Y}} = s_{y \cdot x}\sqrt{1/n + x^2/\Sigma(x^2)} = 0.679(0.251) = 0.170 \text{ pound}$$

instead of the 0.700 pound obtained for the individual. It follows that the required confidence interval is:

$$\text{CI}_{94}\text{: } 13.5 \text{ pounds} \leq \mu_{y \cdot x} \leq 14.5 \text{ pounds},$$

to the nearest one-half pound. This is a narrower interval than is obtained for problem 7.31, as should be expected.

PROBLEMS

1. Compute s_Y and $s_{y \cdot x}$ for the data for Figures 7.11*A* and *G* and relate their comparative sizes to the scatter diagrams.

2. Work problem 1 for Figures 7.11*B* and *G*. Does the downward trend of the points on a scatter diagram, as contrasted with an identical upward trend, have anything to do with the comparison between s_Y and $s_{y \cdot x}$?

Ans. *B*: $s_Y = 16.0$, $s_{y \cdot x} = 3.1$; *G*: $s_Y = 9.2$, $s_{y \cdot x} = 9.8$.

3. Referring to Figure 7.11*E* and the associated data, compute and compare s_Y and $s_{y \cdot x}$ as in problem 1. Could you have predicted from the scatter diagram that they would be of essentially the same magnitude? Why?

4. By visual inspection of Figures 7.11*C*, *D*, and *F* what do you conclude about the comparative sizes of s_Y and $s_{y \cdot x}$ for each figure?

5. For the two sets of data in problem 7, section 7.1, compute the percentage reduction in the standard deviation of the Y_i achieved if variability is measured about the linear trend line rather than about the line $\hat{Y} = \bar{y}$ for each set. Discuss the two results obtained for the two sets in terms of the curvilinear trend in one set.

6. Use the data of problem 1, section 7.1, to estimate the average Y for $X = 13$. Also compute the standard deviation of this estimate, first considering a group with $X = 13$ and then for an individual with $X = 13$.

Ans. $\hat{Y} = 345.7$ when $X = 13$; $s_{\hat{Y}} = 0.53$; 1.01.

7. Use the data of set *B*, problem 7, section 7.1, to place 92 per cent confidence limits on the log(weight) of the 7-day-old bee larvae of the kind represented by this sample. Interpret these limits.

8. Compute the 99 per cent confidence interval on β for problem 1, section 7.1, and draw appropriate conclusions. *Ans.* CI_{99}: $21.2 \leq \beta \leq 23.0$.

9. The following data express the farm population (as defined for the 1950 census) as a percentage of the total U. S. population:

Year:	1940	1941	1942	1943	1944	1945	1946	1947	1948
Per Cent:	21.8	21.5	20.6	18.8	17.7	17.3	17.9	17.9	16.9
				1949	1950	1951			
				16.7	16.0	15.0			

These are not sampling data, but the fitting of a trend line to these data may be useful anyway. For example, if the war years, 1943 to 1945, inclusive, are

ignored, the downward trend in percentage farm population is quite closely represented by a straight line. Make a scatter diagram of the above data, omit 1943, 1944, and 1945 from further consideration, fit a linear trend line by the method of least squares, and then read from the line the approximate percentages for the years omitted. Would the discussions of estimates of the standard deviation and the formulas given in this book for them be appropriate here? Give reasons for your answer.

10. Referring to problem 9, could you use the equation obtained there to predict satisfactorily the percentage farm population for 1953? For 1960? Justify your answers.

7.4 COEFFICIENTS OF LINEAR CORRELATION

It is not always desirable—or even appropriate—to obtain an equation for the linear relation between the two types of measurements being studied, as was done earlier in this chapter. It may be better to describe the relationship as linear, and to give a standard, unitless, measure of its strength, or closeness. This is the purpose of a coefficient of linear correlation.

Although correlation coefficients are widely used, and often without attention to the satisfaction of necessary assumptions, it should be kept in mind that, strictly speaking, both X and Y must be random variables which follow normal frequency distributions. This will be assumed to be true in the following discussion of this section.

It has been seen that the variance of the observed Y's about the least-squares regression line depends on the size of $\Sigma(Y - \hat{Y})^2$, in which $\hat{Y} = \bar{y} + bx$. Hence the magnitude of this variance depends on $\Sigma(Y - \bar{y} - bx)^2 = \Sigma(y - bx)^2 = \Sigma(y^2 - 2bxy + b^2x^2)$. But the last summation can be computed in three parts as follows:

$$\begin{aligned}\Sigma(y^2 - 2bxy + b^2x^2) &= \Sigma(y^2) - 2b\Sigma(xy) + b^2\Sigma x^2,\\ &= \Sigma(y^2) - 2\left[\frac{\Sigma(xy)}{\Sigma(x^2)}\right]\cdot[\Sigma(xy)]\\ &\quad + \left[\frac{\Sigma(xy)}{\Sigma(x^2)}\right]^2\cdot\Sigma(x^2),\\ &= \Sigma(y^2) - \frac{2[\Sigma(xy)]^2}{\Sigma(x^2)} + \frac{[\Sigma(xy)]^2}{\Sigma(x^2)};\end{aligned}$$

hence,

$$\Sigma(Y - \hat{Y})^2 = \Sigma(y^2) - \frac{[\Sigma(xy)]^2}{\Sigma(x^2)}. \tag{7.41}$$

Therefore, it is clear that the observed variability of the Y's about the regression line will be large or small according to the size of $[\Sigma(xy)^2 \div \Sigma(x^2)]$ compared with the size of $\Sigma(y^2)$. If $[\Sigma(xy)]^2/\Sigma(x^2)$ is multiplied by $\Sigma(y^2)/\Sigma(y^2)$—which equals 1 and only changes the form of the quantity by which it is multiplied—it follows from (7.41) that the $\Sigma(Y - \hat{Y})^2$ can be expressed as follows:

$$\Sigma(Y - \hat{Y})^2 = \Sigma(y^2)\left[1 - \frac{[\Sigma(xy)]^2}{\Sigma(x^2)\cdot\Sigma(y^2)}\right].$$

Clearly, the quantity $\dfrac{[\Sigma(xy)]^2}{\Sigma(x^2)\cdot\Sigma(y^2)}$ has the following statistical features:

(*a*) Its value cannot be less than zero nor more than one because it is essentially zero or positive, and if it exceeded one the sum of squares of deviations from the trend line would be negative, which is absurd.

(*b*) If this quantity is near zero there is about as much scatter of the sample points about the trend line as about the horizontal line $\hat{Y} = \bar{y}$; hence there is little or no linear trend.

(*c*) If this quantity is near one there is very little scatter about the regression line; hence the sample points lie quite close to that line.

(*d*) As the size of this quantity varies—for different samples—from zero to one the scatter of the sample points about the least-squares regression line varies from a completely trendless, shot-gun, pattern to a perfect fit to a linear trend line.

(*e*) This quantity is unitless so that the features noted above are true regardless of the units in which Y and X are measured.

(*f*) In its present form this quantity cannot distinguish between positive and negative slopes of trend lines, but its square root would have the same sign as b and would make this distinction if the square root of the denominator were always taken as positive.

It follows then that the square root noted in *f*,

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\cdot\Sigma(y^2)}}, \tag{7.42}$$

is a unitless number within the range $-1 \leq r \leq +1$ which indicates the direction and strength of the observed linear trend. This number, r, is called *the product-moment coefficient of linear correlation* between two measurements X and Y. It obviously is subject to sampling variations and therefore has a sampling distribution. It is a sampling estimate of a corresponding population parameter indicated

by the Greek letter ρ (rho), which succinctly describes the degree of scatter of the population points about the true linear trend line as, for example, in Figure 7.21. If $\rho = 1$, all the points will lie on the regression line. Since they do not—in Figure 7.21—there is sampling error in the estimation of ρ, and hence the r varies from sample to sample. This is similar to the situation when the true regression coefficient, β, was being estimated from samples.

If the ρ is zero, all the sampling estimates r_i will not be zero but will have a sampling distribution which is nearly normal in form. In such circumstances it can be shown that the following ratio follows the t-distribution with $n - 2$ D/F. Thus

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}} \tag{7.43}$$

can be used in the usual manner to test the hypothesis $H_0(\rho = 0)$. As was seen in earlier discussions, H_0 will be rejected whenever the size of t becomes so great that it is unreasonable—according to some predetermined standard—to believe that this t is the product of sampling variation.

It is more difficult to place a confidence interval on ρ than on β because r is not nearly normally distributed when $\rho \neq 0$. This process of computing a confidence interval on ρ will be discussed and developed somewhat heuristically by means of the empirical data found in Table 7.41. These data were obtained by drawing random

TABLE 7.41

OBSERVED SAMPLING DISTRIBUTIONS OF r AND $z = 1/2 \log_e [(1 + r)/(1 - r)]$ FOR $n = 12$ AND $\rho = +.749$

r-interval	f	z-interval	f
.890–1.000	7	1.70–1.89	1
.790– .889	32	1.50–1.69	3
.690– .789	48	1.30–1.49	9
.590– .689	52	1.10–1.29	22
.490– .589	23	0.90–1.09	43
.390– .489	12	0.70–0.89	56
.290– .389	8	0.50–0.69	29
.190– .289	5	0.30–0.49	19
.090– .189	2	0.10–0.29	7
−.010– .089	1	−0.10–0.09	1
Total	190	Total	190

samples of pairs of values of X and Y from Table 7.21 for which $\rho = .749$. This population was considered to be approximately a normal bivariate population. The 190 sampling r's thus obtained are summarized in Table 7.41 along with the corresponding z's (see discussion of z below Figure 7.41). The distribution of r for a ρ so large as this is definitely skewed, as can be seen to some extent in Table 7.41 and in Figure 7.41.

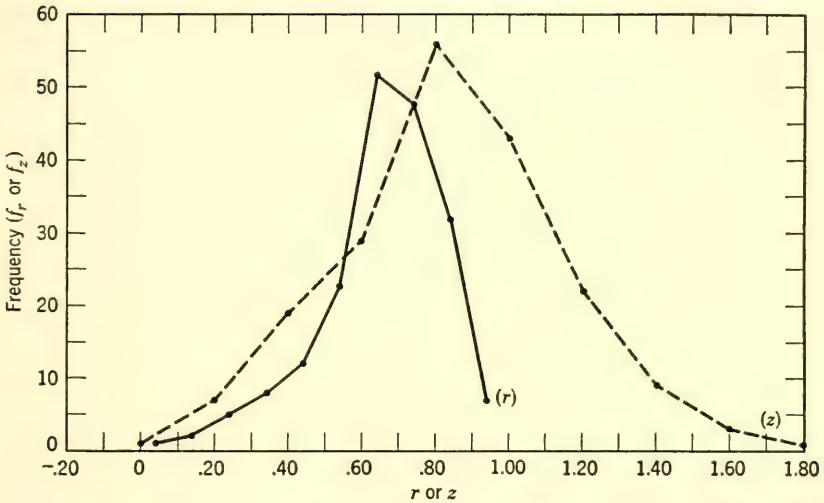


Figure 7.41. Sampling frequency distribution of the correlation coefficient, r, and of the corresponding $z = (1/2) \log_e [(1 + r)/(1 - r)]$. $n = 12$.

It was found by R. A. Fisher that under these circumstances it is helpful to use the following function of r:

$$(7.44) \quad z = (1/2) \log_e \left(\frac{1 + r}{1 - r}\right) = (2.30259/2) \log_{10} \left[\frac{1 + r}{1 - r}\right]$$

$$= 1.1513 \log_{10} \left[\frac{1 + r}{1 - r}\right].$$

because its sampling distribution is essentially normal in all important features even when ρ is definitely $\neq 0$. Moreover, its variance is given by $\sigma_z^2 = 1/(n - 3)$. This is not a sampling estimate but the true variance of z. It follows that, as a good approximation, the quantity $y = (z - z_\rho)/\sigma_z$, where z_ρ is the z corresponding to ρ in (7.44), is normally distributed. Hence, Table III gives the probabilities needed in tests of hypotheses regarding ρ or in the calculation of

confidence intervals on ρ. For example, consider the first sample drawn for Table 7.41. The $n = 12$ and $r = .668$; therefore, by formula 7.44,

$$z = 1.1513 \log_{10}\left(\frac{1.668}{0.332}\right) = 1.1513 \log_{10} 5.024 = 0.807, \text{ and}$$

$$\sigma_z = 1/\sqrt{9} = 0.333.$$

Then, since $y = (0.807 - z\rho)/0.333$ is a member of a standard normal population, the probability distribution of Table III can be used. If a confidence coefficient .95 is chosen, the inequality

$$-1.96 \leq \frac{0.807 - z_\rho}{0.333} \leq +1.96$$

requires that

$$0.154 \leq z_\rho \leq 1.460$$

unless a 1 in 20 chance has occurred in this sample. The corresponding 95 per cent confidence interval on ρ is obtained by using formula 7.44 and solving for the ρ, which now replaces r. Thus,

$$z_{\rho_1} = \text{lower limit} = 0.154 = \frac{1}{2}\log_e\left(\frac{1+\rho_1}{1-\rho_1}\right); \text{ or}$$

$$\frac{1+\rho_1}{1-\rho_1} = e^{0.308}. \quad \text{But, } \log_{10} e^{0.308} = 0.308 \log_{10} e$$

$$= 0.308(0.4343) = 0.134; \text{ and}$$

$$\text{Anti-log } 0.134 = 1.36 = \frac{1+\rho_1}{1-\rho_1}.$$

Hence $2.36\rho_1 = 0.36$ so that $\rho_1 = .155$. Similarly $\rho_2 =$ upper limit of the 95 per cent confidence interval $= .898$; therefore, the 95 per cent confidence interval on ρ is

$$.155 \leq \rho \leq .898,$$

which is a very wide interval but does include the true ρ, known in this case to be .749. If a relatively narrow confidence interval is needed, it is apparent that a rather large sample must be taken.

Figure 7.41 shows that the sample correlation coefficient varies over a considerable range even when ρ is as large as .749. As a matter of fact one sample r out of 190 was negative in spite of the relatively high positive correlation. This figure shows also, to a useful degree, the normalizing effect of the transformation $z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$. Given

a large number of sample correlations, the z-curve would become approximately normal in shape, as can be imagined from Figure 7.41.

Problem 7.41. Finney and Barmore (*Cereal Chemistry,* Vol. 25 [1948], page 299) have reported that the linear correlation between the per cent of protein in Nebred wheat flour and the loaf volume of bread baked therefrom was $r = .94$ on a sample of 30 pairs of measurements. What useful information does this provide?

The mere fact that $r^2 = (.94)^2 = .8836$ tells us that 88.36 per cent of the original sum of squares of the loaf volumes (Y) about their mean, $\bar{y}$, can be associated with the linear increase of that measurement with increasing protein concentration in the flour (X). Loaf volume is an important factor when the quality of bread is judged, and it is important to know what affects it.

It is inconceivable that such a large correlation coefficient would be obtained accidentally on thirty random observations; but, to illustrate the method, the hypothesis $H_0(\rho = 0)$ will be tested. It is seen that

$$t = \frac{.94}{\sqrt{\frac{1 - .8836}{28}}} = \frac{.94}{0.065} = 14.5,\ 28\ D/F.$$

Such a large t would occur by chance almost never; hence the hypothesis $H_0(\rho = 0)$ is decisively rejected. We know without even seeing the scatter diagram that the sample points lie closely about a linear regression line which has an upward trend. It also is apparent that the loaf volume from Nebred flour meeting the conditions of this experiment could be predicted quite accurately from a knowledge of its protein concentration.

There are some circumstances under which it is desirable to determine if two random samples probably were drawn from the same bivariate population as regards one, or both, of β and ρ. For example, it might be of interest to learn if one method of raising turkeys produces a more consistent relationship between the 16-week and the 28-week weights so that we could cull at 16 weeks of age with more confidence. Such an improvement in the relationship between these variables would indicate that the true coefficient of linear correlation, ρ, had been increased by the new methods. It also might be that superior poultry husbandry could increase the amount by which a weight advantage at 16 weeks of age would be followed by a weight advantage at 28 weeks of age. In the population considered earlier in this chapter, a one-pound advantage in weight at 16 weeks of age

was reduced, on the average, to only a one-half-pound advantage at 28 weeks of age. Thus β was 1/2. It might be that the size of β could be increased by superior breeding and handling.

If two samples—from two methods of breeding and raising turkeys, for example—have resulted in the computation of b_1, b_2, r_1, and r_2, the testing of the two hypotheses $H_0(\beta_1 = \beta_2)$ and $H_0(\rho_1 = \rho_2)$ can be carried out as follows:

For $H_0(\beta_1 = \beta_2)$:

(*a*) Pool the $\Sigma(Y_i - \hat{Y}_i)^2$ and the $\Sigma(Y_j - \hat{Y}_j)^2$ from the two samples, pool the degrees of freedom, and calculate

$$\text{pooled } s_{y \cdot x} = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(Y_j - \hat{Y}_j)^2}{n_1 + n_2 - 4}}.$$

(*b*) Compute $\sqrt{\dfrac{1}{\Sigma(x_i^2)} + \dfrac{1}{\Sigma(x_j^2)}}$.

(*c*) Multiply the standard deviation from *a* by the result obtained in *b*. This is the estimated standard deviation of $(b_1 - b_2)$, which will be called $s_{b_1 - b_2}$.

(*d*) Compute $t = (b_1 - b_2)/s_{b_1 - b_2}$, assign it $n_1 + n_2 - 4$ degrees of freedom, and interpret as before with respect to the acceptance or the rejection of $H_0(\beta_1 = \beta_2)$.

For $H_0(\rho_1 = \rho_2)$:

(*a*) Transform the r_1 and the r_2 to z_1 and z_2, respectively, in the manner described earlier in this chapter.

(*b*) Compute $\sigma_{z_1 - z_2} = \sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}$.

(*c*) Calculate $y = |z_1 - z_2|/\sigma_{z_1 - z_2}$ and consider this ratio as a normally distributed quantity in deciding whether or not it is so large that the hypothesis $H_0(\rho_1 = \rho_2)$ should be rejected.

If it seems appropriate after a hypothesis has been rejected, confidence intervals can be determined for the difference $\beta_1 - \beta_2$, but not for $\rho_1 - \rho_2$.

It is useful at times to have a convenient tabular procedure for computing b and r when the data are sufficiently numerous to justify the use of frequency distribution tables. Such data rarely would

TABLE 7.42

A Two-Way Frequency Distribution Table for Pairs of Measurements of the Weights of Turkeys

	X (16-week weight)															
28-week weight	4.0 to 4.5	4.6 to 5.1	5.2 to 5.7	5.8 to 6.3	6.4 to 6.9	7.0 to 7.5	7.6 to 8.1	8.2 to 8.7	8.8 to 9.3	9.4 to 9.9	10.0 to 10.5	f_Y	d_Y	$f_Y \cdot d_Y$	$f_Y \cdot d_Y^2$	$f_{XY} \cdot d_X \cdot d_Y$
17.5–18.0										1		1	4	4	16	12
16.9–17.4							2	4	0	3	1	10	3	30	90	51
16.3–16.8					1	2	7	4	4	4	2	24	2	48	96	56
15.7–16.2						4	7	6	9	4		30	1	30	30	32
15.1–15.6				1	6	11	11	21	19	4		73	0	0	0	0
14.5–15.0			3	4	11	13	8	14	3	2		58	−1	−58	58	33
13.9–14.4			2	4	11	11	12	6	6			52	−2	−104	208	70
13.3–13.8		1	2	2	3	5	5	2	2			22	−3	−66	198	72
12.7–13.2	1	0	1	0	1	2	1					6	−4	−24	96	76
12.1–12.6				1	2	1						4	−5	−20	100	40
f_X	1	1	8	12	35	49	53	57	43	18	3	280		−160	892	442
d_X	−6	−5	−4	−3	−2	−1	0	1	2	3	4					
$f_X \cdot d_X$	−6	−5	−32	−36	−70	−49	0	57	86	54	12	+11				
$f_X \cdot d_X^2$	36	25	128	108	140	49	0	57	172	162	48	925				

come from sampling studies, but perhaps they occur in practice often enough to justify the inclusion here of a method for obtaining the r and the b.

As in Chapter 2, the computations will be carried out in units of the class intervals. A two-way frequency distribution table is needed because two variables are involved. These matters, and others, are illustrated and discussed by means of 280 pairs of observations of 16- and 28-week weights of female turkeys similar to those studied earlier in this chapter. The symbol, X, is used to denote the 16-week weights and Y will stand for the 28-week weights, as before. Now that two variables are being considered simultaneously, frequencies in the X-classes will be symbolized by f_X, those for the Y-classes by f_Y. When it is desirable to indicate both the X and the Y for a class of data, f_{YX} will denote the frequency in that "cell" in the two-way table. Also, there may be two different lengths of class interval, I_X and I_Y for X and Y, respectively. With these symbols in mind, the following formulas are seen to be analogous to those used previously for b and r:

$$b = \frac{\Sigma(f_{XY} \cdot d_X \cdot d_Y) - [(\Sigma f_X \cdot d_X)(\Sigma f_Y \cdot d_Y)]/\Sigma f}{\Sigma(f_X \cdot d_X^2) - (\Sigma f_X \cdot d_X)^2/\Sigma f_X}, \text{ and}$$

$$r = \frac{\text{same numerator as that above for } b}{\sqrt{(\text{same as denominator above})(\text{same with } Y \text{ replacing } X)}}.$$

The data of Table 7.42 are arranged in a two-way frequency distribution table to provide a relatively easy basis for calculating b and r from their formulas as given above.

The following computations are derived from the summaries in Table 7.42:

$$\Sigma(f_Y \cdot d_Y^2) - (\Sigma f_Y \cdot d_Y)^2/\Sigma f_Y = 800.5714, \text{ and its square root} = 28.65.$$

$$\Sigma(f_X \cdot d_X^2) - (\Sigma f_X \cdot d_X)^2/\Sigma f_X = 924.5679, \text{ and its square root} = 30.41.$$

$$\Sigma(f_{XY} \cdot d_X \cdot d_Y) - [(\Sigma f_X \cdot d_X)(\Sigma f_Y \cdot d_Y)]/\Sigma f = 448.2857.$$

PROBLEMS

1. Calculate the r for the data of problem 1, section 7.1.

2. Calculate as in problem 1 for the data of problem 2, section 7.1. Given $\Sigma X^2 = 18{,}484$, $\Sigma XY = 727.99$, $\Sigma Y^2 = 28.6918$. *Ans.* $r = +.96$.

3. Compute $\Sigma(Y - \hat{Y})^2$ for the data of problem 3, section 7.1 by using the formula: $\Sigma(Y - \hat{Y})^2 = (1 - r^2) \cdot \Sigma(y^2)$.

4. In the formula of the previous problem, take $\Sigma(y^2) = 100$ and plot the left member of this equation on the vertical scale against r on the horizontal scale. Take r from -1 to $+1$ by increments of 0.2.

5. Reynolds, Bond, and Kirkland (*USDA Tech. Bull.* 861) give the following information on the relation between the cost of hauling logs and the length of the haul in miles over high-grade dirt or gravel roads:

Miles hauled:	1	2	3	4	5	6	7	8	9
Cost/1000 cu ft ($):	0.35	0.44	0.53	0.62	0.71	0.81	0.90	0.99	1.08
Miles hauled:	10	11	12	13	14	15	16	17	18
Cost/1000 cu ft ($):	1.17	1.26	1.36	1.45	1.54	1.63	1.72	1.82	1.90
Miles hauled:	19	20							
Cost/1000 cu ft ($):	2.00	2.09							

Compute a coefficient of linear correlation between length of haul and cost per 1000 cubic feet of volume, and draw conclusions. Is this really a proper use of correlation analysis? Would a regression analysis be better?

6. The persons mentioned in problem 5 gave the following data on the cost of producing 1000 cubic feet of hardwood logs in relation to the breast-high diameter of the logs:

Diam. (in.):	10	11	12	13	14	15	16	17
Cost ($):	12.70	12.63	12.38	12.03	11.62	11.32	11.10	10.84
Diam. (in.):	18	19	20	21	22	23	24	25
Cost ($):	10.63	10.49	10.40	10.28	10.13	10.04	9.96	9.88

Make a scatter diagram of these data, compute r, and discuss it in terms of the scatter diagram. Given: $\Sigma x^2 = 340$; $\Sigma XY = 3019.59$; $\Sigma(y^2) = 14.3129$.

Ans. $r = -.97$.

7. Compute $s_{y \cdot x}$ and s_Y for the data of problem 5, with Y = cost per 1000 cubic feet.

8. Calculate as in problem 7 for the data of problem 6.

Ans. $s_{y \cdot x} = 0.23$; $s_Y = 0.98$.

9. The *Yearbook of Labour Statistics* for 1943–1944 gives the following average daily wages of Chilean copper workers, in pesos:

Year:	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939
Wage:	11.89	11.26	11.75	11.33	12.80	13.31	14.77	16.37	21.31	23.20	25.34

$\Sigma(xy) = 155.74$; $\Sigma Y = 173.33$; $\Sigma(Y)^2 = 2996.1887$.

Construct a scatter diagram of these data, calculate s_Y and $s_{y \cdot x}$, and discuss their sizes relative to the graph.

10. Compute r for the data of problem 9—ignoring the fact that the year is not a random variable—and relate the size of the r to the appearance of the scatter diagram. Let $X = 1$ for 1929, 2 for 1930, etc. *Ans.* $r = +.91$.

11. Estimate the average wage for the year 1940 from the data of problem 9, using r in the computation of the standard deviation of this estimate.

12. Solve problem 1, and then test $H_0(\rho = 0)$ and draw appropriate conclusions. *Ans.* $r = .999$, $y \cong 10$, reject H_0 decisively.

13. For the data of problem 6 compute the CI_{95} on β, and then interpret this interval in a practical way.

14. For the data of problem 6 compute the CI_{90} on ρ and interpret this interval. *Ans.* $-.96 \leq \rho \leq -.80$.

15. Suppose that two random samples of 15 observations each have resulted in the computation of $r_1 = .75$ and $r_2 = .65$. Test $H_0(\rho_1 = \rho_2)$ and draw appropriate conclusions. Also compute the CI_{95} for each parameter, ρ_1 and ρ_2, and interpret these intervals. Can these interpretations be related to the test of H_0?

16. Draw a random sample of 30 observations from Table 7.21, compute the CI_{90} on ρ, and discuss the meaning of this interval.

17. Draw a random sample of 30 from Table 7.21 and test the hypothesis: $H_0(\rho = 0)$. How frequently would this procedure result in the rejection of H_0 when $\rho \neq 0$ (as in this case) at the 5 per cent level of rejection?

18. Draw two random samples of size 30 from Table 7.21 and test $H_0(\rho_1 = \rho_2)$, assuming that the first sample is from a bivariate normal population with $\rho = \rho_1$, and similarly for the second sample and $\rho = \rho_2$.

7.5 RANK CORRELATION

Sometimes it is either necessary or convenient to correlate the ranks of X's with those of their corresponding Y's. It may be that the X's and the Y's are only ranks in the first place, or it may be merely convenient to use ranks instead of four- or five-digit decimals, for example.

The practice of correlating ranks is both older and broader in its applications than is sometimes realized. Karl Pearson apparently was of the opinion that the idea of correlating ranks originated with Francis Galton during his studies of inheritance. Sometimes C. Spearman is credited with doing much to develop rank-correlation methods, especially as applied in psychological studies. It is his coefficient, r_s, which will be discussed specifically below. The works of M. G. Kendall, and others, recently have increased the use of ranks in statistics to a considerable degree, but no attempt will be made herein to give an exhaustive treatment of this subject. The interested reader is referred to Kendall's book, *Rank Correlation Methods,* published by Charles Griffin and Company, London.

The calculation of the Spearman, or rank-difference, coefficient of linear correlation (r_s) will be illustrated by means of the following pairs of ranks of students in two mathematics courses. Each pair gives the respective ranks of that student in statistics (X) and in mathematics of finance (Y). For example, the first student listed

ranked second in his class in statistics on the final examination, but ranked fifth in the final examination in mathematics of finance.

	Student															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X (statistics):	2	7	4	1	10	8	15	9	16	5	6	12	11	13	3	14
Y(finance):	5	4	3	2	9	7	16	6	15	12	8	11	10	13	1	14
Change in rank (d):	3	3	1	1	1	1	1	3	1	7	2	1	1	0	2	0

It is seen from these data that there is a general but imperfect tendency for a student's grades to rank about the same in both subjects, that is, a student's grade in statistics has some relation to his grade in mathematics of finance. If the relationship is basically linear, it can be measured rather simply and satisfactorily by means of the following formula for what is called the Spearman, or rank-difference, coefficient of correlation:

$$r_s = 1 - \frac{6\Sigma(d^2)}{n(n^2 - 1)} \tag{7.51}$$

where d is the difference between successive pairs of ranks (in the above illustration) or, in general, between the ranks of X_i and Y_i, i varying from 1 to n. For the data on ranks in statistics and in mathematics of finance,

$$d_1 = 2 - 5 = -3, \quad d_2 = 7 - 4 = +3, \cdots, \quad d_{16} = 14 - 14 = 0;$$

hence, $\Sigma(d^2) = 92$, and $r_s = 1 - 553/15(224) = 0.865$.

If there are ties for ranks, each X (or Y) so tied is given the mean of the ranks involved in the tie. For example, if two X's are tied for ranks 1 and 2, each X is given a rank of 1.5; if three Y's are tied among themselves for ranks 1, 2, and 3, each is considered to have rank 2.

It can be shown that r_s never has a size outside the range -1 to $+1$, regardless of the types of measurements involved or their sizes. It is seen from formula 7.51 that, if each Y has exactly the same rank as its corresponding X, all of the d's are 0 and hence $\Sigma(d^2) = 0$ and $r_s = 1$. If the ranks are perfectly reversed (1 with 16, 2 with 15, etc.), $r_s = -1$.

Kendall discusses such matters as confidence intervals for rank-correlation coefficients in his book (reference above) as well as introducing the coefficient tau (τ), which he prefers to the Spearman coefficient, r_s. These matters will not be discussed further here, but

the reader again is invited to consult Kendall's book on this subject if interested.

PROBLEMS

1. Solve problem 5 of section 7.4, using the Spearman coefficient, r_s.

2. Solve problem 6 of section 7.4, using the Spearman coefficient, r_s.

Ans. $r_s = -1$.

3. Compute r_s for the data of problem 9, section 7.4, letting $X = 1$ for 1929, $X = 2$ for 1930, etc., and setting $Y =$ wage.

4. Compute the rank-difference coefficient of linear correlation for the pairs of observations in Table 7.22. *Ans.* $r_s = .68$.

5. Make up a problem for which $r_s = +1$, also for $r_s = -1$, and $r_s = +.5$. Then make up another set for each case different from each of the others.

6. A sampling study in cereal chemistry gave the following product-moment linear correlations:

$$\text{Sample 1:} \quad n_1 = 44, \; r_1 = -0.93$$

$$\text{Sample 2:} \quad n_2 = 44, \; r_2 = -0.81.$$

Test $H_0(\rho_1 = \rho_2)$ and draw appropriate conclusions.

Ans. $y = 2.40$; $P \cong .017$; reject H_0.

7. Referring to problem 6, how small could the sample size become and still result in the rejection of H_0 at the 5 per cent point if the r's stayed the same size?

8. If $r_1 = -.93$, as in problem 7, could $r_2 = -.90$ ever result in the rejection of the H_0 of problem 6 at the 5 per cent level for *any* sized sample? If so, what size would n_1 and n_2 have to be if they were equal? *Ans.* $n_1 = n_2 = 225$.

9. It has been stated that each of the ratios $(b - \beta)/s_b$ and $\dfrac{r}{\sqrt{(1 - r^2)/(n - 2)}}$ follows the t-distribution with $(n - 2)$ degrees of freedom under random sampling with a given n. Show that these two quantities are algebraically identical if $\beta = 0$; and hence that testing H_0 $(\beta = 0)$ is identical to testing H_0 $(\rho = 0)$.

10. Suppose that the following results were obtained from two samples (involving different methods of some sort), each containing 20 observations:

$$\text{Method 1:} \quad r_1 = .40, \; H_0 \; (\rho_1 = 0) \text{ accepted.}$$

$$\text{Method 2:} \quad r_2 = .60, \; H_0 \; (\rho_2 = 0) \text{ rejected, } P < .01.$$

$$\text{Yet } H_0 \; (\rho_1 = \rho_2) \text{ is accepted readily because } P > .40.$$

Explain how such results are not contradictory. Also, determine what sizes n must have in order that each of these three hypotheses will be rejected at the 5 per cent point if the correlations stay as they are.

Ans. First H_0, $n_1 = 25$; second H_0, $n_2 = 11$; third H_0, $n_1 = n_2 = 109$.

You are given the following two-variable frequency distribution table as the basis for solving problems 11 to 15 below. These data are derived from records of heights and weights of 9-year-old Kansas girls in certain schools. These data

were obtained from the Department of Home Economics, Kansas Agricultural Experiment Station, through the courtesy of Dr. Abby Marlatt.

Weight in kilograms (Y)	Height in Centimeters (X) 123 to 126	127 to 130	131 to 134	135 to 138	139 to 142	143 to 146	147 to 150	151 to 154
650–689								1
610–649								1
570–609							3	0
530–569							0	4
490–529					1	4	2	3
450–489					1	4	9	4
410–449			1		3	7	12	6
370–409			1	5	6	14	13	5
330–369		1	2	9	23	17	14	4
290–329	1	1	17	25	32	13	4	
250–289	1	15	32	19	5			
210–249	5	8	5	2				
Mean weight	246.6	259.9	285.4	306.8	336.5	378.7	409.9	458.1
Standard deviation	29.1	27.2	35.6	38.6		57.9	109.6	82.8

11. Plot the mean weights above against the midpoints of the height classes, and decide therefrom if the assumption of a linear relationship between these two variables seems acceptable.

12. Ignore any indication of non-linearity of trend and compute r and b by the methods of this chapter. What conclusions can you draw from these estimates?

Ans. $b = 0.77$, $r = .72$.

13. Compute the standard deviation not given in the above table by the method of Chapter 2 adjusted so as to take account of the fact that this is supposed to be a sample.

14. Each height class has some kind of one-variable frequency distribution of the weights within the height class. Hence the above data constitute several samples of weights within height classes. Theoretically, these weight distributions within height classes must have equal population variances. Plot the standard deviations against the midpoints of the corresponding height classes and decide therefrom—if you can—whether or not that is a good assumption in this case.

15. For the weight class 290 to 329 kilograms compute the coefficient of variation for the heights, taking the point of view maintained in Chapter 2.

REVIEW PROBLEMS

1. Define the term *percentile* and explain how it can be associated with the relative cumulative frequency distribution of a group of measurements.

2. Calculate the arithmetic mean, standard deviation, and coefficient of variation for the following data on the carotene content of pasture grasses, in milligrams per gram:

X: 0.22 0.13 0.23 0.36 0.44 0.26 0.11 0.23 0.26
0.26 0.20 0.16 and 0.20

3. What is the median carotene content for the data of problem 2? The mean deviation?

4. Compute the geometric mean of 32 and 90. Of 2, 7, and 30. Of the X's of problem 2. *Ans.* 53.8; 7.49; 0.22.

5. You are given the following information regarding a group of 2000 weights (in grams): $\mu = 800$, $md = 700$, $Q_1 = 500$, $Q_3 = 900$, extreme weights are 350 and 1300; and the upper limits of the 15th, 35th, 80th, and 90th percentiles are 400, 540, 950, and 1050, respectively. Sketch a graph of the *r.c.f.* curve.

6. Given the following scores made on an Ohio Psychological Test, construct a frequency distribution with equal class intervals and compute μ and σ. These scores are necessarily integers.

83	69	30	26	53	60	44	36	68	71	55	52	45	62	42	47
70	62	28	46	42	45	38	45	75	79	73	105	80	81	68	65
48	52	38	77	26	71	31	24	51	55	67	41	36	67	106	37
60	48	74	98	62	33	83	108	74	35	38	35	38	112	66	85
48	44	100	55	77	78	21	94	35	75	71	69	61	50	70	47
65	103	100	70	60	30	97	86	54	71	87	68	64	54	45	30
52	49	78	51	91	63	45	46	90	42	68	34	79	76	39	38
64	46	34	43	57	76	31	60	34	105	17	31	67	73	53	99
68	54	37	99	43	24	50	58	104	64	54	38	96	53	57	35
52	73	66	39	59	70	91	88	60	44	82	72	56	76	71	30
59	50	100	77	129	46	86	88	36	78	61	58	40	37	65	72
103	63	46	70	48	48	57	83	51	29	51	32	37	100	43	47
53	41	107	115	64	59	26	48	40	61	37	70	49	62	88	42
69	49	71	57	87	63	101	69	50	75	69	48	59	49	96	67
63	71	75	56	78	40	81	59	74	110	57	28	50	68	63	55
61	30	95	116	75	71	31	34	77	60	84	68	70	36	65	27
63	49	41	79	66	73	53	99	98	79	89	27	87	37	48	75
80	109	43	46	91	77	61	44	58	53	45	87	96	64	84	87
116	35	105	43	75	22	37	49	56	60	74	38	38	28	57	29
57	34	61	27	62	71	53	44	88	76	61	45	45	41	33	57
58	83	82	67	75	29	71	77	50	47	102	83	47	64	57	75
94	38	38	107	65	25	51	28	53	80	79	55	47	57	76	49
92	32	39	89	70	52	34	41	31	77	57	44	56	41	39	42
81	70	68	69	80	48	46	38	83	65	33	57	14	42	32	78
51	55	50	52	75	57	65	74	40	63	44	59	38	60	64	35
50	65	37	76	82	100	48	69	47	54	33	35	61	74	37	37
35	42	128	35	47	57	59	46	91	80	81	78	74	53	39	66

58	63	40	55	46	46	40	38	58	63	32	42	56	30	85	50
41	74	43	55	93	33	60	72	54	81	66	56	36	60	92	39
31	81	41	38	28	62	51	86	38	61	48	85	53	82	26	32
48	46	40	51	54	28	66	72	48	75	69	69	82	56	30	57
96	87	63	43	45	38	82	43	62	31	66	80	97	78	36	60
91	97	59	40	45	78	89	28	67	79	53	82	37	98	56	68
66	33	36	43	80	72	51	54	30	34	36	77	54	63	66	45
29	29	59	70	83	45	108	78	37	48	36	33	97	43	58	89
60	67	55	64	72	99	91	75	46	52	59	39	18	54	91	76
29	63	95	41	28	45	44	94	57	34	86	36	36	69	55	58
67	86	82	42	48	62	109	48	81							

Ans. $\mu \cong 59.7$; $\sigma \cong 21.9$.

7. Construct an *r.c.f.* curve for the data of problem 6 and obtain from it evidence regarding the normality of the distribution of these test scores.

8. In what percentile would a person who made a score of 101 rank in the test of problem 6? *Ans.* 97th.

9. What are the modal and the median test scores, respectively, for problem 6?

10. Calculate the median for the Ohio test scores of problem 6 by grouping them into about 12 classes of equal length. *Ans.* $md \cong 51.6$.

11. *The Year Book of Labour Statistics* for 1943–1944 gives the following percentages of unemployed in the United States and Sweden during 1941, by monthly averages:

U.S.:	15.3	14.0	12.1	8.8	5.3	2.9	2.2	1.9
Sweden:	17.1	16.4	15.1	13.1	10.6	9.3	7.8	7.5

U.S.:	0.5	1.3	3.7	5.3
Sweden:	7.3	8.2	10.0	13.0

In which country was the level of unemployment relatively more stable during that year? Justify answer statistically.

12. The USDA publication, *Agricultural Statistics,* 1946, lists the following tax levies for the 48 states, in dollars per acre:

1.01	0.89	0.59	2.73	1.77	2.19	1.06	2.26	1.05	0.73
0.76	1.06	0.54	0.96	0.83	1.18	0.33	0.24	0.24	0.37
0.40	0.33	0.81	0.27	0.16	0.40	0.24	0.18	0.62	0.37
0.43	0.22	0.45	0.32	0.33	0.25	0.15	0.11	0.52	0.06
0.20	0.04	0.08	0.35	0.17	0.43	0.27	1.00		

Compute the median tax per acre from an array of these data. Also compute the range and the midrange. Compare the latter with the median and draw any possible conclusions. *Ans.* $md = 40$; range $= 2.69$; MR $= 1.38$.

13. Referring to the data of problem 12, in what decile would a state rank if its tax rate were 0.40 dollars per acre? What percentage of the states would have a higher rate?

14. Suppose that a sample of 15 differences in yield between two varieties of corn grown side by side on 15 pairs of plots has been found to have an arithmetic

mean = 10 bushels per acre, with the standard deviation = s = 13 bushels per acre. Is one variety superior to the other or not? Give reasons for answer.

Ans. $t \cong 3.00$; $14D/F$, $P \cong .01$.

15. Suppose that a large number of tractor gears has been produced and that 90 per cent of them are classifiable as acceptable. If a sample of 10 gears is taken at random from this group, which of the following is more likely to occur? (*a*) The sample will contain less than 90 per cent "acceptable" and will, therefore, give a pessimistic picture of the quality of the whole batch. Or (*b*) the sample will contain at least 90 per cent "acceptable" and hence will, if anything, overestimate the quality of the batch.

16. The depth of deterioration (in inches) is used as an index of the merchantable volume of timber remaining in fire-killed Douglas fir. Kimmey and Furniss (*USDA Tech. Bull.* 851) report a study made in western Oregon and Washington on such timber. The following data on old-growth trees were read from one of their graphs:

Years after fire:	5	10	15	20	25	30	35	40	45
Depth in inches:	2.2	3.6	4.8	6.3	7.7	9.1	10.8	12.2	13.8

Years after fire:	50	55	60
Depth in inches:	15.5	17.5	19.8

What do you conclude from a regression analysis is the average increase in depth of deterioration per decade? Given $\Sigma XY = 5123.0$, $\Sigma Y^2 = 1617.09$.

Ans. 3.0 inches = point estimate; CI_{95}: $0.25 \leq \beta \leq 0.35$.

17. Economists sometimes speak of commodities with elastic or inelastic prices, meaning generally that a commodity which is slow to change price in the face of changes in demand has an elastic price. If you adjust prices for inflation and for depression, and if demand is measured by per capita consumption of a given commodity, the definition of an elastic, or inelastic, price can be made more specific. For example, if the slope of the linear trend line relating adjusted price (Y) to consumption per capital (X) is less than unity, the price can be called elastic. If β is greater than 1, the price then is called inelastic. Given the following data regarding whole milk and cream, would the price be classified as elastic according to the above definition after due allowance for sampling error?

Adjusted price ($/cwt)	1.88	2.06	2.07	2.26	2.36	2.29	2.48	2.19
Consumption per capita (cwt)	3.43	3.50	3.72	3.93	4.12	4.32	4.20	4.00

18. In the preceding problem a definition of an elastic price was based upon the size of the regression coefficient, β. What information would it add to this discussion to include the size of the correlation coefficient?

19. Following are adjusted farm beef prices (per hundredweight) and consumption per capita (hundredweight) for the 10-year period indicated:

Year:	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946
Price:	6.67	6.67	7.52	7.79	8.32	8.63	8.62	7.94	8.71	9.06
Consumption:	0.55	0.54	0.54	0.55	0.60	0.61	0.53	0.55	0.59	0.61

Compute b and r, describe the price as elastic or inelastic, and bring the size of r into your discussion as suggested in problem 18.

20. Given the following data on turkeys, solve as in problem 19:

Year:	1930	1931	1932	1933	1934	1935	1936	1937
Price (cents/lb) Y:	15.9	18.4	14.8	13.8	16.1	20.1	15.4	17.2
Consumption (lb) X:	1.8	1.7	2.1	2.4	2.2	2.1	2.7	2.7

Year:	1938	1939	1940	1941	1942	1943	1944	1945
Price (cents/lb) Y:	17.9	16.5	15.9	18.8	22.2	23.6	25.0	24.2
Consumption (lb) X:	2.7	3.0	3.5	3.5	3.7	3.3	3.3	4.3

Year:	1946	1947
Price (cents/lb) Y:	22.6	18.7
Consumption (lb) X:	4.5	4.5

Given: $\Sigma X^2 = 175.38$, $\Sigma XY = 1041.46$, $\Sigma(y^2) = 203.7361$.

Ans. CI_{95}: $1.0 \leq \beta \leq 3.4$, $r = .58$.

21. Referring to problems 19 and 20, were the beef or the turkey prices relatively more stable during the period 1937 to 1946? Give statistical evidence for your answer.

22. The earliness with which chickens obtain their feathers is economically important to persons who raise broilers because it affects the rapidity and cleanness of dressing. Early feathering, a sex-linked characteristic, is chiefly dependent upon one gene locus on the sex chromosome. Its inheritance can be described diagrammatically as follows:

1. $\frac{L}{\text{none}}$ = late-feathering female,

2. $\frac{l}{\text{none}}$ = early-feathering female,

3. $\frac{L}{L}$ or $\frac{L}{\quad}$ = late-feathering male,

4.

5. $\frac{l}{l}$ = early-feathering male.

If late-feathering females are mated to early-feathering males, what is the expected number of: (*a*) late-feathering females among the offspring, (*b*) late-feathering males, (*c*) early-feathering chicks of either sex among 1000 offspring?

Ans. (*a*) none; (*b*) 500; (*c*) 500, all females.

23. Suppose that the late-feathering males in a flock can be assumed to have two-thirds of type *Ll* and one-third of type *LL*. If these males are mated to

early-feathering females, what is the probability that a fertile egg selected at random will hatch into an early-feathering chick, sex disregarded?

24. Referring to problem 22, suppose parents 2 and 4 are mated. If 4 fertile eggs are to be incubated and all can be assumed to produce a live chick, what is the probability that at least one early-feathering chick of each sex will be hatched so that you could hope to develop a line of early-feathering chickens? *Ans.* 55/128.

25. Some educators believe that tests can be developed which measure a persons general ability to think critically. No particular field of subject matter is involved. The following data are test scores derived from such a testing program. The students were asked—but not required—to indicate the level of their father's annual income for the preceding year. All samples have $n = 20$ and are assumed to be from normal populations.

	Freshmen						Juniors and Seniors			
	Men			Women			Men		Women	
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(3)
	41	44	34	43	41	24	39	36	38	31
	31	42	39	44	36	38	35	42	47	45
	36	31	39	42	20	32	34	32	26	33
	36	31	41	26	26	35	25	40	34	29
	26	34	36	33	25	35	27	23	45	28
	25	28	35	41	34	20	31	41	24	39
	34	29	36	41	38	26	44	31	39	25
	24	35	33	28	42	32	45	32	43	41
	28	29	21	35	33	20	48	35	35	44
	29	34	31	28	34	34	39	40	33	18
	28	38	33	28	28	31	37	36	31	35
	37	27	26	24	34	33	29	37	27	46
	35	20	28	30	20	22	36	32	28	38
	32	42	35	36	45	24	35	45	29	49
	33	28	30	38	37	35	41	32	32	30
	35	36	32	38	32	29	35	43	33	42
	30	34	30	31	35	28	26	38	37	40
	49	38	21	45	40	21	39	44	32	37
	38	42	25	40	24	37	37	35	40	36
	34	26	39	37	34	41	43	46	38	13
$\bar{x}$:	33.0	33.4	32.2	35.4	32.9	29.9	36.2	37.0	34.6	35.0

(1) = not over \$5000. (2) = \$5001 to \$10,000. (3) = none stated.

Use any or all of these data for problems 25 to 32 below.

Do the two samples for freshmen men and women whose fathers earned not more than \$5000 indicate that freshmen women think more critically than freshmen men, if this test is assumed to be reliable?

26. Referring to problem 25, how do freshmen men compare with junior and senior men whose fathers are in the lower income group? Show how the G-test helps answer this question. *Ans.* $G = 0.13$; $P > .10$; accept $H_0(\mu_F = \mu_{\text{J and S}})$.

27. Solve as in problem 26 for the higher income group of fathers.

28. You might wonder if the group which refused (or neglected) to reveal the fathers income are from a different population as regards scores in critical thinking. Make a study of this matter for freshmen women by means of the G-test.

Ans. Low income vs. undeclared: $G = 0.26$, $P < .010$.
High income vs. undeclared: $G = 0.09$, $P \gg .10$.

29. Do freshmen women from the higher income group belong to a population with a lower mean score than that of the freshmen women from the lower income group? Use the G-test.

30. Solve problem 29, using the t-test instead of the G-test. (See other tables for these D/F.) Given: Σx^2 for (1) = 804.80; for (2) = 953.80.

Ans. $t = 1.16$, 38 D/F, $P > .05$.

31. Use the G-distribution to set a 90 per cent confidence interval on the true mean score of freshmen men whose fathers make $5000 or less per year. Does this interval make it possible to test the hypothesis $H_0(\mu = 0)$?

32. Compute CI_{90}'s on the true means of the populations sampled by columns 4 and 6 (from left) in the table above, and draw all appropriate conclusions.

Ans. Col. 4: $33.2 \leq \mu_1 \leq 37.6$. Col. 6: $27.7 \leq \mu_2 \leq 32.1$ by G-distribution.

The following data record the thiamin-content, in micrograms per gram of meat (dry fat-free basis) in raw pork loin after various periods of storage (temperature not over 10°F). (These data made available through the courtesy of Dr. Beulah Westerman, Department of Foods and Nutrition, Kansas State College.)

	Period of Storage (Weeks)					
	0	12	24	40	56	72
	126.88	81.47	91.51	104.78	76.99	93.57
	98.83	69.14	69.58	69.04	79.22	94.25
	106.55	119.44	98.17	84.84	74.28	114.03
	91.73	75.65	81.49	105.20	121.34	99.65
	68.35	65.41	77.05	70.06	83.58	88.49
	95.41	111.89	102.43	111.17	97.14	77.19
	111.67	80.93	91.87	86.30	97.21	116.62
	78.30	76.94	88.62	71.01	72.91	87.38
	118.50	111.26	102.31	100.85	65.03	91.94
Mean	99.58	88.01	89.23	89.25	85.30	95.90
range	58.53	54.03	25.38	42.13	56.31	39.43

33. Compute through the G-distribution the CI_{90} on the true thiamin concentration (micrograms per gram) in raw, unstored, pork loin produced under the conditions maintained during the sampling which produced the above data. Draw all appropriate conclusions.

34. Solve as in problem 33 for raw pork loin stored 12 weeks.

Ans. CI_{90}: $76.9 \leq \mu \leq 99.1$ micrograms per gram.

35. Does cold storage (at, or below, 10°F) of raw pork loin for 12 weeks reduce the thiamin concentration, according to the evidence from the above data and the G-test?

36. Does the concentration of thiamin in raw pork loin increase between the fifty-sixth and the seventy-second week of cold storage (at, or below, 10°F), or is the observed average increase of 10.60 micrograms per gram probably just a sampling accident? *Ans.* $G = 0.221$, $n = 9$, $P > .10$; sampling accident.

37. It appears from the sampling data above that the thiamin concentration in raw pork loin decreases during the first 12 weeks of storage, stays about the same through the fifty-sixth week of such storage, and returns to about the original concentration by the end of the seventy-second week of storage. Is this actually the case, according to G-tests, or could the observed results reasonably be assigned to sampling error?

The following data are from the same source as those immediately above, and were taken during the same general experimentation. They record the riboflavin concentration in raw pork loin instead of the thiamin content just studied. These data are to be employed in the solution of problems 38 through 43.

	Period of Storage (Weeks)					
	0	12	24	40	56	72
	3.42	4.31	5.98	5.17	4.08	5.39
	2.86	3.52	4.84	4.19	3.22	5.02
	2.99	3.47	5.14	4.87	4.03	5.51
	2.24	3.47	4.72	4.55	4.19	5.03
	2.02	3.43	4.52	4.58	3.35	4.25
	2.17	4.07	2.91	4.28	3.35	4.18
	1.69	3.52	3.73	4.33	5.23	4.80
	2.09	3.48	3.61	4.29	4.91	3.51
	1.57	3.84	3.60	5.14	5.80	4.63
Mean	2.34	3.68	4.34	4.60	4.24	4.70
range	1.85	0.88	3.07	0.98	2.58	2.00

38. Make a scatter diagram with Y = mean riboflavin concentration and X = weeks of storage. Is the trend in the bivariate population of X's and Y's probably linear for these times of storage?

39. Would the above data cause you to accept, or to reject, the hypothesis that the riboflavin concentration in raw pork loin is increased by 12 weeks of storage at, or below, 10°F?

40. Can the apparent drop in riboflavin concentration between the fortieth and the fifty-sixth weeks of storage reasonably be assigned to sampling accidents? *Ans.* $G = 0.202$, $n = 9$, $P > .10$; accept $H_0(\mu_{40} = \mu_{56})$.

41. Use the t-distribution to set a 95 per cent confidence interval on the true riboflavin concentration in raw, unstored, pork loin of the kind sampled here. You are given that $\Sigma X = 21.05$, and $\Sigma X^2 = 52.3121$.

42. Solve problem 41 by means of the G-distribution, and compare the result with that obtained from the t-distribution.

Ans. From G: $1.9 \leq \mu_0 \leq 2.8$; from t: $1.8 \leq \mu_0 \leq 2.9$.

43. According to the evidence above, 40 weeks might be an optimum storage period for increasing riboflavin. The means for 0 and 40 weeks differ by 2.36. Use the G-distribution to place a CI_{95} on the true gain due to 40 weeks of storage, and draw conclusions.

44. A recessive lethal will destroy an organism only if carried by both chromosomes of a pair. Suppose that l_1 is such a lethal, and that the following mating has been made: $L_1l_1 \times L_1l_1$. What is the probability that among the first 10 offspring none will be killed by this lethal? *Ans.* .057.

45. Suppose that a flock of chickens carries the lethal mentioned in problem 44, and that the owner wishes to so select his future breeding stock that this lethal will disappear as rapidly as possible from his flock. He knows that some of his chickens are carriers, that is, are L_1l_1. New stock which he raises cannot be designated as L_1L_1 or as L_1l_1 until they have produced some (perhaps many) offspring. Hence new members of the flock will be mated to known L_1l_1's and then will be eliminated from the flock if *any* of their offspring are victims of the lethal because this will show that they are carrying that gene. How many offspring should the owner see from a chicken *without the appearance of the lethal* before accepting that chicken as being L_1L_1 and hence not a carrier of the lethal? Since he never can be absolutely positive, assume that he is willing to run a risk of 1 in 50 of reaching such a conclusion erroneously.

46. Suppose that a trait which is of economic interest to a sheep breeder is determined by two genes, R and S, believed to be carried on two different chromosomes. It also is believed that R is completely dominant to r and similarly for S with respect to s. It is supposed that only those animals showing both dominant characteristics are of special interest. If the breeder's hypotheses are correct, the mating $RrSs \times RrSs$ should produce 9/16 of its offspring with both the R and the S genes, 3/16 with R but not S, 3/16 with S but not R, and 1/16 with neither R nor S. Suppose that all four possibilities are distinguishable and that the following offspring have been recorded:

82 are R and S (called RS); 36 are R but not S (called Rs); 28 are S but not R (called rS), and 14 are neither R nor S (called rs).

Given these results, would you accept the hypothesis stated above, namely, $H_0(9RS{:}3Rs{:}3rS{:}1rs)$? Give reasons.

Ans. $\chi^2 = 3.644$, 3 D/F, $P \gg .11$; accept H_0.

47. What is the probability that *both* of two CI_{95}'s on μ obtained from two random samples from the same normal population will include μ? Since the μ would lie in the overlap of these two intervals (if both did include μ), and since this overlap would be shorter than either interval in many cases, and never longer, would you do a better job of estimating μ by using two random samples and considering this overlap? Would the probability of an error of the first kind be reduced if this process were used to test an H_0? Give reasons for answers.

The following numbers are measurements of basal metabolism (in calories/square meter of surface area/hour), and are to be used in answering problems 48 to 53 below. These data were derived from measurements provided through

the courtesy of Mrs. Ada Seymour and the Department of Home Economics, Kansas Agricultural Experiment Station. All ages are to the nearest birthday.

Age Class	n	CI_{95} on μ by t	mean, $\bar{x}$	$s_{\bar{x}}$
10–11	45	42.96–44.72	43.84	0.441
12–14	46	37.27–39.33		0.516
15–16	52		34.59	0.350
17	65	33.69–35.11		0.354
18	90	32.59–33.77		0.295
19	91	32.05–33.70		
20	73	32.67–34.00		0.333
21–25	175		32.82	0.185
26–29	55	31.66–33.01		0.338
30–34	73	31.90–32.92	32.41	
35–39	57	32.36–33.82	33.09	0.362
40–44	53	32.00–33.38		0.346
45–49	56	30.75–31.97		0.304
50–59	62	30.67–32.03		0.341
60 and over	33	30.06–32.12		

48. Fill in the two CI_{95}'s omitted above and state what information they yield.

49. Graph the CI_{95}'s versus age (on the horizontal axis) so as to produce a figure from which you could read, approximately, the confidence interval on true mean basal metabolism for any age, with a confidence coefficient .95. This is to be applied only to Kansans, of course.

50. Compute the two missing standard deviations in the above table.

51. Test the hypothesis that Kansas women between the ages of 35 and 39 have a higher average basal metabolism than those in the age interval from 30 to 34 years.

52. According to the "Mayo Foundation Normal Standards," published in July of 1936 in the *American Journal of Physiology,* the mean basal for 17-year-old females is 37.82 calories per square meter per hour. According to the table above do the Kansas girls fit that norm, or do they probably have a lower average metabolism rate? How confident can you be of your answer when allowance is made for sampling error in the above table, but none is allowed for the Mayo Standard?

53. Assuming that the records for those persons in the age group 21 to 25 years are normally distributed, estimate the range for this sample of 175.

54. Suppose that 147 freshmen, 18 years of age, have taken a test designed to measure their ability to think critically, and have taken this test at the beginning and also at the end of their freshmen year. Their progress during the year is measured by the difference between these two scores. Given that $\Sigma Y = 712$ and $\Sigma y^2 = 5567.40$, test the hypothesis that freshmen of the sort so sampled make *some* improvement in critical thinking during the year in so far as this is measured by the test administered. Consider that t has 30 D/F.

Ans. $t = 9.67$, P nearly zero; $\mu \neq 0$.

55. Suppose that two varieties of corn have been grown at the same experimental farm during the same year, and that the following plot yields, in pounds, have been obtained:

No. 1: 12.1 12.8 15.2 14.0 13.5 13.6 14.3 12.9 13.9 and 14.7
No. 2: 14.6 12.9 15.6 14.3 14.8 13.4 13.8 15.3 16.0 and 14.5

These field weights have been corrected for moisture content so that the variety yields per acre can be compared directly with these data. Use the G-test to test the hypothesis $H_0(\mu_1 = \mu_2)$, where the μ's are the true means of the varieties.

56. The following data simulate those which might be obtained from an experimental comparison of the effectiveness of two fertilizers on the yield of orange trees in pounds per tree:

Nitrogen (N):	74	89	90	72	78	76	84	79	81	76	and	80
N + Potash:	103	102	97	80	87	92	91	78	83	89	and	92

The two groups of trees (one for N and the other for N + P) were assumed with good reason to be on equivalent areas of land before the two fertilizers were applied. Test the hypothesis that the addition of potash does not affect yield. *Ans.* $G = 0.488$, $n = 11$, $P \cong .002$; reject hypothesis.

57. Referring to problem 56, use the G-test to place a 92 per cent confidence interval on the true difference in average yield produced by adding potash under these circumstances, and draw appropriate conclusions.

58. The following numbers are the pounds of tobacco per acre yielded, on the average, in the United States during the years indicated. Make a scatter diagram and decide if the trend toward increasing yield can be reasonably considered as linear if this is taken to be a sample.

Year:	1932	1933	1934	1935	1936	1937	1938	1939
Yield:	725	789	852	905	807	895	866	940
Year:	1940	1941	1942	1943	1944	1945	1946	1947
Yield:	1036	966	1023	964	1116	1094	1182	1142

59. Referring to problem 58, again assume that this is a sample from a bivariate population and compute, and interpret, the CI_{90} on β, the true slope of the regression line.

60. Solve as in problem 59, after substituting ρ, the true coefficient of linear correlation, for β. *Ans.* CI_{90}: $.86 \leq \rho \leq .98$.

61. The following data are the numbers of sugar-maple trees tapped each year and the resulting pounds of sugar and sirup. If these data can be regarded as a sample, did the production per tree change during this period in any orderly manner; and, if so, how?

Year:	1929	1930	1931	1932	1933	1934	1935
Trees (1000's):	12,951	13,158	12,092	12,064	12,009	12,099	12,341
Pounds (1000's):	3724	5856	3589	3748	3269	3488	4673

Year:	1936	1937	1938	1939	1940	1941
Trees:	11,500	11,339	11,380	10,313	9,957	9,785
Pounds:	3122	3276	3475	2881	3031	2384
Year:	1942	1943	1944	1945	1946	1947
Trees:	9,847	9,281	8,681	7,336	8,000	8,568
Pounds:	3569	3133	3133	1228	1700	2344

62. The cumulative and *r.c.f.* distributions given below are those of the sizes of peach orchards in the Sandhills of North Carolina during 1946. (Data from *Technical Bulletin* 91, North Carolina Agricultural Experiment Station.) Do the sizes of these orchards follow a normal frequency distribution quite well, or is their distribution far from normal?

Number of Trees in Orchard	Cumulative f	*r.c.f.*	Number of Trees in Orchard	Cumulative f	*r.c.f.*
200–299	12	.03	200–1999	121	.37
200–399	23	.06	200–2999	153	.47
200–599	47	.13	200–4999	188	.60
200–799	64	.17	200–9999	225	.79
200–999	79	.22	all orchards	257	1.00

63. Referring to problem 62, what is the median size of orchard? The lower limit of the second quartile?

The following scores on certain academic aptitude tests and the student's grade point average (GPA) at the end of the indicated year are to be the basis for answering the questions in problems 64 to 70, inclusive. These data constitute samples from classes taking a natural science comprehensive course at Kansas State College.

FRESHMEN

ACE-T	ACE-L	ACE-Q	GPA	ACE-T	ACE-L	ACE-Q	GPA
66	33	33	0.11	85	48	37	0.56
101	57	44	0.96	89	53	36	0.68
85	50	35	1.33	100	53	47	0.56
96	56	40	1.11	122	67	55	1.03
115	66	48	1.30	117	74	43	2.33
110	74	36	2.06	96	64	32	2.31
111	70	41	0.06	90	58	32	0.93
62	39	23	1.50	103	63	40	1.58
74	48	26	0.77	41	26	15	0.04
116	85	31	2.64	68	42	26	1.04
102	62	40	1.14	125	74	51	2.24
113	69	44	0.81	111	71	40	0.77
105	62	43	0.48	87	64	23	1.27
81	49	32	1.22	100	65	35	0.59
113	61	52	2.59	114	71	43	2.19
147	101	46	2.54	99	59	40	1.68

FRESHMEN (*Continued*)

ACE-T	ACE-L	ACE-Q	GPA	ACE-T	ACE-L	ACE-Q	GPA
93	58	35	0.97	115	60	55	1.39
59	31	28	0.50	77	46	31	0.50
75	52	23	0.83	89	53	36	1.27
106	63	43	1.24	137	81	56	1.38
37	27	10	0.41	42	30	12	0.96
139	72	67	1.97	125	67	38	1.41
126	80	46	1.83				

JUNIORS

115	65	50	1.42	132	71	61	1.00
100	55	45	0.85	109	58	51	1.95
107	55	52	1.12	129	83	46	1.93
108	72	36	1.64	87	50	37	1.79
115	64	51	0.81	80	45	35	1.35
83	46	37	0.86	110	70	40	2.08
121	71	50	1.11	96	47	49	1.60
82	54	28	1.18	122	70	52	1.93

64. Make scatter diagrams of the total ACE scores (ACE-T's) on the horizontal axis and the GPA's on the vertical for freshmen and also for juniors, using the same coordinate system but different symbols for the two classes.

65. After solving the preceding problem, explain why you agree or disagree with each of the following statements:

(*a*) For freshmen, you would expect to find a positive and useful linear correlation between these two variables; but there also are other important factors affecting the grade point average of a college student.

(*b*) For the juniors represented by this sample, there is little, or no, relationship between the ACE-T score and the grade point average.

(*c*) The freshmen and the juniors fit the same general relationship between ACE-T and GPA; the persons with especially low ACE-T scores simply have been eliminated by the time of the junior year.

(*d*) Given that for freshmen the linear correlation between GPA and ACE-L score is .6, whereas that between GPA and ACE-Q is only .4 for these samples, it is concluded that whatever is measured by the *L*-score definitely is more important than whatever is measured by the *Q*-score.

66. Make a scatter diagram for the ACE-L scores of freshmen against their GPA's. And then do likewise for the juniors, using the same coordinate axes. Draw appropriate conclusions.

67. Solve as in problem 65, parts (*a*) to (*c*), but use the results of problem 66 and change ACE-T to ACE-L wherever used.

68. Compute the Spearman rank-difference correlation, r_s, for each scatter diagram of problem 64. Then consider problem 65 in the light of these correlations. *Ans.* For freshmen, $r_s = .59$; for juniors, $r_s = .14$.

69. Compute the Spearman coefficient of linear correlation for each scatter diagram of problem 66 and then solve problem 67 in the light of these results.

70. Make a scatter diagram for the freshmen and for the juniors, as in problem 64, but use the ACE-Q scores. Draw all appropriate conclusions.

REFERENCES

Arley, Niels, and K. Rander Buch, *Introduction to the Theory of Probability and Statistics,* John Wiley and Sons, New York, 1950.

Dixon, Wilfrid J., and Frank J. Massey, Jr., *Introduction to Statistical Analysis,* McGraw-Hill Book Company, New York, 1952.

Freund, John E., *Modern Elementary Statistics,* Prentice-Hall, New York, 1952.

Hald, A., *Statistical Theory with Engineering Applications,* John Wiley and Sons, New York, 1952.

Snedecor, George W., *Statistical Methods Applied to Experiments in Agriculture and Biology,* Fourth Edition, Iowa State College Press, Ames, Iowa, 1946.

Tables

TABLE I

Squares, Square Roots, and Reciprocals

n	$\sqrt{n}$	$\sqrt{10n}$	n^2	$1/n$	n	$\sqrt{n}$	$\sqrt{10n}$	n^2	$1/n$
0.1	0.32	1.00	0.01	10.000	4.6	2.14	6.78	21.16	.217
0.2	0.45	1.41	0.04	5.000	4.7	2.17	6.86	22.09	.213
0.3	0.55	1.73	0.09	3.333	4.8	2.19	6.93	23.04	.208
0.4	0.63	2.00	0.16	2.500	4.9	2.21	7.00	24.01	.204
0.5	0.71	2.24	0.25	2.000	5.0	2.24	7.07	25.00	.200
0.6	0.77	2.45	0.36	1.667	5.1	2.26	7.14	26.01	.196
0.7	0.84	2.65	0.49	1.429	5.2	2.28	7.21	27.04	.192
0.8	0.89	2.83	0.64	1.250	5.3	2.30	7.28	28.09	.189
0.9	0.95	3.00	0.81	1.111	5.4	2.32	7.35	29.16	.185
1.0	1.00	3.16	1.00	1.000	5.5	2.35	7.42	30.25	.182
1.1	1.05	3.32	1.21	0.909	5.6	2.37	7.48	31.36	.179
1.2	1.10	3.46	1.44	0.833	5.7	2.39	7.55	32.49	.175
1.3	1.14	3.61	1.69	0.769	5.8	2.41	7.62	33.64	.172
1.4	1.18	3.74	1.96	0.714	5.9	2.43	7.68	34.81	.169
1.5	1.22	3.87	2.25	0.667	6.0	2.45	7.75	36.00	.167
1.6	1.26	4.00	2.56	0.625	6.1	2.47	7.81	37.21	.164
1.7	1.30	4.12	2.89	0.588	6.2	2.49	7.87	38.44	.161
1.8	1.34	4.24	3.24	0.556	6.3	2.51	7.94	39.69	.159
1.9	1.38	4.36	3.61	0.526	6.4	2.53	8.00	40.96	.156
2.0	1.41	4.47	4.00	0.500	6.5	2.55	8.06	42.25	.154
2.1	1.45	4.58	4.41	0.476	6.6	2.57	8.12	43.56	.152
2.2	1.48	4.69	4.84	0.455	6.7	2.59	8.19	44.89	.149
2.3	1.52	4.80	5.29	0.435	6.8	2.61	8.25	46.24	.147
2.4	1.55	4.90	5.76	0.417	6.9	2.63	8.31	47.61	.145
2.5	1.58	5.00	6.25	0.400	7.0	2.65	8.57	49.00	.143
2.6	1.61	5.10	6.76	0.385	7.1	2.66	8.43	50.41	.141
2.7	1.64	5.20	7.29	0.370	7.2	2.68	8.49	51.84	.139
2.8	1.67	5.29	7.84	0.357	7.3	2.70	8.54	53.29	.137
2.9	1.70	5.39	8.41	0.345	7.4	2.72	8.60	54.76	.135
3.0	1.73	5.48	9.00	0.333	7.5	2.74	8.66	56.25	.133
3.1	1.76	5.57	9.61	0.323	7.6	2.76	8.72	57.76	.132
3.2	1.79	5.66	10.24	0.312	7.7	2.77	8.77	59.29	.130
3.3	1.82	5.74	10.89	0.303	7.8	2.79	8.83	60.84	.128
3.4	1.84	5.83	11.56	0.294	7.9	2.81	8.89	62.41	.127
3.5	1.87	5.92	12.25	0.286	8.0	2.83	8.94	64.00	.125
3.6	1.90	6.00	12.96	0.278	8.1	2.85	9.00	65.61	.123
3.7	1.92	6.08	13.69	0.270	8.2	2.86	9.06	67.24	.122
3.8	1.95	6.16	14.44	0.232	8.3	2.88	9.11	68.89	.120
3.9	1.97	6.24	15.21	0.256	8.4	2.90	9.17	70.56	.119
4.0	2.00	6.32	16.00	0.250	8.5	2.92	9.22	72.25	.118
4.1	2.02	6.40	16.81	0.244	8.6	2.93	9.27	73.96	.116
4.2	2.05	6.48	17.64	0.238	8.7	2.95	9.33	75.69	.115
4.3	2.07	6.56	18.49	0.233	8.8	2.97	9.38	77.44	.114
4.4	2.10	6.63	19.36	0.227	8.9	2.98	9.43	79.21	.112
4.5	2.12	6.71	20.25	0.222	9.0	3.00	9.49	81.00	.111

TABLE I (*Continued*)

SQUARES, SQUARE ROOTS, AND RECIPROCALS

n	$\sqrt{n}$	$\sqrt{10n}$	n^2	$1/n$	n	$\sqrt{n}$	$\sqrt{10n}$	n^2	$1/n$
9.1	3.02	9.54	82.81	.110	13.6	3.69	11.66	184.96	.074
9.2	3.03	9.59	84.64	.109	13.7	3.70	11.70	187.69	.073
9.3	3.05	9.64	86.49	.108	13.8	3.72	11.75	190.44	.072
9.4	3.07	9.70	88.36	.106	13.9	3.73	11.79	193.21	.072
9.5	3.08	9.75	90.25	.105	14.0	3.74	11.83	196.00	.071
9.6	3.10	9.80	92.16	.104	14.1	3.76	11.87	198.81	.071
9.7	3.11	9.85	94.09	.103	14.2	3.77	11.92	201.64	.070
9.8	3.13	9.90	96.04	.102	14.3	3.78	11.96	204.49	.070
9.9	3.15	9.95	98.01	.101	14.4	3.79	12.00	207.36	.069
10.0	3.16	10.00	100.00	.100	14.5	3.81	12.04	210.25	.069
10.1	3.18	10.05	102.01	.099	14.6	3.82	12.08	213.16	.068
10.2	3.19	10.10	104.04	.098	14.7	3.83	12.12	216.09	.068
10.3	3.21	10.15	106.09	.097	14.8	3.85	12.17	219.04	.068
10.4	3.22	10.20	108.16	.096	14.9	3.86	12.21	222.01	.067
10.5	3.24	10.25	110.25	.095	15.0	3.87	12.25	225.00	.067
10.6	3.26	10.30	112.36	.094	15.1	3.89	12.29	228.01	.066
10.7	3.27	10.34	114.49	.093	15.2	3.90	12.33	231.04	.066
10.8	3.29	10.39	116.64	.093	15.3	3.91	12.37	234.09	.065
10.9	3.30	10.44	118.81	.092	15.4	3.92	12.41	237.16	.065
11.0	3.32	10.49	121.00	.091	15.5	3.94	12.45	240.25	.065
11.1	3.33	10.54	123.21	.090	15.6	3.95	12.49	243.36	.064
11.2	3.35	10.58	125.44	.089	15.7	3.96	12.53	246.49	.064
11.3	3.36	10.63	127.69	.088	15.8	3.97	12.57	249.64	.063
11.4	3.38	10.68	129.96	.088	15.9	3.99	12.61	252.81	.063
11.5	3.39	10.72	132.25	.087	16.0	4.00	12.65	256.00	.062
11.6	3.41	10.77	134.56	.086	16.1	4.01	12.69	259.21	.062
11.7	3.42	10.82	136.89	.085	16.2	4.02	12.73	262.44	.061
11.8	3.44	10.86	139.24	.085	16.3	4.04	12.77	265.69	.061
11.9	3.45	10.91	141.61	.084	16.4	4.05	12.81	268.96	.061
12.0	3.46	10.95	144.00	.083	16.5	4.06	12.85	272.25	.061
12.1	3.48	11.00	146.41	.083	16.6	4.07	12.88	275.56	.060
12.2	3.49	11.05	148.84	.082	16.7	4.09	12.92	278.89	.060
12.3	3.51	11.09	151.29	.081	16.8	4.10	12.96	282.24	.060
12.4	3.52	11.14	153.76	.081	16.9	4.11	13.00	285.61	.059
12.5	3.54	11.18	156.25	.080	17.0	4.12	13.04	289.00	.059
12.5	3.55	11.22	158.76	.079	17.1	4.14	13.08	292.41	.058
12.7	3.56	11.27	161.29	.079	17.2	4.15	13.11	295.84	.058
12.8	3.58	11.31	163.84	.078	17.3	4.16	13.15	299.29	.058
12.9	3.59	11.36	166.41	.078	17.4	4.17	13.19	302.76	.057
13.0	3.61	11.40	169.00	.077	17.5	4.18	13.23	306.25	.057
13.1	3.62	11.45	171.61	.076	17.6	4.20	13.27	309.76	.057
13.2	3.63	11.49	174.24	.076	17.7	4.21	13.30	313.29	.056
13.3	3.65	11.53	176.89	.075	17.8	4.22	13.34	316.84	.056
13.4	3.66	11.58	179.56	.075	17.9	4.23	13.38	320.41	.056
13.5	3.67	11.62	182.25	.074	18.0	4.24	13.42	324.00	.056

TABLE I (*Continued*)

SQUARES, SQUARE ROOTS, AND RECIPROCALS

n	$\sqrt{n}$	$\sqrt{10n}$	n^2	$1/n$	n	$\sqrt{n}$	$\sqrt{10n}$	n^2	$1/n$
18.1	4.25	13.45	327.61	.055	21.6	4.65	14.70	466.56	.046
18.2	4.27	13.49	331.24	.055	21.7	4.66	14.73	470.89	.046
18.3	4.28	13.53	334.89	.055	21.8	4.67	14.76	475.24	.046
18.4	4.29	13.56	338.56	.054	21.9	4.68	14.80	479.61	.046
18.5	4.30	13.60	342.25	.054	22.0	4.69	14.83	484.00	.045
18.6	4.31	13.64	345.96	.054	22.1	4.70	14.87	488.41	.045
18.7	4.32	13.67	349.69	.053	22.2	4.71	14.90	492.84	.045
18.8	4.34	13.71	353.44	.053	22.3	4.72	14.93	497.29	.045
18.9	4.35	13.75	357.21	.053	22.4	4.73	14.97	501.76	.045
19.0	4.36	13.78	361.00	.053	22.5	4.74	15.00	506.25	.044
19.1	4.37	13.82	364.81	.052	22.6	4.75	15.03	510.76	.044
19.2	4.38	13.86	368.64	.052	22.7	4.76	15.07	515.29	.044
19.3	4.39	13.89	372.49	.052	22.8	4.77	15.10	519.84	.044
19.4	4.40	13.93	376.36	.052	22.9	4.79	15.13	524.41	.044
19.5	4.42	13.96	380.25	.051	23.0	4.80	15.17	529.00	.043
19.6	4.43	14.00	384.16	.051	23.1	4.81	15.20	533.61	.043
19.7	4.44	14.04	388.09	.051	23.2	4.82	15.23	538.24	.043
19.8	4.45	14.07	392.04	.050	23.3	4.83	15.26	542.89	.043
19.9	4.46	14.11	396.01	.050	23.4	4.84	15.30	547.56	.043
20.0	4.47	14.14	400.00	.050	23.5	4.85	15.33	552.25	.043
20.1	4.48	14.18	404.01	.050	23.6	4.86	15.36	556.96	.042
20.2	4.49	14.21	408.04	.050	23.7	4.87	15.39	561.69	.042
20.3	4.51	14.25	412.09	.049	23.8	4.88	15.43	566.44	.042
20.4	4.52	14.28	416.16	.049	23.9	4.89	15.46	571.21	.042
20.5	4.53	14.32	420.25	.049	24.0	4.90	15.49	576.00	.042
20.6	4.54	14.35	424.36	.049	24.1	4.91	15.52	580.81	.041
20.7	4.55	14.39	428.49	.048	24.2	4.92	15.56	585.64	.041
20.8	4.56	14.42	432.64	.048	24.3	4.93	15.59	590.49	.041
20.9	4.57	14.46	436.81	.048	24.4	4.94	15.62	595.36	.041
21.0	4.58	14.49	441.00	.048	24.5	4.95	15.65	600.25	.041
21.1	4.59	14.53	445.21	.047	24.6	4.96	15.68	605.16	.041
21.2	4.60	14.56	449.44	.047	24.7	4.97	15.72	610.09	.040
21.3	4.62	14.59	453.69	.047	24.8	4.98	15.75	615.04	.040
21.4	4.63	14.63	457.96	.047	24.9	4.99	15.78	620.01	.040
21.5	4.64	14.66	462.25	.047	25.0	5.00	15.81	625.00	.040

TABLE II

Mantissas for Common Logarithms

N	0	1	2	3	4	5	6	7	8	9
1.0	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
1.1	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
1.2	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
1.3	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
1.4	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
1.5	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
1.6	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
1.7	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
1.8	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
1.9	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
2.0	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
2.1	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
2.2	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
2.3	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
2.4	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
2.5	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
2.6	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
2.7	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
2.8	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
2.9	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
3.0	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
3.1	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
3.2	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
3.3	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
3.4	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
3.5	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
3.6	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
3.7	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
3.8	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
3.9	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
4.0	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
4.1	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
4.2	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
4.3	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
4.4	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
4.5	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
4.6	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
4.7	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
4.8	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
4.9	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
5.0	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
5.1	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
5.2	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
5.3	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
5.4	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396

TABLE II (*Continued*)

MANTISSAS FOR COMMON LOGARITHMS

N	0	1	2	3	4	5	6	7	8	9
5.5	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
5.6	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
5.7	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
5.8	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
5.9	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
6.0	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
6.1	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
6.2	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
6.3	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
6.4	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
6.5	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
6.6	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
6.7	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
6.8	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
6.9	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
7.0	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
7.1	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
7.2	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
7.3	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
7.4	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
7.5	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
7.6	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
7.7	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
7.8	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
7.9	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
8.0	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
8.1	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
8.2	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
8.3	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
8.4	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
8.5	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
8.6	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
8.7	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
8.8	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
8.9	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
9.0	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
9.1	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
9.2	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
9.3	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
9.4	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
9.5	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
9.6	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
9.7	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
9.8	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
9.9	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996

TABLE III

FREQUENCY AND RELATIVE CUMULATIVE FREQUENCY DISTRIBUTIONS FOR THE STANDARD NORMAL POPULATION GIVEN FOR THE ABSCISSAS FROM $\lambda = -3.00$ TO $\lambda = +3.00$

Abscissas λ	Ordinates y	$r.c.f.$	Abscissas λ	Ordinates y	$r.c.f.$	Abscissas λ	Ordinates y	$r.c.f.$
−3.00	.004	.001	−1.14	.208	.127	−0.30	.381	.382
−2.90	.006	.002	−1.12	.213	.131	−0.28	.384	.390
−2.80	.008	.003	−1.10	.218	.136	−0.26	.386	.397
−2.70	.010	.003	−1.08	.223	.140	−0.24	.388	.405
−2.60	.014	.005	−1.06	.227	.145	−0.22	.389	.413
−2.50	.018	.006	−1.04	.232	.149	−0.20	.391	.421
−2.40	.022	.008	−1.02	.237	.154	−0.18	.393	.429
−2.30	.028	.011	−1.00	.241	.159	−0.16	.394	.436
−2.25	.032	.012	−0.98	.247	.164	−0.14	.395	.444
−2.20	.035	.014	−0.96	.252	.169	−0.12	.396	.452
−2.15	.040	.016	−0.94	.256	.174	−0.10	.397	.460
−2.10	.044	.018	−0.92	.261	.179	−0.08	.398	.468
−2.05	.049	.020	−0.90	.266	.184	−0.06	.398	.476
−2.00	.054	.023	−0.88	.271	.189	−0.04	.399	.484
−1.95	.060	.026	−0.86	.276	.195	−0.02	.399	.492
−1.90	.066	.029	−0.84	.280	.200	0.00	.399	.500
−1.85	.072	.032	−0.82	.285	.206	+0.02	.399	.508
−1.80	.079	.036	−0.80	.290	.212	0.04	.399	.516
−1.75	.086	.040	−0.78	.294	.218	0.06	.398	.524
−1.70	.094	.045	−0.76	.299	.224	0.08	.398	.532
−1.66	.101	.048	−0.74	.303	.230	0.10	.397	.540
−1.62	.107	.053	−0.72	.308	.236	0.12	.396	.548
−1.58	.114	.057	−0.70	.312	.242	0.14	.395	.556
−1.54	.122	.062	−0.68	.317	.248	0.16	.394	.564
−1.50	.130	.067	−0.66	.321	.255	0.18	.393	.571
−1.48	.133	.069	−0.64	.325	.261	0.20	.391	.579
−1.46	.137	.072	−0.62	.329	.268	0.22	.389	.587
−1.44	.141	.075	−0.60	.333	.274	0.24	.388	.595
−1.42	.146	.078	−0.58	.337	.281	0.26	.386	.603
−1.40	.150	.081	−0.56	.341	.288	0.28	.384	.610
−1.38	.154	.084	−0.54	.345	.295	0.30	.381	.618
−1.36	.158	.087	−0.52	.348	.302	0.32	.379	.626
−1.34	.163	.090	−0.50	.352	.309	0.34	.377	.633
−1.32	.167	.093	−0.48	.356	.316	0.36	.374	.641
−1.30	.171	.097	−0.46	.359	.323	0.38	.371	.648
−1.28	.176	.100	−0.44	.362	.330	0.40	.368	.655
−1.26	.180	.104	−0.42	.365	.337	0.42	.365	.663
−1.24	.185	.107	−0.40	.368	.345	0.44	.362	.670
−1.22	.190	.111	−0.38	.371	.352	0.46	.359	.677
−1.20	.194	.115	−0.36	.374	.359	0.48	.356	.684
−1.18	.199	.119	−0.34	.377	.367	0.50	.352	.691
−1.16	.204	.123	−0.32	.379	.374	0.52	.348	.698

TABLE III (*Continued*)

FREQUENCY AND RELATIVE CUMULATIVE FREQUENCY DISTRIBUTIONS FOR THE STANDARD NORMAL POPULATION GIVEN FOR THE ABSCISSAS FROM $\lambda = -3.00$ TO $\lambda = +3.00$

Abscissas λ	Ordinates y	$r.c.f.$	Abscissas λ	Ordinates y	$r.c.f.$	Abscissas λ	Ordinates y	$r.c.f.$
0.54	.345	.705	1.04	.232	.851	1.54	.122	.938
0.56	.341	.712	1.06	.227	.855	1.58	.114	.943
0.58	.337	.719	1.08	.223	.860	1.62	.107	.947
0.60	.333	.726	1.10	.218	.864	1.66	.101	.952
0.62	.329	.732	1.12	.213	.869	1.70	.094	.955
0.64	.325	.739	1.14	.208	.873	1.75	.086	.960
0.66	.321	.745	1.16	.204	.877	1.80	.079	.964
0.68	.317	.752	1.18	.199	.881	1.85	.072	.968
0.70	.312	.758	1.20	.194	.885	1.90	.066	.971
0.72	.308	.764	1.22	.190	.888	1.95	.060	.974
0.74	.303	.770	1.24	.185	.893	2.00	.054	.977
0.76	.299	.776	1.26	.180	.896	2.05	.049	.980
0.78	.294	.782	1.28	.176	.900	2.10	.044	.982
0.80	.290	.788	1.30	.171	.903	2.15	.040	.984
0.82	.285	.794	1.32	.167	.907	2.20	.035	.986
0.84	.280	.800	1.34	.163	.910	2.25	.032	.988
0.86	.276	.805	1.36	.158	.913	2.30	.028	.989
0.88	.271	.811	1.38	.154	.916	2.40	.022	.992
0.90	.266	.816	1.40	.150	.919	2.50	.018	.994
0.92	.261	.821	1.42	.146	.922	2.60	.014	.995
0.94	.256	.826	1.44	.141	.925	2.70	.010	.997
0.96	.252	.831	1.46	.137	.928	2.80	.008	.997
0.98	.247	.836	1.48	.133	.931	2.90	.006	.998
1.00	.241	.841	1.50	.130	.933	3.00	.004	.999
1.02	.237	.846						

TABLE IV

RELATIVE CUMULATIVE FREQUENCY DISTRIBUTION OF t SHOWING THE PROPORTIONS OF ALL SAMPLING t_i WITH THE SAME DEGREES OF FREEDOM WHICH ARE LESS THAN THE t SHOWN IN COLUMN 1 ON THE LEFT

	Degrees of Freedom											
t	8	9	10	11	12	14	16	18	20	22	26	30
−5.0	.001	.000										
−4.6	.001	.001	.000	.000	.000							
−4.2	.002	.001	.001	.001	.001	.000	.000	.000	.000			
−3.8	.003	.002	.002	.001	.001	.001	.001	.001	.001	.000	.000	.000
−3.4	.005	.004	.003	.003	.003	.002	.002	.002	.001	.001	.001	.001
−3.0	.009	.007	.007	.006	.006	.005	.004	.004	.004	.003	.003	.003
−2.8	.012	.010	.009	.009	.008	.007	.006	.006	.006	.005	.005	.004
−2.6	.016	.014	.013	.012	.012	.010	.010	.009	.009	.008	.008	.007
−2.4	.022	.020	.019	.018	.017	.015	.014	.014	.013	.013	.012	.011
−2.2	.030	.028	.026	.025	.024	.023	.021	.021	.020	.019	.019	.018
−2.0	.040	.038	.037	.035	.034	.033	.031	.030	.030	.029	.028	.027
−1.8	.055	.053	.051	.050	.049	.047	.045	.044	.043	.043	.042	.041
−1.6	.074	.072	.070	.969	.068	.066	.065	.064	.063	.062	.061	.060
−1.4	.100	.098	.096	.095	.093	.092	.090	.089	.088	.088	.087	.086
−1.2	.132	.130	.129	.128	.127	.125	.124	.123	.122	.121	.121	.120
−1.0	.173	.172	.170	.169	.169	.167	.166	.165	.165	.164	.163	.163
−0.8	.223	.222	.221	.220	.220	.219	.218	.217	.217	.216	.216	.215
−0.6	.283	.282	.281	.280	.280	.279	.278	.278	.278	.277	.277	.277
−0.4	.350	.349	.349	.348	.348	.348	.347	.347	.347	.347	.346	.346
−0.2	.423	.423	.423	.423	.422	.422	.422	.422	.422	.422	.421	.421
0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.2	.577	.577	.577	.577	.578	.578	.578	.578	.578	.578	.579	.579
0.4	.650	.651	.651	.652	.652	.652	.653	.653	.653	.653	.654	.654
0.6	.717	.718	.719	.720	.720	.721	.722	.722	.722	.723	.723	.723
0.8	.777	.778	.779	.780	.780	.781	.782	.783	.783	.784	.784	.785
1.0	.827	.828	.830	.831	.831	.833	.834	.835	.835	.836	.837	.837
1.2	.868	.870	.871	.872	.873	.875	.876	.877	.878	.879	.879	.880
1.4	.900	.902	.904	.905	.907	.908	.910	.911	.912	.912	.913	.914
1.6	.926	.928	.930	.931	.932	.934	.935	.936	.937	.938	.939	.940
1.8	.945	.947	.949	.950	.951	.953	.955	.956	.957	.957	.958	.959
2.0	.960	.962	.963	.965	.966	.967	.969	.970	.970	.971	.972	.973
2.2	.970	.972	.974	.975	.976	.977	.979	.979	.980	.981	.981	.982
2.4	.978	.980	.981	.982	.983	.985	.986	.986	.987	.987	.988	.989
2.6	.984	.986	.987	.988	.988	.990	.990	.991	.991	.992	.992	.993
2.8	.988	.990	.991	.991	.992	.993	.994	.994	.994	.995	.995	.996
3.0	.991	.993	.993	.994	.994	.995	.996	.996	.996	.997	.997	.997
3.4	.995	.996	.997	.997	.997	.998	.998	.998	.999	.999	.999	.999
3.8	.997	.998	.998	.999	.999	.999	.999	.999	.999	1.000	1.000	1.000
4.2	.998	.999	.999	.999	.999	1.000	1.000	1.000	1.000			
4.6	.999	.999	1.000	1.000	1.000							
5.0	.999	1.000										

SOME FREQUENTLY USED t'S CORRESPONDING TO PRE-ASSIGNED PROBABILITIES OF OCCURRENCE DURING RANDOM SAMPLING

	Degrees of Freedom											
$P(\lvert t \rvert > t_0)$	8	9	10	11	12	14	16	18	20	22	26	30
.100	1.86	1.83	1.81	1.80	1.78	1.76	1.75	1.73	1.72	1.72	1.71	1.70
.050	2.31	2.26	2.23	2.20	2.18	2.14	2.12	2.10	2.09	2.07	2.06	2.04
.010	3.36	3.25	3.17	3.11	3.06	2.98	2.92	2.88	2.84	2.82	2.78	2.75
.001	5.04	4.78	4.59	4.44	4.32	4.14	4.02	3.92	3.85	3.79	3.71	3.65

TABLE V

Relative Cumulative Frequency Distribution of χ^2 Showing Proportion of All Sampling χ^2 with Same Degrees of Freedom Which Are Greater Than the χ^2 Shown on the Left

χ^2	Degrees of Freedom 1	2	3	χ^2	Degrees of Freedom 1	2	3
0.40	.53			6.00	.014	.050	.11
0.50	.48			6.25	.012	.044	.10
0.60	.44			6.50	.011	.039	.090
0.70	.40			6.75	.009	.034	.082
0.80	.37			7.00	.008	.030	.072
0.90	.34			7.50	.006	.023	.057
1.00	.32			8.00	.005	.018	.046
1.25	.26	.53		8.50	.003	.014	.036
1.50	.22	.47		9.00	.003	.011	.029
1.75	.19	.42		9.50	.002	.009	.024
2.00	.16	.37		10.00	.002	.007	.019
2.25	.13	.33	.50	10.50	.001	.005	.015
2.50	.11	.29	.47	11.00	.000	.004	.012
2.75	.10	.25	.43	11.50		.003	.010
3.00	.084	.22	.39	12.00		.002	.007
3.25	.071	.20	.35	12.50		.002	.006
3.50	.061	.18	.32	13.00		.002	.005
3.75	.053	.15	.29	13.50		.001	.004
4.00	.046	.13	.26	14.00		.001	.003
4.25	.039	.12	.24	14.50		.001	.002
4.50	.034	.10	.21	15.00		.001	.002
4.75	.029	.092	.19	15.50		.000	.002
5.00	.025	.082	.17	16.00			.001
5.25	.022	.073	.15	17.00			.001
5.50	.019	.064	.14	18.00			.000
5.75	.016	.057	.12				

TABLE VI

VALUES OF THE FUNCTION, $y = (1/\sqrt{2\pi}) \cdot e^{-w}$

w	y	w	y	w	y
0	.399	1.8	.066	4.2	.006
0.1	.361	1.9	.060	4.4	.005
0.2	.327	2.0	.054	4.6	.004
0.3	.295	2.1	.049	4.8	.003
0.4	.267	2.2	.044	5.0	.003
0.5	.242	2.3	.040	5.2	.002
0.6	.219	2.4	.036	5.4	.002
0.7	.198	2.5	.033	5.6	.001
0.8	.179	2.6	.030	5.8	.001
0.9	.162	2.7	.027	6.0	.001
1.0	.147	2.8	.024	6.2	.001
1.1	.133	2.9	.022	6.4	.001
1.2	.120	3.0	.020	6.6	.001
1.3	.109	3.2	.016	6.8	.000
1.4	.099	3.4	.013		
1.5	.089	3.6	.011		
1.6	.081	3.8	.009		
1.7	.073	4.0	.007		

TABLE VII

Binomial Coefficients: $C_{n,r} = n!/r!(n-r)!$

	n															
r	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	10	15	21	28	36	45	55	66	78	91	105	120	136	153	171	190
3	10	20	35	56	84	120	165	220	286	364	455	560	680	816	969	1140
4	5	15	35	70	126	210	330	495	715	1001	1365	1820	2380	3060	3876	4845
5	1	6	21	56	126	252	462	792	1287	2002	3003	4368	6188	8568	11628	15504
6		1	7	28	84	210	462	924	1716	3003	5005	8008	12376	18564	27132	38760
7			1	8	36	120	330	792	1716	3432	6435	11440	19448	31824	50388	77520
8				1	9	45	165	495	1287	3003	6435	12870	24310	43758	75582	125970
9					1	10	55	220	715	2002	5005	11440	24310	48620	92378	167960
10						1	11	66	286	1001	3003	8008	19448	43758	92378	184756
11							1	12	78	364	1365	4368	12376	31824	75582	167960
12								1	13	91	455	1820	6188	18564	50388	125970
13									1	14	105	560	2380	8568	27132	77520
14										1	15	120	680	3060	11628	38760
15											1	16	136	816	3876	15504
16												1	17	153	969	4845
17													1	18	171	1140
18														1	19	190
19															1	20
20																1

TABLE VIII

Factorials and Their Logarithms

n	$n!$	$\log n!$	n	$n!$	$\log n!$	n	$n!$	$\log n!$
2	2	0.3010	18	640.24×10^{13}	15.8063	34	295.23×10^{36}	38.4702
3	6	0.7782	19	121.65×10^{15}	17.0851	35	103.33×10^{38}	40.0142
4	24	1.3802	20	243.29×10^{16}	18.3861	36	371.99×10^{39}	41.5705
5	120	2.0792	21	510.91×10^{17}	19.7083	37	137.64×10^{41}	43.1387
6	720	2.8573	22	112.40×10^{19}	21.0508	38	523.02×10^{42}	44.7185
7	5040	3.7024	23	258.52×10^{20}	22.4125	39	203.98×10^{44}	46.3096
8	40320	4.6055	24	620.45×10^{21}	23.7927	40	815.92×10^{45}	47.9116
9	362880	5.5598	25	155.11×10^{23}	25.1906	41	334.53×10^{47}	49.5244
10	3628800	6.5598	26	403.29×10^{24}	26.6056	42	140.50×10^{49}	51.1477
11	399.17×10^{5}	7.6012	27	108.89×10^{26}	28.0370	43	604.15×10^{50}	52.7811
12	479.00×10^{6}	8.6803	28	304.89×10^{27}	29.4841	44	265.83×10^{52}	54.4246
13	622.70×10^{7}	9.7943	29	884.18×10^{28}	30.9465	45	119.62×10^{54}	56.0778
14	871.78×10^{8}	10.9404	30	265.25×10^{30}	32.4237	46	550.26×10^{55}	57.7406
15	130.77×10^{10}	12.1165	31	822.28×10^{31}	33.9150	47	258.62×10^{57}	59.4127
16	209.23×10^{11}	13.3206	32	263.13×10^{33}	35.4202	48	124.14×10^{59}	61.0939
17	355.69×10^{12}	14.5511	33	868.33×10^{34}	36.9387	49	608.28×10^{60}	62.7841

TABLE IX

PROBABILITY DISTRIBUTION OF $G = |\bar{x} - \mu|/(\text{RANGE})$ FOR A SAMPLE OF SIZE n FROM A NORMAL POPULATION

Sample Size, n	Probability that G will be greater than table value					
	.100	.050	.020	.010	.002	.001
2	3.157	6.353	15.910	31.828	159.16	318.31
3	0.885	1.304	2.111	3.008	6.77	9.58
4	0.529	0.717	1.023	1.316	2.29	2.85+
5	0.388	0.507	0.685+	0.843	1.32	1.58
6	0.312	0.399	0.523	0.628	0.92	1.07
7	0.263	0.333	0.429	0.507	0.71	0.82
8	0.230	0.288	0.366	0.429	0.59	0.67
9	0.205	0.255	0.322	0.374	0.50	0.57
10	0.186	0.230	0.288	0.333	0.44	0.50
11	0.170	0.210	0.262	0.302	0.40	0.44
12	0.158	0.194	0.241	0.277	0.36	0.40
13	0.147	0.181	0.224	0.256	0.33	0.37
14	0.138	0.170	0.209	0.239	0.31	0.34
15	0.131	0.160	0.197	0.224	0.29	0.32
16	0.124	0.151	0.186	0.212	0.27	0.30
17	0.118	0.144	0.177	0.201	0.26	0.28
18	0.113	0.137	0.168	0.191	0.24	0.26
19	0.108	0.131	0.161	0.182	0.23	0.25+
20	0.104	0.126	0.154	0.175−	0.22	0.24

The above table was derived from Table 9, page 66, vol. 34, *Biometrika*, in an article by E. Lord, with the permission of the publishers of *Biometrika*.

TABLE X

PROBABILITY DISTRIBUTION OF $G = |\bar{x}_1 - \bar{x}_2|/(\text{MEAN RANGE})$ FOR TWO SAMPLES EACH OF SIZE n FROM THE SAME NORMAL POPULATION

Sample Size, n	Probability that G will be greater than tabular value					
	.100	.050	.020	.010	.002	.001
2	2.322	3.427	5.553	7.916	17.81	25.23
3	0.974	1.272	1.715−	2.093	3.27	4.18
4	0.644	0.813	1.047	1.237	1.74	1.99
5	0.493	0.613	0.772	0.896	1.21	1.35+
6	0.405+	0.499	0.621	0.714	0.94	1.03
7	0.347	0.426	0.525+	0.600	0.77	0.85−
8	0.306	0.373	0.459	0.521	0.67	0.73
9	0.275−	0.334	0.409	0.464	0.59	0.64
10	0.250	0.304	0.371	0.419	0.53	0.58
11	0.233	0.280	0.340	0.384	0.48	0.52
12	0.214	0.260	0.315+	0.355+	0.44	0.48
13	0.201	0.243	0.294	0.331	0.41	0.45−
14	0.189	0.228	0.276	0.311	0.39	0.42
15	0.179	0.216	0.261	0.293	0.36	0.39
16	0.170	0.205−	0.247	0.278	0.34	0.37
17	0.162	0.195+	0.236	0.264	0.33	0.35+
18	0.155+	0.187	0.225+	0.252	0.31	0.34
19	0.149	0.179	0.216	0.242	0.30	0.32
20	0.143	0.172	0.207	0.232	0.29	0.31

The above table was derived from the same article as Table IX, also with the permission of the publishers of *Biometrika*.

Index