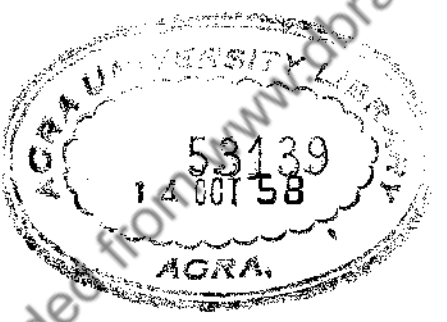
ELEMENTARY STATISTICAL ANALYSIS

By S. S. WILKS



PRINCETON UNIVERSITY PRESS
PRINCETON, NEW JERSEY

Reprinted with corrections but
without change of pagination

JANUARY 1951
SEVENTH PRINTING 1954
EIGHTH PRINTING 1956

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

This book has been prepared for a one-semester basic course in elementary statistical analysis which, at Princeton, is the introductory course for all fields of statistical application, and is usually taken in the freshman year. It is especially designed for those who intend to go into the biological and social sciences. It presupposes one semester of elementary mathematical analysis covering topics such as those included in the first half of F. L. Griffin's Introduction to Mathematical Analysis. The material has been developed from two years of experience with such a course.

An effort has been made throughout the book to emphasize the role played in statistical analysis by a sample of measurements and a population from which the sample is supposed to have arisen. Only three chapters are devoted to elementary descriptive statistics of a sample of measurements. In these three chapters the idea of a population is presented on a purely intuitive basis. Probability concepts are then introduced. This makes it possible to use these basic concepts at an early stage in dealing more critically with the idea of a population and sampling from a population. Considerable attention is given to the application of sampling principles to the simpler problems of statistical inference such as determining confidence limits of population means and difference of means, making elementary significance tests, testing for randomness, etc. No attempt has been made here (in fact, there is not enough time in one semester!) to go into analysis of variance and more sophisticated problems of statistical inference. An elementary treatment of analysis of pairs of measurements including least squares methods is presented. Special effort has been made throughout the book to keep the mathematics elementary and to state specifically at which points the mathematics is too advanced to present.

The course in which this material has been used has been conducted satisfactorily (not ideally!) without the use of a computing laboratory. The problems in the exercises have been selected so that computations can be carried out effectively by the use of a small handbook of tables such as C. D. Hodgman's Mathematical Tables from Handbook of Chemistry and Physics.

The author would like to express his appreciation to: Professor R. A. Fisher and Messrs. Oliver and Boyd for permission to use the material in Table 10.2; to Professor E. S. Pearson and the Biometrika Office for permission to reprint Figures 10.1 and 13.6; to Dr. C. Eisenhart and Miss Freda S. Swed for permission to use the material in Table 12.3; and to the College Entrance Examination Board for permission to reprint Figure 13.5.

Finally, the author takes this opportunity to acknowledge the benefit of many helpful discussions he has had with his colleagues, Professors A. W. Tucker and J. W. Tukey, and Professor F. Mosteller of Harvard University during the preparation of the material. He is also indebted to Drs. K. L. Chung, R. Otter and D. F. Votaw, who assisted with the preparation of Chapters 4, 6, 7

and 8; to Mr. J. G. C. Templeton, who checked the computations and proofread the manuscript; to Mr. R. B. Murphy who drew the figures; and to Mrs. Frances Purvis who typed the manuscript.

The material in this book is still in a tentative form. Any errors or weaknesses in presentation are solely the responsibility of the author. Corrections, criticisms, and expressions of other points of view on the teaching of such a course will be gratefully received.

S. S. Wilks

Princeton, New Jersey
September
1948

CONTENTS

CHAPTER 1. INTRODUCTION

1.1 General Remarks.

To many persons the word statistics means neatly arranged tables of figures and bar charts printed in financial sections of newspapers or issued by almanac publishers and government agencies. They have the impression that statistics are figures used by persons called statisticians to prove or disprove something. There is plenty of ground for this impression. Anyone who tries to make sense out of a set of observational or experimental data is assuming the role of a statistician, no matter whether he is a business executive, a medical research man, a biologist, a public opinion poller or an economist. Some sets of data are very simple and the implications and conclusions inherent in the data are obvious. Other data, however, are complex and may trick and confuse the statistical novice, even though he may be an expert in the subject matter field from which the data came. The only way to reduce this confusion is through scientific methods of collecting, analyzing and interpreting data. Such methods have been developed and are available. The fact that expert statisticians well-versed in these methods can and do come out with sound conclusions from a given set of data which differ very little from one statistician to another is evidence that there are no real grounds for the naive claim that statistics can prove anything. Some of the most dangerously deceptive uses of statistics occur in situations where correct conclusions are drawn and which seem to depend on the statistics, when in fact, the statistics have little if anything to do with the original question. While mathematics cannot protect a person from this danger directly, familiarity with numerical analysis will make it easier to spot such hidden fallacies.

Modern statistical method is a science in itself, dealing with such questions as: How shall a program of obtaining data be planned so that reliable conclusions can be made from the data? How shall the data be analyzed? What conclusions are we entitled to draw from the data? How reliable are the conclusions? To try to present all the statistical methods that are known and used at present would be an encyclopaedic venture which would lead us deeply into statistical theory and many subject matter fields. However, there is a

body of fundamental concepts and elementary methods which can be presented in a beginning course. The purpose of this course is to do just this, and to illustrate the concepts and methods on simple examples and problems from various fields.

You will ask at this juncture what kinds of situations come up which involve these fundamental concepts and elementary methods. Series or sets of raw statistical observations or measurements arise in many ways and in many different fields. The number of observations or measurements needed or feasible varies tremendously from one situation to another. In some cases, as in peacetime firing of large caliber naval guns, only a very small sample of measurement (of where the shells actually fall) can be obtained because of cost. In other cases, as for example in a Gallup poll for a presidential election, the number of observations runs into the thousands. In some situations the sample comprises the entire population of measurements or observations which could be made, particularly in census-type work where complete enumerations of populations are made. Federal and state government agencies and national associations compile data on entire populations of objects, e.g., births, deaths, automobile registrations, number of life insurance policies, etc., etc.

There are two general types of statistical observations: (1) quantitative and (2) qualitative. We shall discuss these separately.

1.2 Quantitative Statistical Observations.

By quantitative statistical observations we mean a sequence or set of numerical measurements or observations made on some or all of the objects in a specified population of objects. If the observations are made on some of the objects we call the set of observations a sample. Let us illustrate by some examples.

Suppose a men's clothing store proprietor writes down from sales slips the sizes of men's overcoats sold every other week for September and October. He would end up with a list of numbers that might run something like this: 36, 42, 44, 30, 40, 36, ... and so on for 145 numbers. The list of numbers written down constitutes a sample of sizes from the population of overcoats he has sold during September and October.

By making an analysis of a series of specimens from a certain deposit of ore, for percent of iron, a chemist might turn up with 1 measurement on each of five specimens something like this: 28.2, 27.6, 29.3, 28.2, 30.1. This is

body of fundamental concepts and elementary methods which can be presented in a beginning course. The purpose of this course is to do just this, and to illustrate the concepts and methods on simple examples and problems from various fields.

You will ask at this juncture what kinds of situations come up which involve these fundamental concepts and elementary methods. Series or sets of raw statistical observations or measurements arise in many ways and in many different fields. The number of observations or measurements needed or feasible varies tremendously from one situation to another. In some cases, as in peacetime firing of large caliber naval guns, only a very small sample of measurements (of where the shells actually fall) can be obtained because of cost. In other cases, as for example in a Gallup poll for a presidential election, the number of observations runs into the thousands. In some situations the sample comprises the entire population of measurements or observations which could be made, particularly in census-type work where complete enumerations of populations are made. Federal and state government agencies and national associations compile data on entire populations of objects, e.g., births, deaths, automobile registrations, number of life insurance policies, etc., etc.

There are two general types of statistical observations: (1) quantitative and (2) qualitative. We shall discuss these separately.

1.2 Quantitative Statistical Observations.

By quantitative statistical observations we mean a sequence or set of numerical measurements or observations made on some or all of the objects in a specified population of objects. If the observations are made on some of the objects we call the set of observations a sample. Let us illustrate by some examples.

Suppose a men's clothing store proprietor writes down from sales slips the sizes of men's overcoats sold every other week for September and October. He would end up with a list of numbers that might run something like this: 38, 42, 44, 30, 40, 36, ... and so on for 145 numbers. The list of numbers written down constitutes a sample of sizes from the population of overcoats he has sold during September and October.

By making an analysis of a series of specimens from a certain deposit of ore, for percent of iron, a chemist might turn up with 1 measurement on each of five specimens something like this: 28.2, 27.6, 29.3, 28.2, 30.1. This is

a sample of iron percentage measurements from five specimens out of an extremely large population of possible specimens from the deposit.

A record-keeping bridge player might keep track of the number of honor cards he gets in 200 bridge hands finding some such sequence as: 9, 5, 7, 2, 4, 8, 0, 3 and so on for 200 numbers. He would therefore accumulate honor card counts in a sample of 200 hands out of a population of indefinitely many hands which could be conceivably dealt under a given shuffling and cutting practice.

A quality control inspector interested in maintaining control of the inner diameter of bushings turned out by an automatic lathe would pick a bushing every 30 minutes and measure its inner diameter, obtaining some such sequence as this: 1.001", .998", .999", 1.001", 1.002", etc. He is selecting a sample of bushings out of the population of bushings being manufactured by this lathe.

A Princeton personnel researcher goes through the record cards of all 246 freshmen who took Mathematics 103 and writes down two numbers for each student; his College Board mathematics score and his final group in Mathematics 103. His sequence of pairs of numbers (arranged alphabetically with respect to students' names) might run like this (680, 3+), (740, 1-), (530, 5), (620, 3), (510, 6) and so on for 246 pairs of scores. In this case the sample would consist of all of the freshmen in the population of freshmen who took Mathematics 103.

Notice that in this last example each quantity in the sequence consists of two measurements. We could mention many examples in which the sequence would contain not only pairs, but sets of three, four or more measurements.

We could continue with dozens of such examples. It is to be noted that in every example mentioned, the series of statistical measurements may be regarded as a sample of measurements from a population of measurements. In general, there are two kinds of populations: finite populations and indefinitely large populations. For example, the undergraduates now enrolled at Princeton constitute a finite population. The licensed hunters of Pennsylvania form a finite population. The sequence of numbers of dots obtained by rolling a pair of dice indefinitely many times is an indefinitely large population. In the case of the dice, the indefinitely large population consisting of the sequence of dots is generated by successively rolling the dice indefinitely many times. The population in this case depends on various factors, such as the dice themselves (which may be slightly biased), the method of throwing them and the surface on which they are thrown. If the dice were "perfectly true", and if they were thoroughly shaken before throwing and if they were thrown on a "perfect" table top, we can imagine having an "ideal" population. We can use probability theory (to be discussed in later

sections) to predict characteristics of this "ideal" population, such as the fraction of rolls of two dice in which 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 dots will occur "in the long run", the fraction of sets of three rolls of two dice in which 6, 2, 11 dots are obtained in that order, and so on. In the case of the lathe turning out bushings, we may essentially consider the population to be indefinitely large, since the population is being generated by the production of one bushing after another, with no consideration of a "last bushing" (which, as a matter of fact, sooner or later will be made). But the important thing about this population of bushings is that it actually changes because of tool wear or changes in raw material from which bushings are made or change of operators, etc. For any particular shift of operators the population may be fairly constant and a sample of inner diameters of bushings taken during that shift may be considered as a sample from an indefinitely large potential population of bushings that might be turned out under the particular conditions of that shift. Even in the case of a finite population of objects, a given sampling procedure might be such that when applied to a relatively small number of objects in the population it essentially begins to generate a population different from the finite population one thinks he is sampling. For example, if one should take every 20th residence listed in the Princeton telephone directory and call the number for information about that residence, one has, on the face of it, a sampling procedure which might be expected to yield information from which one could make accurate inferences about the population of Princeton residences with telephones. Actually, there will be a substantial number of residences for which there will be no response. If we take the sample of residences in which a response is obtained, our sampling procedure is not sampling the population of residences with telephones -- it is sampling the population of residences with telephones in which telephones are answered. These two populations of residences are actually different. For example, the second tends to have larger families and more old people and other stay-at-home types of people in them. Of course, if we make enough repeated telephone calls to the residences who did not answer the telephone originally, we would then be sampling the first population.

What is supposed to be done with samples of measurements? The main reason for keeping track of such measurements is not simply to accumulate a lot of numbers, but, in general, to try to learn something about the main features of the set of numbers -- their average, how much they vary from one another, etc., -- for the purpose of making inferences about the population from which

they can be considered as having been "drawn". None of these measurement-makers want to get any more data than necessary to make these inferences. Once he has what he thinks is a pretty sound inference as to what the population is (that is, a "reasonably" accurate description of it from the sample) he can then begin to consider what ought to be done (perhaps nothing) to change it in some direction or other which will be to his advantage, or more often, to use this information elsewhere.

The clothing store proprietor can find out from a sample whether he is stocking the right distribution of sizes of overcoats; the chemist (or rather his boss) can use the results of his sample of analyses to help decide whether the iron ore is worth mining; the bridge player can satisfy his curiosity as to how frequently various numbers of honor cards are obtained (since he presumably does not want to try to figure these things out mathematically on the assumption of perfect shuffling); the quality control expert can see whether the inner diameters of his bushings are being kept within the specified tolerances and if not whether the holes are being made too large or too small and by how much; the personnel researcher can determine how high the relationship or correlation is between the College Board mathematics test and the final group in Mathematics 103 and whether it is high enough to make useful predictions as to how well each entering freshman can be expected to do on Mathematics 103 from a knowledge of his College Board mathematics score.

Evidently, condensing the sample data in some way is vital in any one of these problems. The first thing that has to be learned in statistics is how to condense the sample data and present it satisfactorily. The main thing that has to be learned is what kind of inferences or statements can be made from the sample about the population sampled and how reliable these inferences are. The simplest thing that can be done in condensing and describing samples of quantitative data is to make frequency distributions and describe them by calculating certain kinds of averages. Such quantities calculated from samples for describing samples are called statistics. Similarly, populations are described by population parameters.

Only rarely is it possible to know precisely the values of population parameters, simply because only rarely does one ever have the data for the entire population. The usual situation is that one only has a sample from the population. Hence the usual problem is to calculate statistics from the sample frequency distribution and then try to figure out from the values of these statistics

what the values of the parameters of the population are likely to be. In case of extremely large samples, the statistics of properly drawn samples will have values very close to those of the corresponding population parameters. For example, the average of a very large sample of measurements "randomly drawn" from a population will be very close to the average of the entire population of measurements. But in the case of small samples the discrepancies become larger, and the problem of inferring the values of population parameters from sample statistics becomes more complicated and has to be settled by means of probability theory.

There is a source of information in sequences of observations which is particularly useful in such fields as analysis of data from scientific experiments and industrial research, development and production. This is the information contained in the way in which measurements jump about from value to value as one goes through the sequence of sample measurements in the order in which they are made. The usual frequency distribution analysis (to be studied in Chapters 2 and 3) does not take account of this information. But we shall discuss this kind of sequence analysis in Chapter 12.

1.3 Qualitative statistical observations.

By qualitative statistical observations we mean a sequence of observations in which each observation in the sample (as well as the population) belongs to one of several mutually exclusive classes which are likely to be non-numerical. Let us consider some examples.

A person tosses a coin 50 times and obtains some such sequence as H, H, T, T, H, T and so on (H=heads, T=tails). He is essentially drawing a sample of 50 tosses out of an indefinitely large population of tosses, and is making an observation on each toss as to whether it is an H or a T.

A movie producer polling agent stationed at the exit of the Princeton Playhouse asking outgoing moviegoers whether or not they liked Movie X (just seen), might get a sequence of 100 answers starting off like this: Yes, Yes, No, Yes, No, Yes, Yes and so on. (He probably wouldn't stop at this simple question, however, since he would probably at least want to know why he or she liked it or not.) The answers here are qualitative; they are either yes or no. The data accumulated are responses from a sample of moviegoers out of the population of moviegoers who saw Movie X at the Playhouse.

A Washington traffic analyst interested in out-of-DC cars coming into Washington during July 1948, might place traffic-counting clerks for three one-hour periods on each odd-numbered day in July at each of the major highway entrances into Washington to check license plates and record the state (ignoring Virginia and Maryland perhaps) for each car. The record for each clerk would be a sequence of state names or initials (or tally marks on a list of states). These clerks are drawing samples of out-of-DC cars out of the population of out-of-DC cars going into Washington during July, 1948.

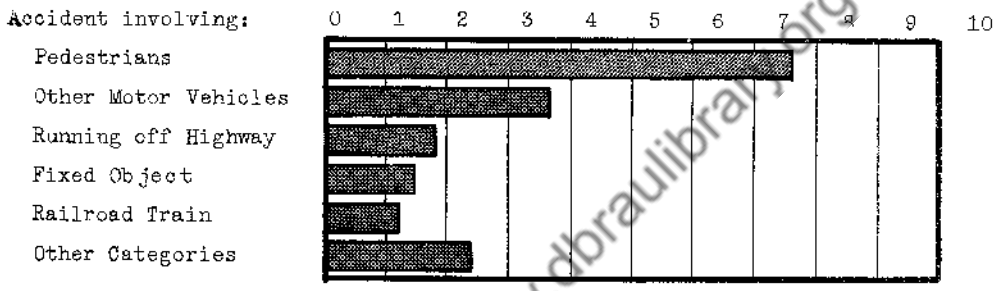
As in the case of quantitative statistical data, we can have samples of observations, each of which consists of pairs, triplets, or any number of qualitative observations. For example, if a public opinion questionnaire with ten opinion questions with yes, no or no-opinion as possible answers on each question were submitted to each of 100 people, the results would be 100 sets, with ten observations in each set.

For qualitative observations the problems of how much data to gather are similar to those for quantitative observations. However, the problems of condensing the data and making analyses of them are, in general, simpler than those for quantitative data. These problems of condensing the data in a sample of qualitative observations is one of counting frequencies and computing percentages with which observations fall into the various mutually exclusive classes. For example in a public opinion poll, the analysis of the results on a particular question amounts to counting the number of answers in the "yes", "no" and "no-opinion" classes and calculating the percentage in each class. (See any one of the Gallup Poll newspaper releases.) (There are, of course, other problems of cross-tabulation of analyses, such as finding the percentage of those answering "no" to question B who answered "yes" on question A, and so on.) If one has very large "properly drawn" samples of qualitative observations, the analysis essentially stops with counting and percentage analysis, and perhaps in presenting them graphically. In large "properly drawn" samples, the percentages calculated from the sample (the sample statistics) will be approximately the same as the percentages for the entire population (the population parameters) as one could check if one had the population available. But if the samples are small, one is then faced with the problem -- just as in the case of small samples of quantitative data -- of worrying about the accuracy with which one can estimate the population percentages by using the sample percentages.

Everyone is familiar with graphical presentations of percentages or frequencies calculated from qualitative data. Examples can be found in magazines, newspapers, posters, folders, information booklets, etc. There are two or three basic types of graphs or charts, of which there are many colored and pictorial variations. The first is the bar chart with horizontal or vertical bars, of which the following are examples:

Motor-Vehicle Traffic Fatalities (in 1000's) in the U.S. in 1943

(From Statistical Abstract of the U.S., 1944-45)



Ordinary Life Insurance Death Benefits in the U.S. in 1946

(From The Institute of Life Insurance)

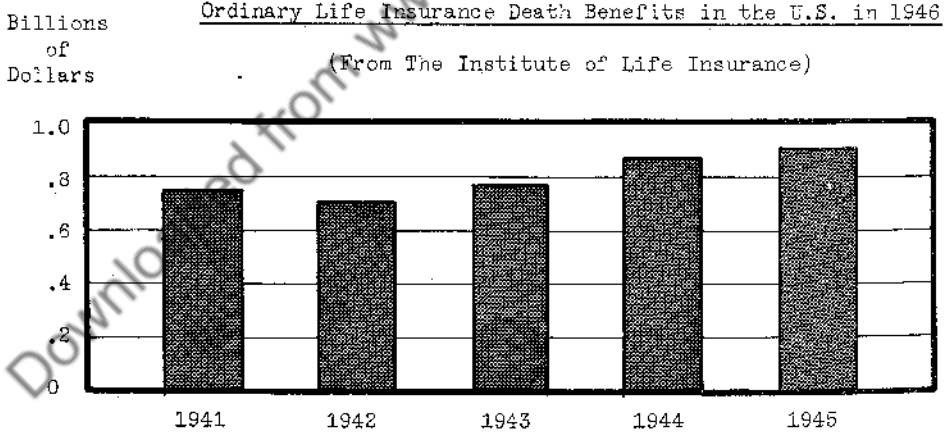
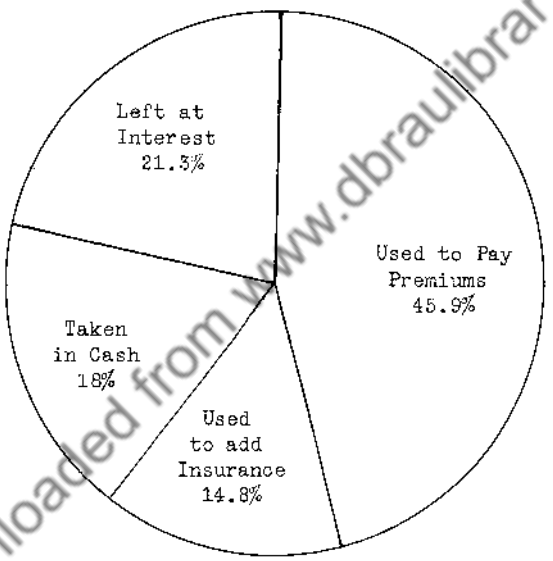


Figure 1.1

In many popular presentations these monotonous black bars are replaced by rows of men, autos, piles of dollars, or other symbols suggesting the subject matter which is being described. Often bar charts are used for presenting percentages or totals for a series of years -- one bar being used for each year.

There are many elaborations of bar charts from which one can make easy comparisons of numbers in each of several categories for two or more years. For example if one had the percentages of death by various major causes for New Jersey for 1946 and 1947, one could construct a bar chart in which there is a pair of bars side by side for each cause of death, one bar for 1946 and another for 1947.

The second type of chart is the familiar pie chart which is particularly useful in showing how the total or 100% of anything is divided up into certain classes. For example, here is a typical pie chart (without the coloring or cross-hatching):



How the 1945 life insurance dividends (dollars) were used by policyholders in 1945

(Institute of Life Insurance Data)

Figure 1.2

Exercise 1.

(These questions are listed mainly to provoke discussion - oral and written).

1. If, in an opinion survey, you were asked to select a sample of 100 Princeton undergraduates who are sons of Princeton alumni, how would you select the sample? What is the population sampled? If these 100 undergraduates are mailed a questionnaire and only 60 undergraduates fill them out and return them, what population would be sampled?

2. Suppose you were asked to undertake a study of electric current consumption in private homes of Princeton during December, 1947, on the basis of a sample of 300 households. How would you proceed with the selection of the sample of addresses for such purpose? What is the population sampled?

3. In investigating an allegedly biased six-sided die, suggest a procedure for getting a sample of numbers you would need. What would be the population for this die?

4. In studying the burning life of a new type of 40-watt bulb being made in small quantities for experimental purposes, a suitable sample of measurements would consist of what? What is the population in such a study?

5. Consider the washers being made by an automatic machine for a certain kind of precision instrument. What is the population of washers? How would you select a sample of washers?

6. Indicate how you would undertake to get a sample of 200 sentence lengths in studying sentence length used by Margaret Mitchell in Gone With the Wind. What is the population here?

7. Describe briefly how a radio audience researcher for Station WOR, investigating the amount of WOR day-time listening in Trenton, might select a sample of 500 homes from telephone subscribers in Trenton. What is the population in this example? If the investigator calls these 500 homes by telephone and gets a response from only 400 of them, what population is actually being sampled?

8. In studying the tensile (breaking) strength of 12-gauge aluminum fence wire being turned out continuously at a factory and cut into 1000-foot lengths (plus

a few feet) for coiling, suggest a practical procedure for getting a sample of measurements which you could use. What would be the population?

9. The Fish and Game Commission of a certain state wants a sample of 1000 of its population of licensed hunters to fill out a post card questionnaire. It has dozens of books of license stubs from agents all over the state giving names and addresses of the hunters. What would you consider to be a satisfactory method of drawing a sample of names of licensed hunters? If only 750 of the hunters returned the filled-out post cards, what population would be sampled? How do you think this population would compare with the entire population of licensed hunters of the state?

10. A small radio transmitter is designed so it may be used to generate a certain automatic signal. In making a detailed study of the length of this signal and how it varies for a single specified unit, how would you draw a sample of signals from a given unit? What is the population in this case?

11. Suppose you had 2500 ballots filled out in a public opinion poll on a sample of 2500 voters in City X. In the ballot there is an item consisting of a list of 6 potential presidential candidates, and each respondent is asked to check one name among them whom he would like to see as president. How would you condense the results on this item for the 2500 ballots and present the results? Suppose in a second item the respondent checks whether he is a Republican, Democrat or Other. How would you present the presidential choice data so as to show how it varies with political affiliation? What is the population in this example?

12. If you are asked to find out from a sample of cars the extent to which Princeton car owners use the various brands of automobile tires, how would you proceed to collect the data? How would you condense it and exhibit it when you got it? What is the population of tires for the procedure you would use?

13. The percentage of life insurance policies sold by United States companies in 1944 in each of the following categories: Whole life, Limited payment life, Endowment, All Other, were 22, 36, 21, 21, respectively. The percentages for 1942 were 26, 36, 16, 22. Present this material graphically so it is easy to make comparisons within each category for the two years. In this example, what is the population? What is the sample?

14. If a family should keep an accurate record of the disposition of its incomes for a year, and should find the number of dollars for each of the categories: food, clothing, rent, entertainment, savings, all other, how would you present the data graphically? If it were collected for three successive years, how would you present it graphically, so as to make easy year-to-year comparisons within each category?

CHAPTER 2. FREQUENCY DISTRIBUTIONS

2.1 Frequency Distributions for Ungrouped Measurements -- an example.

Raw statistical data as pointed out in Section 1.2, usually consists of a series of readings or measurements. As an example, we shall take the weights (to the nearest .01 ounce) of zinc coating of 75 galvanized iron sheets of a given size as given in Table 2.1 (from A.S.T.M. Manual on Presentation of Data, American Society for Testing Materials, 1947, p. 4):

TABLE 2.1

Weights (in ounces) of Zinc Coatings of 75 Galvanized Iron Sheets

1.47	1.60	1.58	1.56	1.44
1.62	1.60	1.58	1.39	1.35
1.52	1.38	1.32	1.65	1.53
1.77	1.73	1.62	1.62	1.38
1.55	1.70	1.47	1.53	1.46
1.53	1.60	1.42	1.47	1.44
1.38	1.60	1.45	1.34	1.47
1.37	1.48	1.34	1.58	1.43
1.64	1.51	1.44	1.49	1.64
1.46	1.53	1.56	1.56	1.50
1.63	1.59	1.48	1.54	1.61
1.54	1.50	1.48	1.57	1.42
1.53	1.60	1.55	1.67	1.57
1.34	1.54	1.64	1.47	1.75
1.60	1.57	1.57	1.63	1.47

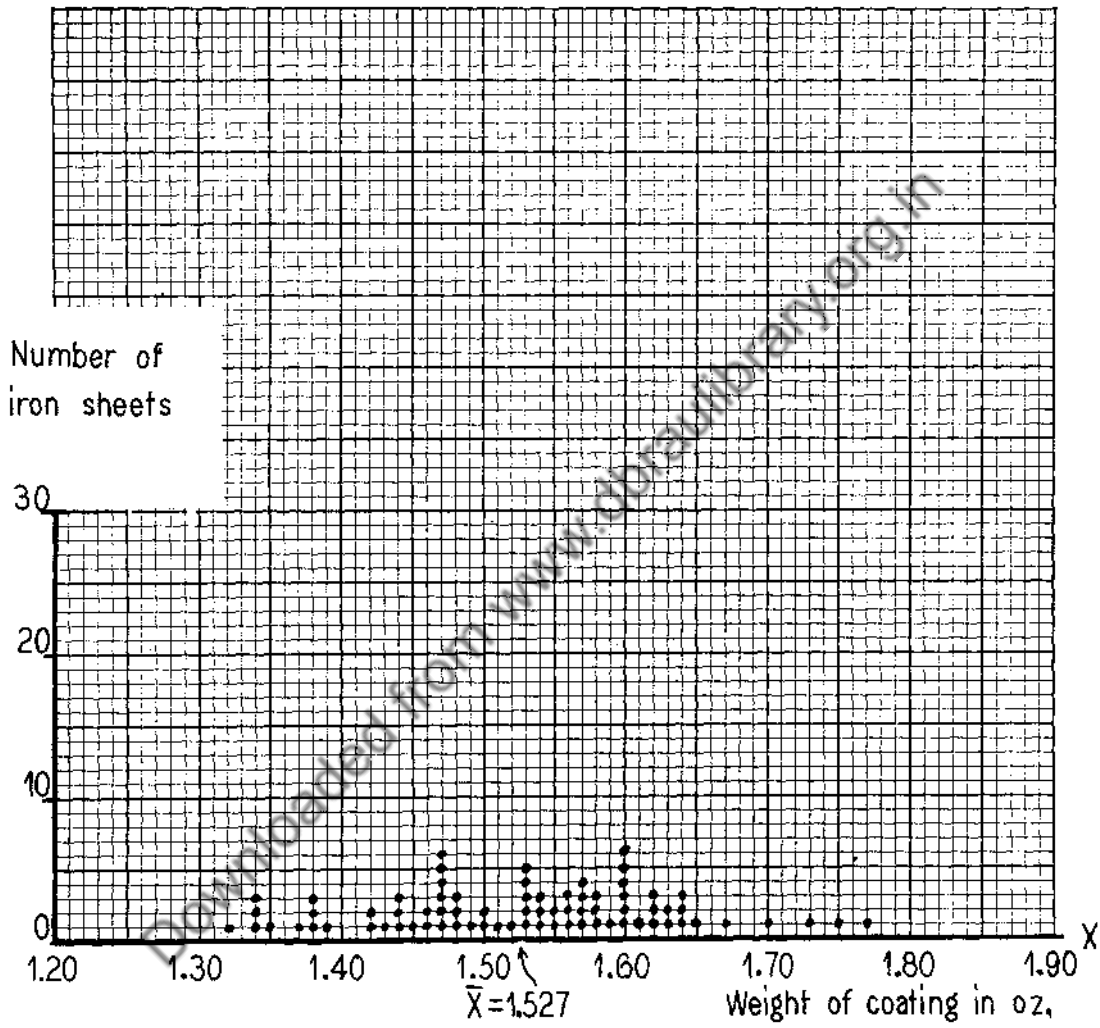
These 75 zinc coating weights were measured on a sample of small iron sheets of the same size by a chemical technique. The 75 measurements were a sample of chemical determinations from a (theoretically) indefinitely large population of chemical determinations which might have been made from galvanized iron sheets at that time. Just by looking at the 75 measurements themselves, one cannot tell whether the variation from 1.32 ounces to 1.77 ounces is due mainly to variations in the weights of zinc actually deposited on the iron sheets or to variations in chemical technique, or both. This question would have to be settled by an elaborate experiment. This kind of question always arises in

connection with measurements, although our common knowledge can sometimes answer it for us. But we shall proceed as though the variations in the measurements are due mainly to variations in the weights of the actual coatings.

The 75 measurements in Table 2.1 may be considered therefore as a sample from an indefinitely large population of measurements that might have been taken. Thus, if we had taken a further sample of 75 sheets we would have obtained another set of 75 numbers, and so on for any number of samples of 75 we can imagine as having been taken. We will consider this sampling problem later in the course. Our job at present is to describe the information furnished by the sample of 75 numbers in Table 2.1.

A simple graphical representation of these 75 numbers is given by the dot frequency diagram shown in Figure 2.1, in which each dot represents an observation. The graphical display in Figure 2.1, although giving a quick picture of the data and showing how it tends to "bunch up" in the middle, is ordinarily not as useful for descriptive purposes as a cumulative graph as shown in Figure 2.2, which can be readily plotted from the information shown in a dot frequency diagram. In Figure 2.2, the ordinate of the "step-like" cumulative graph for any given abscissa gives the frequency (or percent) of iron sheets having zinc coating weight less than or equal to that particular abscissa. The left-hand scale of ordinates gives cumulative frequency and the right-hand scale gives cumulative percent. As an example, we note that the ordinate at the abscissa 1.58 is 53, as read on the frequency scale, or 70.7, as read on the percent scale. This means that there are 53 iron sheets (or 70.7%) having zinc coat weights less than or equal to 1.58 ounces.

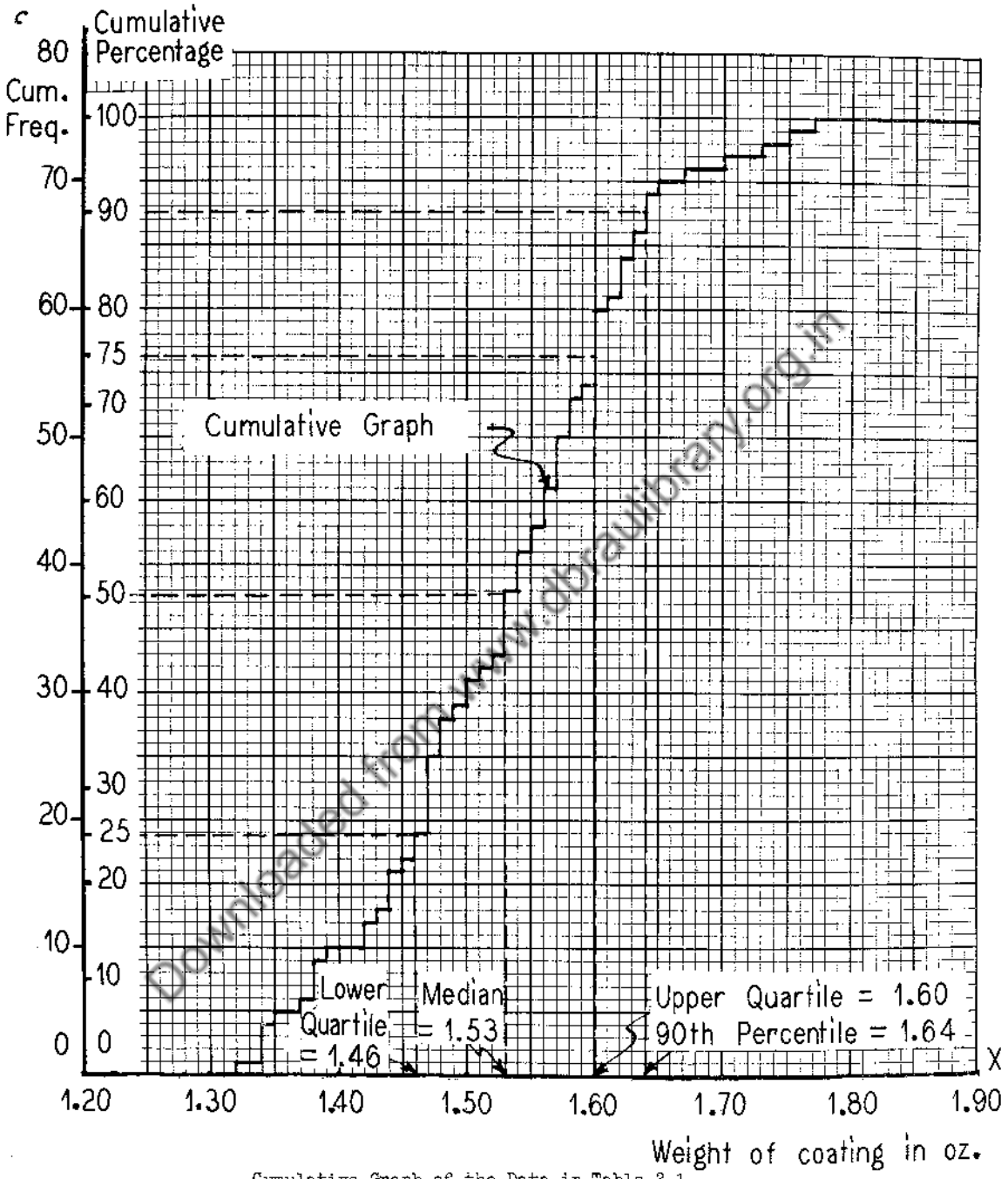
Note that the ordinate for any given abscissa at which a jump occurs, is to be extended to the top of the jump. For example, the ordinate corresponding to 1.58 is 53 and not 50. Conversely, one may find the abscissa corresponding to any given ordinate (read from either of the two scales). Strictly speaking, the only values of the ordinates at which abscissas are actually defined are those at which horizontal "steps" occur, and then the abscissa for that ordinate is the abscissa corresponding to the left-hand end of the "step". For example, the ordinate 10 (on the cumulative frequency scale) corresponds to the abscissa 1.39 (and not 1.42). However, if we take any value p on the percent scale then draw a horizontal line to the right until we strike the step-like graph (either the vertical dotted portions of the graph or the left-hand end of a "step") and then draw a straight line vertically downward until we strike the



Dot Frequency Diagram of the Measurements in Table 2.1

(The Point Marked $\bar{X}$ = 1.527 is the Mean of the Distribution and Will be Explained in Section 3)

Figure 2.1



Cumulative Graph of the Data in Table 2.1

Figure 2.2

horizontal axis, the point of intersection on the horizontal axis is called the p-th percentile. This means that approximately p percent of the sample measurements are less than or equal to the value of the p-th percentile. For example, the 90th percentile is 1.64. The actual number of cases less than or equal to 1.64 is 69 (or 92%), while the actual number less than 1.64 is 66 (or 88%). The desired percentage (90%) falls between these two percentages. The 50th percentile is called the median, and in this case is 1.53. Actually, the number of measurements less than or equal to the median is 38 (or 50.7%), while the number less than 1.53 is 33 (or 44%).

The 25th percentile is called the lower quartile and the 75th percentile is called the upper quartile. The difference between these two quartiles is called the inter-quartile range, which includes approximately 50% of the sample measurements. The lower and upper quartiles in Figure 2.2 are 1.46 and 1.60, respectively. The inter-quartile range is therefore 1.60 - 1.46 = .14.

The difference between the least and greatest measurements in the sample is called the range of the sample. In the case of the data in Table 2.1 and Figure 2.1, the range is 1.77 - 1.32 = .45.

Exercise 2.1.

(In these problems use graph paper ruled with 10 divisions per inch).

1. A class of twenty students made the following grades on a mid-term test: 30, 26, 31, 20, 33, 40, 7, 36, 28, 15, 18, 24, 22, 21, 28, 22, 25, 46, 29, 27. Make a dot frequency diagram and a cumulative graph of these grades. Determine the range, upper and lower quartiles, inter-quartile range and median. Indicate the quartiles and median on the cumulative graph.

2. The following table gives the Vickers Hardness numbers of 20 shell cases (Pedlar Data):

66.3	61.3	62.7	60.4	60.2
64.5	66.5	62.9	61.5	67.8
65.0	62.7	62.2	64.8	65.8
62.2	67.5	67.5	60.9	63.8

Make a dot frequency diagram and a cumulative graph of these numbers. Determine the range, upper and lower quartiles, inter-quartile range and median. Indicate the quartiles and median on the cumulative graph.

3. The number of words per sentence in 60 sentences taken from a certain section of Toynbee's A Study of History were as follows:

24	44	26	39	34
39	54	28	73	96
46	80	25	26	21
22	35	7	42	34
51	36	17	41	55
20	23	22	11	36
48	15	27	44	16
58	21	70	50	40
39	43	42	20	35
60	18	12	69	40
28	12	15	20	43
19	19	65	41	66

Make a dot frequency diagram and a cumulative graph of these numbers. Determine the range, upper and lower quartiles, inter-quartile range and median. Indicate the quartiles and median on the cumulative graph.

4. The following table (from Grant) gives the yield point (in units of 1000 lb/sq. inch) for each of 40 steel castings:

64.5	67.5	67.5	64.5
66.5	73.0	68.0	75.0
68.5	71.0	67.0	69.0
68.0	69.5	72.0	71.0
69.5	72.0	71.0	68.5
66.5	67.5	69.0	68.0
65.0	63.5	65.5	65.0
70.0	68.5	68.5	70.5
64.5	67.0	66.0	63.5
62.0	70.0	71.0	68.5

Make a dot frequency diagram and a cumulative graph of these numbers. Determine the range, upper and lower quartiles, inter-quartile range and median. Indicate the quartiles and median on the cumulative graph.

5. Throw 5 dice 40 times and record the total number of points on each throw. Make a dot frequency diagram of the results, and also a cumulative graph. Also determine the range, median, lower and upper quartiles, and the inter-quartile range. (For the purposes of this problem you can consider 5 throws of one die equivalent to one throw of five dice in case you do not have five dice!)

6. Throw ten pennies 50 times and record the number of heads each time. Make a dot frequency diagram and a cumulative graph of the results. Also determine

the range, the median, the lower and upper quartiles and the inter-quartile range.

7. Shuffle a pack of cards thoroughly and deal off a hand of 13 cards. Record the number of honor cards. Return the cards to the pack and repeat. Do this 40 times. Make a dot frequency diagram and a cumulative graph. Also determine the range, median, lower and upper quartiles and the inter-quartile range.

2.2 Frequency Distributions for Grouped Measurements -- an example.

If there are more than about 25 observations in the sample of data, the construction of dot frequency diagrams or cumulative graphs for ungrouped data often involves more detail than is usually needed for practical purposes. One does not distort the pertinent numerical information provided by the data to amount to anything from a practical point of view if the data are grouped. In this case one can use grouped frequency distributions, and cumulative grouped frequency distributions. To define these grouped distributions we first construct a frequency table. Returning to Table 2.1 (or looking at Figure 2.1) we find the least value in the table to be 1.32 and the largest value to be 1.77. The range is .45. We now divide the range into a number of equal intervals of convenient length. This means that the length should be a "round number". The number of intervals is usually taken to be between 10 and 25. A convenient interval for our example is 0.05, which gives us 10 class intervals or cells. We might also have used 0.04, 0.03, or 0.02, but we would usually avoid 0.0333, 0.035, and other such inconvenient numbers.

We now take the cells to be 1.275 - 1.325, 1.325 - 1.375, 1.375 - 1.425, and so on to 1.725 - 1.775. Note the following features of these cells: (a) each is of length 0.05, (b) the boundaries of each cell end in a 5 and are written with one more decimal than is used in the original data of Table 2.1, (c) the upper boundary of any cell is the same as the lower boundary of the succeeding cell (this is a convenience and will cause no ambiguity since the boundaries are written to one more decimal than is used in the original data in Table 2.1).

The cells are constructed so they will have the following simple midpoints respectively: 1.30, 1.35, 1.40, and so on to 1.75. (We can have all these nice properties since we took a round number for the cell length.)

The frequency table can be exhibited as Table 2.2. The cell boundari are shown in column (a), the midpoints in column (b). The tallied frequencies with which the observations fall into the various cells are shown in column (c) The frequencies are shown in column (d). The relative frequencies in column (e the cumulative frequencies in column (f), and the cumulative relative frequenci in column (g). The last two columns are associated with the upper cell boundar. ies, and their entries are sometimes dropped a half line in the table.

What we are really doing in this grouping procedure is to arbitrarily assign the one measurement in Table 2.1 falling in the cell 1.275 - 1.325 the value 1.30, arbitrarily assign the five measurements in Table 2.1 falling in the cell 1.325 - 1.375 the value 1.35, and so on.

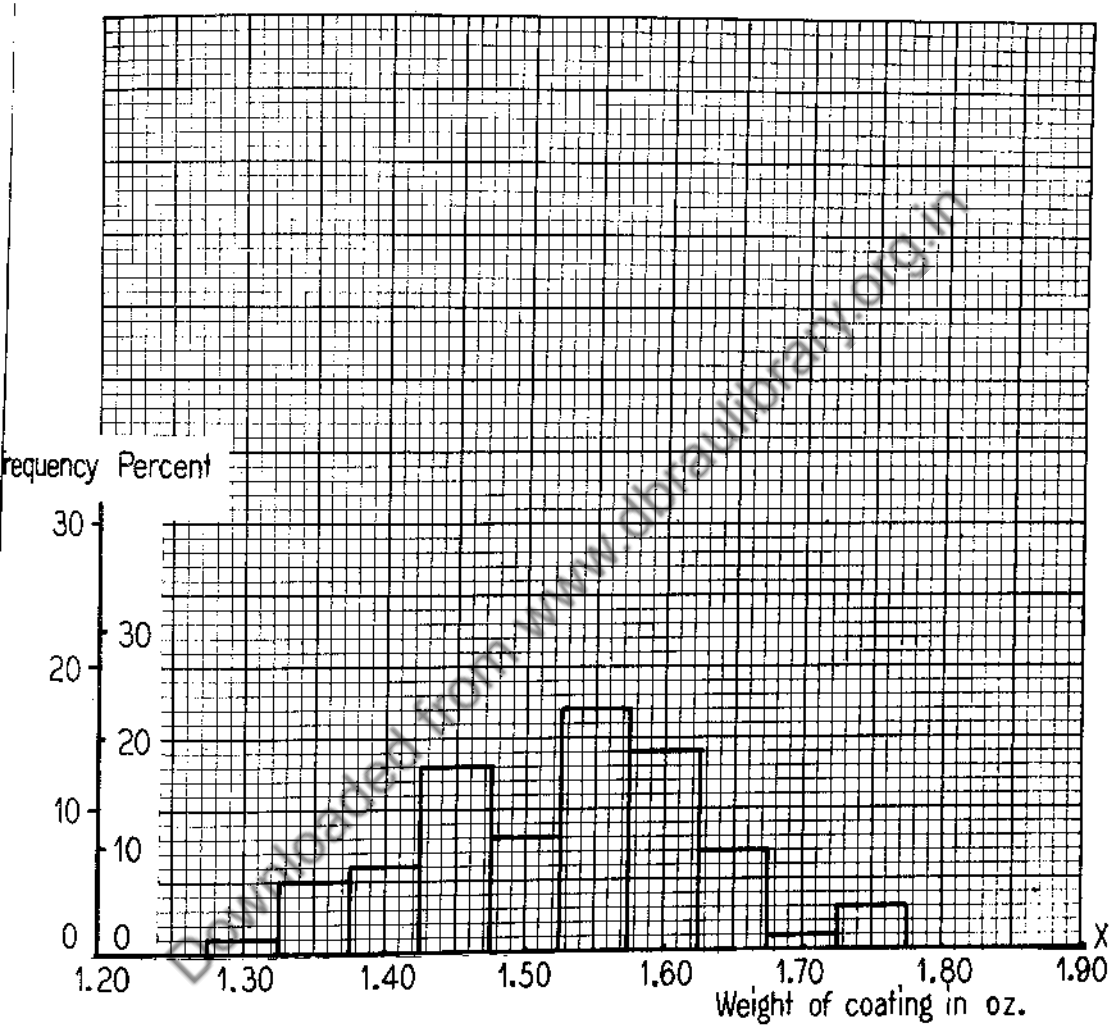
TABLE 2.2

Frequency Distribution for Grouped Measurements of Weights of 75 Zinc Coatings

(a) Cell Boundaries	(b) Cell Midpoints	(c) Tallied Frequency	(d) Frequency	(e) Relative Frequency	(f) Cumulative Frequency	(g) Cumulati Relativ Frequenc
1.275-1.325	1.30	1	1	.013	1	.013
1.325-1.375	1.35	𝍸	5	.067	6	.080
1.375-1.425	1.40	𝍸 1	6	.080	12	.160
1.425-1.475	1.45	𝍸 𝍸 111	13	.173	25	.333
1.475-1.525	1.50	𝍸 111	8	.107	33	.440
1.525-1.575	1.55	𝍸 𝍸 𝍸 11	17	.227	50	.667
1.575-1.625	1.60	𝍸 𝍸 1111	14	.187	64	.854
1.625-1.675	1.65	𝍸 11	7	.093	71	.947
1.675-1.725	1.70	1	1	.013	72	.960
1.725-1.775	1.75	111	3	.040	75	1.000

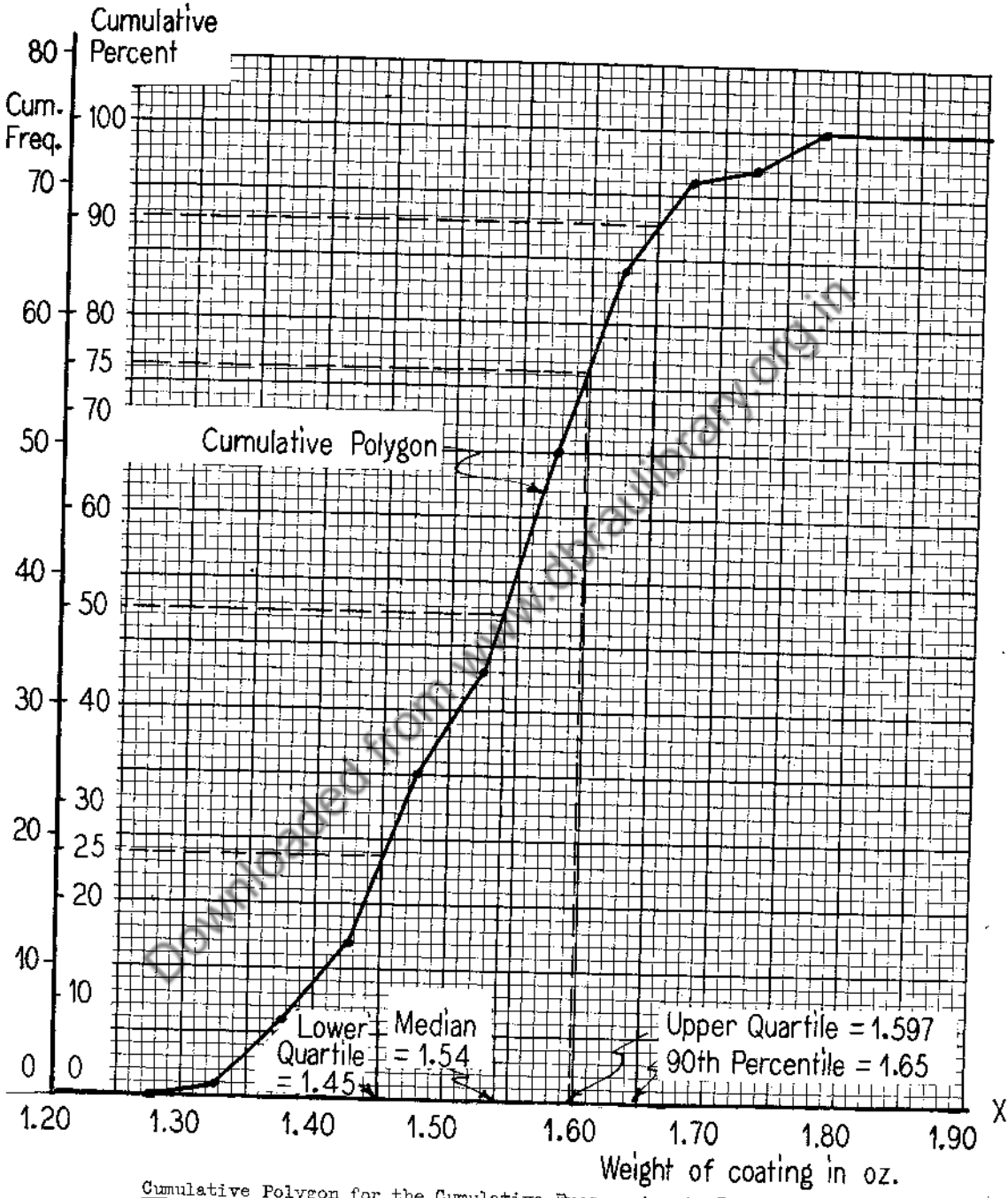
The frequencies [columns (d) and (e)] in Table 2.2 can be represented graphically as a frequency histogram as shown in Figure 2.3. Note that two scale are provided for the ordinates -- one scale refers to frequency the other to rela tive frequency expressed in terms of percent.

A more useful graphical representation of the material in Table 2.2 is given by a cumulative polygon for grouped data as shown by the heavy graph in Figure 2.4. This graph together with the two scales of ordinates is the graphica



Frequency Histogram of Frequencies of Table 2.2

Figure 2.3



Cumulative Polygon for the Cumulative Frequencies in Table 2.2

Figure 2.4

representation of columns (f) and (g) of Table 2.2. In this graph the points are plotted above the upper cell boundaries and not above the cell midpoints.

The cumulative polygon provides a simple and quick graphical procedure for approximately determining percentiles without going to the degree of detail involved in using the cumulative graph in the ungrouped case (Figure 2.2). For example, the 90th percentile is 1.65 ounces. (See dotted lines in Figure 2.4.) This means that approximately 90 percent of the observations have values less than or equal to 1.65. (The actual number of such observations less than or equal to 1.65 ounces is 70 or 93.3%, as will be seen from Table 2.1 and Figure 2.1. The actual number less than 1.65 is 69 (or 92%), so that our grouped percentile does not necessarily have the bracketing characteristic as in the ungrouped percentile. In other words, 90% does not lie between 92% and 93.3%.) In general, the larger the number of observations, and the smaller the cells, the more accurate these percentile approximations will be.

It will be seen from Figure 2.4 that the median is about 1.54, the upper quartile 1.597 and the lower quartile 1.450. Note that the values of these quartiles as determined from the graph in Figure 2.4 are slightly different from the values found from the graph in Figure 2.1. This is due to the effect of grouping.

We have been talking about grouped and ungrouped data -- yet the original data itself could be considered as grouped with cell length equal to .01 if the original measurements in Table 2.1 had been given to three or more decimal places instead of two. The question of deciding how many figures or decimals to keep in a set of measurements arises in most measurement problems, and has to be settled in each case. In the present problem, it may be considered doubtful whether the zinc coating measurements would really have any significance if carried to three or more decimal places. Or even if they did have significance there is such a wide variation of weights from one iron sheet to another that it may be considered as not worthwhile to have weights measured more accurately than to two decimals.

If, therefore, we should consider the two-decimal measurements as grouped from measurements to three or more decimals we could construct a cumulative polygon for cell length of .01 just as we have done for cell length of .05 (Figure 2.4). Actually, the frequency polygon for a grouping with cell length of .01, the cells being centered at 1.32, 1.33, 1.34, and so on to 1.77, could be constructed from the cumulative graph in Figure 2.2 as follows: Take the

midpoint of each "unit part" (.01 inch) on each horizontal portion of the step-like graph in Figure 2.2. The first point will be on the axis of abscissas at 1.315, and the last one will be at the top (100% point) of the ordinate erected for the abscissa 1.775. This cumulative polygon would be simply obtained then from the cumulative graph of Figure 2.2 by trimming off the corners. The general course of the cumulative polygon constructed from the cumulative graph in Figure 2.2 and that of Figure 2.4, as one would expect, is similar except that the cumulative polygon in Figure 2.4 is "smoother".

We can determine from a cumulative polygon approximately the number of cases in the sample lying between two values of the abscissa. For example, suppose we are interested in the number of cases in the sample having zinc coating weight between 1.42 and 1.68 ounces. 1.68 is the 94.8 percentile, and 1.42 is the 14.7 percentile. 94.8 - 14.7 = 80.1%. This difference is approximately the percentage of cases having zinc coatings between 1.42 and 1.68 ounces. The actual number of cases is 59 (or 78.7%).

Exercise 2.2.

(Each student is expected to do at least one of problems No. 8, 9, 10, 11, 12, 13, 14 and to keep a record of the data in the order in which it was obtained -- also a record of all calculations. The record will be needed in future problems.)

1. Each of 300 measurements is given in inches to three decimal places. The largest measurement is 2.062" and the smallest is 1.997". Make up an outline of a frequency table showing cell boundaries and cell midpoints you would use.

2. Measurements on the crushing strength of 270 bricks (in pounds per square inch), are given to the nearest 10 pounds. The largest measurement is 2070 and the smallest is 270. Set up an outline of a frequency table showing cell boundaries and cell midpoints.

3. Suppose the cell midpoints for a given frequency distribution are 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160. Make up an outline of a frequency table showing cell boundaries.

4. Make up an outline of a frequency table you would use in presenting the

heights of all Princeton men of the Class of 1952, if the heights were available to the nearest quarter of an inch.

5. Find the 10th, 40th, 60th and 80th percentiles graphically from Figure 2.4, and compare them with the 10th, 40th, 60th and 80th percentiles as determined graphically from Figure 2.2. Working from Figure 2.1, find the actual percentages of cases less than or equal to each percentile in the two cases. Also find the percentages of cases less than each percentile in the two cases.

Instructions for problems 6 - 14. In each of the problems Nos. 6 to 14 the following operations are to be carried out:

(a) Make frequency table, showing frequency and cumulative frequency distributions, relative frequency and relative cumulative frequency distributions.

(b) Construct a frequency histogram.

(c) Construct a cumulative polygon and find the two quartiles, the median and the inter-quartile range from the graph.

6. In the following table are given the scholastic aptitude scores of the 66 departmental students of a certain department in the Class of 1938:

345	530	556	354	593	574
395	516	479	494	417	494
563	444	629	439	486	560
505	604	490	446	604	464
402	406	730	505	515	549
472	475	611	585	523	541
691	523	468	468	545	468
624	582	574	578	505	629
523	575	420	605	527	607
461	439	596	417	384	490
490	523	585	585	431	549

7. The following table gives 125 observations of a spectral line (Birge Data), where only the last two digits of the reading are recorded. For example, the first reading is actually 65.177 mm, of which only the 77 is recorded.

77	74	73	84	77
78	85	80	81	80
75	69	72	83	79
75	80	79	74	78
70	74	83	72	79
73	81	87	82	79
78	79	78	74	85
83	79	83	81	84
81	88	79	80	78
77	80	85	80	78
72	75	73	85	79
78	82	80	76	76
79	75	83	81	78
82	76	78	78	79
86	79	79	84	74
76	75	77	82	77
79	77	72	77	81
83	75	82	90	77
80	78	83	81	74
79	80	79	75	84
81	74	73	74	86
77	82	75	74	75
75	74	83	76	84
72	84	73	77	77
76	75	81	79	74

See instructions preceding Problem 6.

8. Roll 5 dice and record the total number of dots that appear. Repeat this 50 times. See instructions preceding Problem 6.

9. Shuffle a deck of ordinary playing cards, deal off two hands of 13 cards each (one is yours and the other belongs to your partner) and record the total number of honor cards in the two hands. Repeat this 50 times. See instructions preceding Problem 6.

10. Take 25 pennies (or any other kind of coin), put them in an empty pocket, jingle them thoroughly, take them out, spread them on a table, count the number of heads and record the result. Repeat these operations 60 times. See instructions preceding Problem 6.

11. Take 50 thumbtacks, shake them up, throw them on a table, count the number of tacks that fall point up, and record the result. Repeat this 75 times. See

instructions preceding Problem 6.

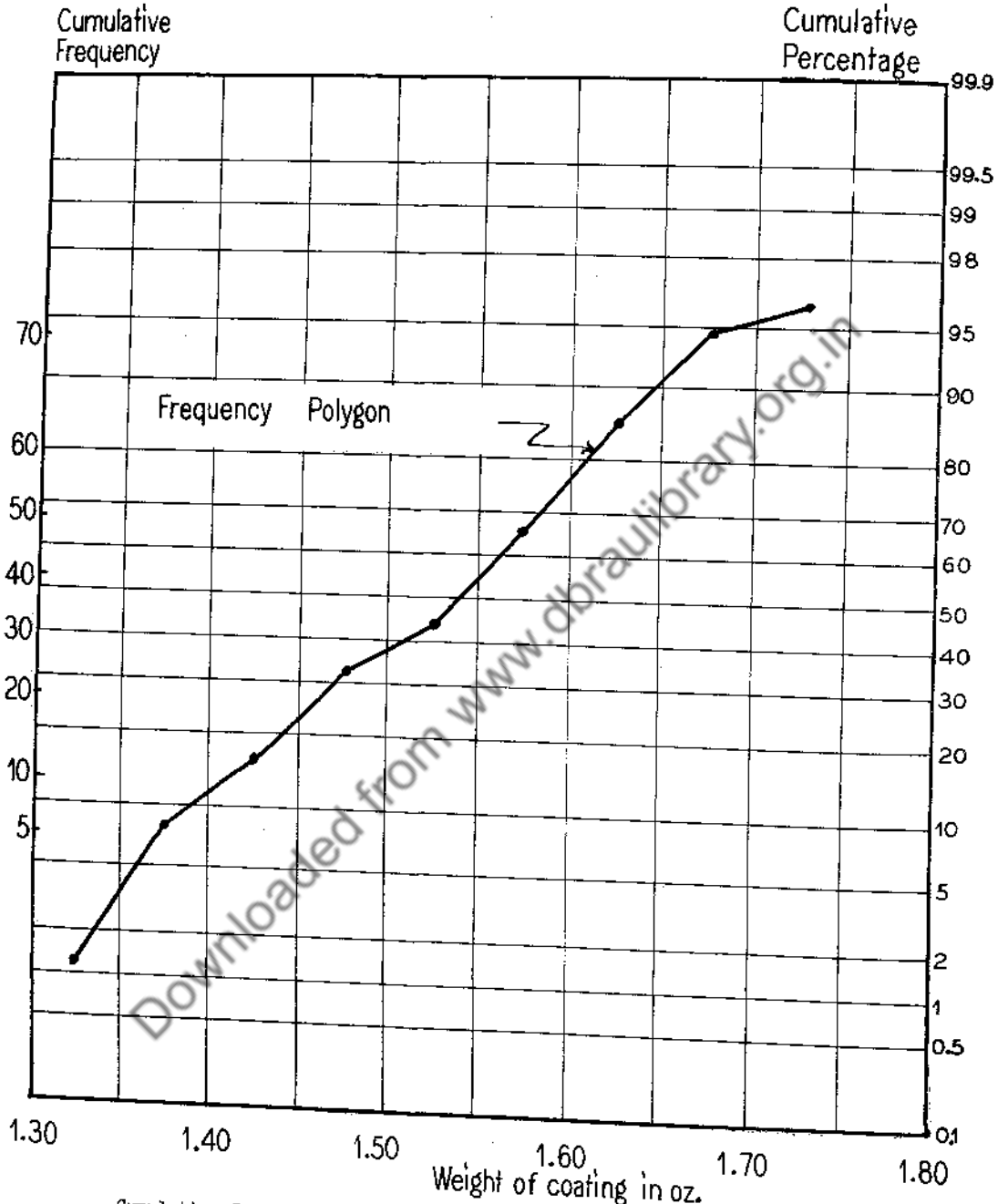
12. Pick up any book (containing no formulas!), open the book at random, count the number of e's on each full line on the 2-page spread and record the number for each line. See instructions preceding Problem 6.

13. Take any mathematical table such as a table of square roots, logarithms or trigonometric functions, in which the entries are blocked off in sets of five. Start with any block in the table you please, note the last digit in each of the five numbers in the block, and add these five digits together. Record the sum (which will be one of the 46 numbers 0, 1, 2, 3, ..., 45). Repeat this for the next block, the next, etc., until you have 60 blocks. See instructions preceding Problem 6.

14. Get 50 metal-rimmed price tags about 1 inch in diameter at any stationery store. Mark 10 tags with the number 3 on each side, 10 tags with the number 4, 10 with the number 5, 10 with the number 6, and 10 with the number 7. Stir up this population of fifty tags thoroughly in a bowl or an empty coat pocket. Draw out a sample of 5 tags, and find the mean of the numbers on the five tags. Put the 5 tags back in the population and draw another sample of 5. Repeat this 80 times. See instructions preceding Problem 6.

2.3 Cumulative Polygons Graphed on Probability Paper.

You have now plotted enough cumulative polygons to realize that they are usually steeper in the middle than at the ends. This is a very general characteristic of cumulative polygons. It is sometimes convenient to plot them on a special kind of graph paper called probability graph paper so that they become approximately straight lines. This is accomplished by stretching the percentage scale for low percentages and for high percentages. If we plot the cumulative frequency polygon shown in Figure 2.4, we obtain the graph shown in Figure 2.5 which is much more nearly a straight line. Cumulative polygons plotted on probability graph paper will be discussed in greater detail in Section 8.3.



Cumulative Polygon of Figure 2.4 Plotted on Probability Graph Paper

Figure 2.5

2.4 Frequency Distributions -- General.

The general aspects of the procedures exemplified in Section 2.1 and 2.2 should be noted. In general, we start off with a set of n measurements of some kind. We may call these measurements X_1, X_2, X_3, ..., X_n and in practice they would be displayed in a table similar to Table 2.1. In Table 2.1, for example, n = 75 and we could take $X_1 = 1.47$, $X_2 = 1.62$, ..., $X_{75} = 1.47$.

In their ungrouped form, these measurements may be represented graphically by n dots in a dot frequency diagram exemplified in Figure 2.1. Here each measurement in the set of n measurements is represented by a dot placed above the value of that particular measurement wherever it occurs along the X-axis. The dot frequency diagram gives a convenient pictorial arrangement of the X's ranged from least to greatest, ties being indicated by dots placed in vertical columns of 2 or more dots.

These n measurements can also be graphically represented, in their ungrouped form, by a cumulative graph. In such a graph, the ordinate erected at any given abscissa simply represents the number of the measurements (i.e., the number of X's among the set X_1, X_2, ..., X_n) which are less than or equal to that given abscissa. This graph can be constructed immediately from a dot frequency diagram, since the n X's are arranged in order from least to greatest in that graph.

The handling of the n individual measurements in ungrouped form becomes too detailed and laborious for practical purposes if the total number of different numerical values actually taken on by the n measurements is very large (more than about 25 for practical purposes). This amounts to saying that working with ungrouped data involves too much detail if the dot frequency diagram has more than about 25 vertical columns of dots (each column containing one or more dots).

Suppose the dot frequency diagram contains more than about 25 columns of dots (and this can be determined easily by looking at the n measurements and seeing how many different numerical values are to be found among them). We then proceed to group the data to simplify matters. To do this, we look through the n measurements and find the least and greatest values. We take the range (the difference between the least and greatest values) and divide it into some number of equal intervals, making sure that the interval length actually chosen is a simple one in terms of the original units of measurement. This usually means that it will not involve a lot of awkward decimals. When there are a very few measurements far out on one or both ends of the distribution, it

will often pay to use a smaller cell length and hence more cells. We then proceed to cut the X-axis up into cells, each cell being an interval of the length chosen, and each cell having a simple and convenient midpoint. Once we have decided the cell length and the midpoints of the cells, we then proceed to arbitrarily assign every measurement falling in a given cell a value equal to the value of X at the midpoint of the cell. The cell boundaries for a given cell are placed one-half of the cell length on each side of the cell midpoint.

We shall use the following notation:

k is number of cells

c is length of each cell

$x_1, x_2, \ldots, x_k$ are the midpoints of the first, second, ..., k-th cell as counted from left to right.

$f_1, f_2, \ldots, f_k$ are the numbers or frequencies of the measurements $X_1, X_2, \ldots, X_n$ in the first, second, ..., k-th cells respectively.

$F_1 = f_1$, $F_2 = f_1 + f_2$, $F_3 = f_1 + f_2 + f_3, \ldots, F_k = f_1 + f_2 + \ldots + f_k = n$, are the cumulative frequencies associated with the upper cell boundaries for the first, second, ..., k-th cells respectively.

$x_1 \pm \frac{1}{2}c$, $x_2 \pm \frac{1}{2}c$, ..., $x_k \pm \frac{1}{2}c$ are the boundaries of the first, second, ..., k-th cells. In actual practice the boundaries are given to one more decimal than the original measurements $X_1, X_2, \ldots, X_n$, and they end in 5. This makes it possible to use the left-hand boundary of a cell as the right-hand boundary of the preceding cell without ambiguity as to which cell a given measurement belongs. These symbols are represented in Figure 2.6.

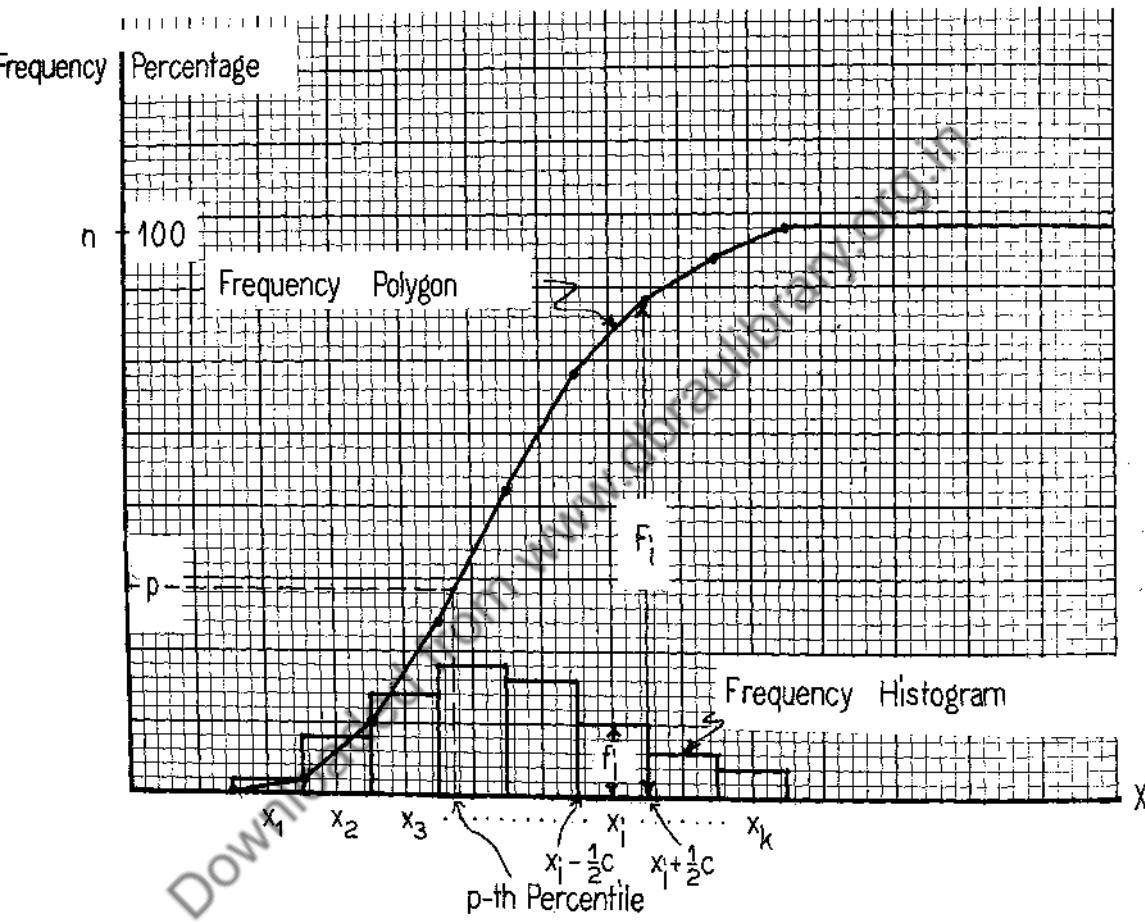
In this grouping operation, what we are really doing is this: every one of the f_1 measurements in the sample of measurements $X_1, X_2, \ldots, X_n$ which falls in the cell $x_1 \pm \frac{1}{2}c$ is arbitrarily given the value x_1, every one of the f_2 measurements which fall in $x_2 \pm \frac{1}{2}c$ is arbitrarily given the value x_2, and so on for all of the cells.

These symbols can be arranged in the form of a general grouped frequency table as shown in Table 2.3.

TABLE 2.3

General Frequency Distribution of Grouped Measurements

(a) Cell No.	(b) Cell Boundaries	(c) Cell Mid-point	(d) Frequency	(e) Relative Frequency	(f) Cumulative Frequency	(g) Relative Cumulative Frequency
1	$(x_1 - \frac{1}{2}c) - (x_1 + \frac{1}{2}c)$	x_1	f_1	$\frac{f_1}{n}$	$f_1 = F_1$	$\frac{F_1}{n}$
2	$(x_2 - \frac{1}{2}c) - (x_2 + \frac{1}{2}c)$	x_2	f_2	$\frac{f_2}{n}$	$f_1 + f_2 = F_2$	$\frac{F_2}{n}$
3	$(x_3 - \frac{1}{2}c) - (x_3 + \frac{1}{2}c)$	x_3	f_3	$\frac{f_3}{n}$	$f_1 + f_2 + f_3 = F_3$	$\frac{F_3}{n}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	$(x_i - \frac{1}{2}c) - (x_i + \frac{1}{2}c)$	x_i	f_i	$\frac{f_i}{n}$	$f_1 + f_2 + \ldots + f_i = F_i$	$\frac{F_i}{n}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
k	$(x_k - \frac{1}{2}c) - (x_k + \frac{1}{2}c)$	x_k	f_k	$\frac{f_k}{n}$	$n = F_k$	1
Total			n	1		



General Frequency Histogram and Frequency Polygon

Figure 2.6

We have deliberately put in the notation for cell No. i because we shall frequently want to refer to a "typical" cell in the table and we can do this by talking about the i-th cell, its midpoint x_i, etc.

The frequency column (d) and cumulative frequency column (f) in Table 2.3 can be represented graphically as a frequency histogram and a cumulative polygon respectively as shown in Figure 2.6. The value of X corresponding to any given percent, say p, as determined by the cumulative frequency polygon is called the p-th percentile of the measurements.

In particular, the 50th percentile is the median, the 25th percentile the lower quartile, the 75th percentile the upper quartile and the difference between the upper and lower quartiles is the inter-quartile range.

Exercise 2.4

1. For each of the problems No. 6 to 14 of Exercise 2.2 which you did, plot the cumulative polygons on probability paper.

2. Express $F_{11} - F_5$, $F_k - F_1$, $F_i - F_{i+j}$, in terms of $f_1, f_2, \ldots, f_k$.

3. If the cell length is doubled, what, approximately, will happen to the entries in the frequency column of a frequency table?

4. What is the largest possible change which can happen to a measurement when changed from its original value to a cell midpoint? Illustrate when c = .05 and the measurements are made to two decimals.

5. Referring to Table 2.3, how many measurements lie between $x_i - \frac{1}{2}c$ and $x_j + \frac{1}{2}c$? Express your answer in terms of the capital F's. Also express it in terms of the lower case f's.

CHAPTER 3. SAMPLE MEAN AND STANDARD DEVIATION

3.1 Mean and Standard Deviation for the Case of Ungrouped Measurements.

In Sections 2.1 and 2.2 we have seen how a given sample of statistical measurements in both the ungrouped and grouped forms can be condensed into tables and graphs, and how information pertaining to percentiles can be obtained from the graphs. For instance, the 50th percentile or median is the "middle" of the distribution of measurements in a certain well-defined sense. The inter-quartile range is an indication of the "scatter" or "spread" of the measurements in a well-defined sense.

There are other important ways of describing the "middle" of the distribution and the "spread" of the distribution. In this section we shall discuss the arithmetic mean or simply the mean of the distribution of sample measurements as another description of the "middle" of the distribution, and the standard deviation of the measurements as another description of the "spread" of the distribution.

3.11 Definition of the mean of a sample (ungrouped).

As a simple example, suppose the weights (in pounds) of five students are 141, 136, 157, 143 and 138. The mean of this sample of five weights is the sum of the weights divided by 5, i.e.

$$\text{mean} = \frac{141 + 136 + 157 + 143 + 138}{5} = \frac{715}{5} = 143 \text{ lbs.}$$

In general, if $X_1, X_2, X_3, \ldots, X_n$ is a sample of n measurements, the sample mean $\bar{X}$ of the X's is defined by the following relation:

(3.1) $$n\bar{X} = X_1 + X_2 + \ldots + X_n.$$

We can write the sample sum $X_1 + X_2 + \ldots + X_n$ more compactly as $\sum_{j=1}^{n} X_j$ where $\sum$ is the Greek letter capital sigma (chosen to correspond to the first letter of the word "sum"). $\sum_{j=1}^{n} X_j$ is to be read: "the sum of X sub j from j = 1 to j = n". Hence, (3.1) can be written more compactly as

(3.2) $$n\bar{X} = \sum_{j=1}^{n} X_j ,$$

from which the formula for the mean $\bar{X}$ is written explicitly as

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \,. \tag{3.3}$$

We shall be using the sample sum $\sum_{j=1}^{n} X_j$ so often that it will be convenient to simply call it $S(X)$, read "sum of X" in which case we may write (3.2) more briefly as

$$n\,\bar{X} = S(X). \tag{3.2a}$$

and (3.3) more briefly as

$$\bar{X} = \frac{1}{n}\,S(X). \tag{3.3a}$$

If we refer to our example of 5 weights we would have $n = 5$, $X_1 = 141$, $X_2 = 136$, $X_3 = 157$, $X_4 = 143$, $X_5 = 138$, $\sum_{j=1}^{5} X_j = 715$ (or $S(X) = 715$) and applying formula (3.2) to the 5 weights would give $5(\bar{X}) = 715$ or the mean is $\bar{X} = \frac{715}{5} = 143$ lbs.

The mean of the sample of 75 measurements in Table 2.1 is given by

$$\begin{aligned} 75(\bar{X}) &= 1.47 + 1.62 + \ldots + 1.47 \\ &= 114.51 \end{aligned}$$

or

$$\bar{X} = 1.527 \text{ ounces.}$$

In other words, applying the formula (3.2) to the measurements in Table 2.1, gives $75(\bar{X}) = 114.51$, and applying (3.3) gives $\bar{X} = 1.527$.

Suppose we take the difference between each X and the mean $\bar{X}$. We have $X_1 - \bar{X}$, $X_2 - \bar{X}$, ..., $X_n - \bar{X}$. If we add these differences we get

$$\begin{aligned} &(X_1 - \bar{X}) + (X_2 - \bar{X}) + \ldots + (X_n - \bar{X}) \\ &= (X_1 + X_2 + \ldots + X_n) - n\bar{X} = 0 \end{aligned}$$

because of (3.1). Hence by using summation notation we have

$$\sum_{j=1}^{n} (X_j - \bar{X}) = 0 \,. \tag{3.4}$$

In other words, the sum of the differences between each measurement in a sample and the mean of all measurements in the sample is equal to zero.

Returning to our example of 5 weights, we note that the differences between the measurements and the mean are (141 - 143), (136 - 143), (157 - 143), (143 - 143) and (138 - 143) or - 2, - 7, + 14, 0, - 5 respectively, and that the

sum of these differences is zero.

3.12 Definition of the standard deviation of a sample (ungrouped).

Considering the example of the 5 weights again, suppose we square the difference between each measurement and the mean and add them. The standard deviation s of the 5 weights is given by the following relation:

$$(5-1)s^2 = (141-143)^2 + (136-143)^2 + (137-143)^2 + (143-143)^2 + (138-143)^2$$

or

$$4\,s^2 = 4^2 + 7^2 + 14^2 + 0^2 + 5^2 = 274.$$

From this we find

$$s^2 = 68.50$$

or

$$s = 8.27\ .$$

More generally, if $X_1, X_2, \ldots, X_n$ is a sample of n measurements, the standard deviation s_X of the sample is defined by

$$(3.5) \qquad (n-1)s_X^2 = (X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \ldots + (X_n-\overline{X})^2.$$

Using the summation notation this can be written more briefly as

$$(3.6) \qquad (n-1)s_X^2 = \sum_{j=1}^{n} (X_j-\overline{X})^2\ .$$

The quantity s_X^2, the square of the standard deviation s_X, is called the variance of the sample. We shall not rewrite (3.5), (3.6) or any similar formula so as to give an explicit formula for the standard deviation s_X. For we can perfectly well talk about the standard deviation s_X given by (3.6) or the variance s_X^2 given by (3.6) without having to write down two formulas.

From the point of view of computation, a formula which is often more convenient when a calculating machine is available can be found from (3.5). For, by squaring each term on the right-hand side of (3.5) we have

$$(3.7) \qquad (n-1)s_X^2 = (X_1^2 - 2X_1\overline{X} + \overline{X}^2) + (X_2^2 - 2X_2\overline{X} + \overline{X}^2) + \ldots + (X_n^2 - 2X_n\overline{X} + \overline{X}^2)$$

or collecting terms

$$(3.8) \qquad (n-1)s_X^2 = (X_1^2 + X_2^2 + \ldots + X_n^2) - 2\,\overline{X}\,(X_1 + X_2 + \ldots + X_n) + n\,\overline{X}^2.$$

But from formula (3.1) it is seen that

$$n\,\overline{X} = (X_1 + X_2 + \ldots + X_n)$$

which, when used in the right-hand side of (3.8) gives

$$(3.9) \qquad (n-1)\, s_X^2 = (X_1^2 + X_2^2 + \ldots + X_n^2) - 2n\bar{X}^2 + n\bar{X}^2.$$

Using summation notation

$$(3.10) \qquad (n-1)\, s_X^2 = \sum_{j=1}^{n} X_j^2 - n\bar{X}^2$$

which is the desired formula. In practice, it is convenient to delay the division by n in calculating $\bar{X}$ and to calculate s_X^2 from the formula

$$(3.11) \qquad (n-1)\, s_X^2 = \sum_{j=1}^{n} X_j^2 - \frac{1}{n}\left[\sum_{j=1}^{n} X_j\right]^2$$

which may be written still more briefly as

$$(3.11a) \qquad (n-1)\, s_X^2 = S(X^2) - \frac{1}{n}\,[\,S(X)\,]^2.$$

It should be noticed that s_X and $S(X)$ are two entirely different symbols and have entirely different meanings. s_X is the standard deviation of the sample and $S(X)$ is the sample sum, i.e., the sum of the measurements in the sample.

As an example, if (3.11a) is applied to the 75 measurements of Table 2.1, we find

$$74\, s_X^2 = [\,(1.47)^2 + (1.62)^2 + \ldots + (1.47)^2\,] - \frac{1}{75}\,[\,1.47 + 1.62 + \ldots + 1.47\,]^2$$

$$= (175.5849) - \frac{1}{75}\,(114.51)^2 = .7510$$

$$s_X^2 = .01015$$

$$s_X = .101$$

You will have noticed that n-1 appears in formula (3.8) for the variance where you might have expected n. One reason for this is that although there are n squares on the right-hand side of (3.8), the sum of these squares actually reduces to n-1 squared quantities. To see this, consider the case of a sample of one measurement X_1. Here $\bar{X} = X_1$ and $(X_1 - \bar{X})^2 = 0$, so that formula (3.8) in this

case is

$$(1-1)\quad s_X^2 = (X_1 - \overline{X})^2 = 0.$$

Now let us look at the case of a sample of two measurements X_1 and X_2. Since $\overline{X} = \frac{X_1+X_2}{2}$, we have for the right-hand of (3.6)

$$(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = (X_1 - \frac{X_1+X_2}{2})^2 + (X_2 - \frac{X_1+X_2}{2})^2$$

$$= \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2.$$

Thus (3.6) reduces to

$$(2 - 1)\quad s_X^2 = \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2,$$

which has only one squared term on the right.

In case of samples of three measurements X_1, X_2, X_3, the right-hand side of (3.6) can be written as

$$(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 = \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2 + \left(\frac{X_1 + X_2 - 2X_3}{\sqrt{6}}\right)^2.$$

In other words, the sum of three squares reduces to the sum of two squares.

It is generally true that $\sum_{j=1}^{n} (X_j - \overline{X})^2$ can be written as the sum of n-1 (and no fewer) squared differences among the sample measurements. For this reason we say that $\sum_{j=1}^{n} (X_j - \overline{X})^2$ has n - 1 degrees of freedom and we use (n - 1) rather than n in formula (3.6) in defining s_X^2.

In the preceding paragraphs we have been talking about sample means and standard deviations. We should remember that in most statistical problems we

work with samples, rather than populations from which these samples are supposed to have been drawn, because it is only rarely feasible or possible to obtain measurements on an entire population, and then only in the case of a *finite* population. If we did have measurements for an entire finite population we could, of course, compute the mean a of the population just as we have calculated the mean $\overline{X}$ of a sample. Similarly for the variance.

Exercise 3.1.

1. Final inspection of nine aircraft before delivery revealed the following numbers of missing rivets: 8, 16, 14, 19, 11, 15, 8, 11, 21. Find the mean, variance and standard deviation of the number of missing rivets per plane.

2. The first ten sentences in Somervell's abridgement of Toynbee's *A Study of History* have the following numbers of words: 55, 19, 11, 39, 9, 12, 15, 28, 46, 24. Find the mean, variance and standard deviation of these sentence lengths.

3. Five 2000-piece lots of a certain electrical device contained the following numbers of defective pieces: 4, 9, 3, 2, 1. Find the mean, variance and standard deviation of the number of defectives.

4. If $X_1 = 1$, $X_2 = 6$, $X_3 = 4$, $X_4 = 7$, $X_5 = 3$, find the value of the following:

(a) $\sum_{j=1}^{5} X_j$ (d) $\sum_{j=1}^{5} (3X_j + 2X_j^2)$

(b) $\sum_{j=1}^{5} X_j^2$ (e) $\sum_{j=1}^{5} X_j (X_j - 1)$

(c) $\sum_{j=1}^{5} (X_j - 2)$ (f) $\sum_{j=1}^{5} (X_j - 1)(X_j + 1)$

5. If $X_1, X_2, \ldots, X_n$ are any numbers and if C is any constant, show that

$$\sum_{j=1}^{n} CX_j = C \sum_{j=1}^{n} X_j .$$

(Check this for the example $X_1 = 1, X_2 = 2, \ldots, X_n = n$, $n = 8$, $C = 5$.)

6. If $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are any two sets of numbers and A and B are any two constants, show that

$$\sum_{j=1}^{n} (AX_j + BY_j) = A\sum_{j=1}^{n} X_j + B\sum_{j=1}^{n} Y_j .$$

(Check this for the example: $A = -6$, $B = 9$, $n = 4$, $X_1 = 2$, $X_2 = 4$, $X_3 = 5$, $X_4 = 6$, $Y_1 = 2$, $Y_2 = -1$, $Y_3 = 0$, $Y_4 = 3$.)

7. Suppose $X_1, X_2, \ldots, X_n$ are n measurements which have mean 10 and standard deviation 3. If a new measurement Y is obtained from each X measurement by the equation $Y = 4X + 2$, what are the mean and variance of $Y_1, Y_2, \ldots, Y_n$?

8. If the mean of measurements $X_1, X_2, \ldots, X_n$ has value A and the standard deviation has value B, and if $Y = aX + b$, what are the mean and standard deviation of $Y_1, Y_2, \ldots, Y_n$, expressed in terms of A, B, a and b?

3.2 Remarks on the Interpretation of the Mean and Standard Deviation of a Sample.

It was found that the mean of the sample of 75 measurements in Table 2.1 is 1.527. Turning to Figure 2.1, it will be seen that the mean (indicated by the arrow) is near the "middle" of the distribution of dots. Actually the mean is at the center of gravity of the distribution. By this we mean that if the dots in Figure 2.1 were all of equal weight and could be imagined as neatly arranged piles of blocks setting on a thin board (the X-axis) then this arrangement would be just balanced by holding a knife-edge under the board at the mean 1.527 ounces. The mean has another property: If we take each measurement minus the mean, we get 75 "discrepancies" or differences; some of these are positive and some are negative, but, as we have seen from expression (3.4), the algebraic sum of all of the differences is equal to zero.

We have seen that the mean of a set of measurements gives us some information about where the "middle" or "center of gravity" of the set of measurements falls, but it gives no information about the "scatter" (or "amount of concentration") of the measurements. For example, the 5 measurements 14, 24.5, 25, 25.5 and 36 have the same mean as the 5 measurements 24, 24.5, 25, 25.5 and 26, but the two sets of measurements have widely different amounts of "scatter". One simple indication of the "scatter" of a set of measurements is the range,

i.e., the largest value minus the smallest. In the two sets of measurements mentioned, the ranges are 22 and 2 respectively. If we always worked with fairly small samples of the same size n (as is the case in industrial quality control, where n = 4 and n = 5 are widely used sample sizes) then we would find the range very convenient. It is difficult, however, to compare a range for one sample size with that for a different sample size. For this and other reasons, the range, in spite of its simplicity, convenience and importance, is used only in rather restricted situations. It is very widely used in the field of industrial quality control.

The inter-quartile range defined in Section 2.1 is only useful when the samples are large enough to establish the quartiles fairly well. For n less than about 25 the quartiles are of doubtful value.

We clearly need a measure of scatter which can be used in samples of any size and in some sense makes use of all the measurements in the sample. There are several measures of scatter that can be used for this purpose, and the most common of these is the standard deviation. For normal (or Gaussian) distributions, to be described roughly below and in greater detail in Chapter 8, the standard deviation is the "natural" measure of scatter.

Many samples of measurements yield cumulative polygons on probability paper which are nearly straight lines. This means that their frequency histograms are fairly symmetrical and bell-shaped, and that:

(a) About 95% of the measurements fall within a distance of $2 \cdot s_X$ (two standard deviations) of $\bar{X}$.

(b) About 68% of the measurements fall within a distance of s_X (one standard deviation) of $\bar{X}$.

(c) About 50% of the measurements fall within a distance of $\frac{2}{3}s_X$ ($0.6745 s_X$ to be more precise) of $\bar{X}$.

For example, in the case of the zinc coating measurements of Table 2.1, (which did give a fairly straight cumulative polygon as you remember from Figure 2.5) we found $\bar{X} = 1.527$ and $s_X = .101$. Within 2(.101) of 1.527 (i.e., between 1.325 and 1.729) there are 71 out of 75 measurements or 94.7% instead of 95%. Within .101 of 1.527 (i.e., between 1.426 and 1.628) there are 52 out of 75 measurements, or 69.3% instead of 68%. Within $\frac{2}{3}$(.101) of 1.527 (i.e., between 1.460 and 1.594) there are 35 out of 75 measurements, or 46.7% instead

of 50%. In all cases the agreement between the "actual" percentages and the "theoretical" percentages is good.

Those distributions of indefinitely large populations whose cumulative polygons are straight lines on probability paper are called normal or Gaussian distributions, and are of great importance in statistics. Since large samples of many kinds of actual measurements give nearly straight cumulative polygons on probability paper, the theory of normal distributions has important practical consequences. In particular, we shall often want to select a normal distribution which "fits" our sample of measurements. Fitting such a distribution depends on knowledge of the mean (which tells where to center the distribution) and the standard deviation (which tells how widely to spread the distribution). We shall show how to fit a normal distribution in Chapter 8.

3.3 The Mean and Standard Deviation for the Case of Grouped Data.

The determination of the mean and standard deviation of a sample of data by the formulas of Section 3.1 are likely to be unnecessarily laborious for a large sample. Large samples of observations are usually treated as grouped data. When computing by hand a sample of more than about 50 measurements (or when using a computing machine a sample of more than about 100 measurements) should almost surely be treated as grouped data.

3.31 An example.

To see how we should proceed in calculating the mean and standard deviation of a grouped distribution, let us return to the data in Table 2.2 as an example. The only quantities that will be needed from Table 2.2 in computing the mean and standard deviation are the cell midpoints and the frequencies which are rewritten as columns (a) and (b) of Table 3.1.

Remember that in grouping the data of Table 2.1 and arranging it in Table 2.2 we are arbitrarily assigning the one measurement falling in the cell $1.30 \pm .025$ the value 1.30, assigning the 5 measurements falling in the cell $1.35 \pm .025$ the value 1.35, and so on. Thus, when the data are grouped, we consider that we have the following measurements: one measurement with the value 1.30, 5 with the value 1.35, 6 with the value 1.40, and so on. The mean $\bar{X}$ of the 75 measurements is obtained by applying formula (3.2). We have

$$75\,\bar{X} = (1.30) + (1.35 + 1.35 + 1.35 + 1.35 + 1.35)$$

$$+ (1.40 + 1.40 + 1.40 + 1.40 + 1.40 + 1.40)$$
$$+ \ldots + (1.75 + 1.75 + 1.75)$$

or

$$75\,\bar{X} = 1(1.30) + 5(1.35) + 6(1.40) + \ldots + 3(1.75) = 114.55\ .$$

Hence, the mean $\bar{X}$ is

$$\bar{X} = \frac{114.55}{75} = 1.527.$$

Notice that the quantity on the right-hand side of the expression for $75\,\bar{X}$ is the sum of the entries in column (c) in Table 3.1.

TABLE 3.1

Table Showing Calculations for Obtaining Mean and Standard Deviation of Sample from Grouped Data

(a) Cell Midpoint x_i	(b) Frequency f_i	(c) $f_i x_i$	(d) $f_i x_i^2$
1.30	1	1.30	1.6900
1.35	5	6.75	9.1125
1.40	6	8.40	11.7600
1.45	13	18.85	27.3325
1.50	8	12.00	18.0000
1.55	17	26.35	40.8425
1.60	14	22.40	35.8400
1.65	7	11.55	19.0575
1.70	1	1.70	2.8900
1.75	3	5.25	9.1875
Total	n = 75	$S(X) = 114.55$	$S(X^2) = 175.7125$

We find the standard deviation of the grouped measurements by applying formula (3.11) (or the briefer formula (3.11a)) to the grouped measurements:

$$74\,s_X^2 = (1.30)^2 + [\,(1.35)^2 + (1.35)^2 + (1.35)^2 + (1.35)^2 + (1.35)^2\,]$$
$$+ \ldots + [\,(1.75)^2 + (1.75)^2 + (1.75)^2\,] - \frac{1}{75}\,(114.55)^2$$
$$= 175.7125 - \frac{1}{75}\,(114.55)^2.$$

Hence

$$s_X^2 = .01022$$

and

$$s_X = .101.$$

Note that the sum of the squared terms in the expression for $74\, s_X^2$ is the total of the entries in column (d) of Table 3.1.

In general, there is a slight difference between $\bar{X}$ as calculated from the ungrouped measurements, and $\bar{X}$ as calculated from the grouped measurements. In the example of the zinc coatings, the values of $\bar{X}$ for the grouped and ungrouped cases are $\frac{114.55}{75}$ and $\frac{114.51}{75}$, respectively. These two quotients have the value 1.527 to three decimal places. Similarly, there are, in general, differences between the values of s_X as calculated from the grouped and ungrouped measurements. In the example of the zinc coatings, the values of s^2 are .01022 and .01015 respectively, for the grouped and ungrouped cases. These discrepancies are due to the grouping operation. In the present example, grouping the data into 10 cells of length .05 ounces does not change the value of $\bar{X}$ or s_X to any practical extent. In any given problem, it is evident that decreasing the size of the cells tends to decrease the effect of grouping.

3.32 The general case.

To find expressions for the mean $\bar{X}$ and standard deviation s_X in the case of a general grouped frequency distribution, we consider the cell midpoints and frequencies in a general grouped frequency table as given in columns (a) and (b) of Table 3.2. We construct column (c) of values of $f_i x_i$, i.e., products of cell midpoints and frequencies. Similarly, we construct a column of values of $f_i x_i^2$, i.e., products of squared cell midpoints and frequencies.

The mean $\bar{X}$ of the grouped measurements is obtained by applying formula (3.2a) to the n individual grouped measurements. But when (3.2a) is applied we get

$$n\bar{X} = \overbrace{(x_1+x_1+\ldots+x_1)}^{f_1 \text{ terms}} + \overbrace{(x_2+x_2+\ldots+x_2)}^{f_2 \text{ terms}} + \ldots + \overbrace{(x_k+x_k+\ldots x_k)}^{f_k \text{ terms}},$$

or

$$n\bar{X} = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k ,$$

which may be written more compactly as

(3.12) $$n\,\overline{X} = \sum_{i=1}^{k} f_i\, x_i \ ,$$

from which $\overline{X}$ may be more explicitly written as

(3.13) $$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i \ .$$

TABLE 3.2

General Table for Finding Mean and Standard Deviation of Sample from Grouped Data

(a) Cell Midpoint x_i	(b) Frequency f_i	(c) $f_i x_i$	(d) $f_i x_i^2$
x_1	f_1	$f_1 x_1$	$f_1 x_1^2$
x_2	f_2	$f_2 x_2$	$f_2 x_2^2$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
x_i	f_i	$f_i x_i$	$f_i x_i^2$
.	.	.	.
.	.	.	.
.	.	.	.
x_k	f_k	$f_k x_k$	$f_k x_k^2$
Total	n	$S(X) = \sum_{i=1}^{k} f_i x_i$	$S(X^2) = \sum_{i=1}^{k} f_i x_i^2$

Thus, in the case of grouped measurements we find the value of $S(X)$ for expression (3.3a) from the formula

$$S(X) = \sum_{i=1}^{k} f_i x_i, \tag{3.14}$$

which is simply the sum of the entries in column (c) of Table 3.2.

The standard deviation s_X is given by applying the formula (3.11a) to the n individual grouped measurements. We have already seen that the value of S(X) for grouped measurements is given by formula (3.14). In the case of $S(X^2)$, we similarly have

$$S(X^2) = \sum_{i=1}^{k} f_i x_i^2, \tag{3.15}$$

which is given by the sum of the entries in column (d) of Table 3.2. Hence, the formula for the standard deviation for grouped data is given by

$$(n-1)\, s_X^2 = \left(\sum_{i=1}^{k} f_i x_i^2\right) - \frac{1}{n}\left[\sum_{i=1}^{k} f_i x_i\right]^2. \tag{3.16}$$

Exercise 3.3.

1. "On" temperaturesat which a certain thermostatic switch operated in 25 trials were as follows (Grant data):

55	54	55	51	53
55	55	54	51	56
55	55	54	53	55
54	53	50	52	56
55	55	50	56	55

Find the mean and variance of the "on" temperatures.

2. Suppose 1000 pieces of enameled ware are inspected and the number of surface defects on each piece is recorded. If the distribution of number of defects is

No. of defects (cell midpoint) x_i	Frequency f_i
0	600
1	310
2	75
3	13
4	2

Find the mean and variance of the numbers of defects.

3. In a germination experiment, 80 rows of cabbage seed with 10 seeds per row were incubated. The distribution of number of cabbage seed which germinated per row was as follows (Tippett data):

No. seeds germinated per row (cell midpoint) x_i	Frequency of rows f_i
0	6
1	20
2	28
3	12
4	8
5	6

Find the mean and variance of the numbers of seeds germinating per row.

4. A pair of cheap plastic dice were thrown 100 times and the distribution of the total number of dots obtained was as follows:

No. of dots per throw (cell midpoint) x_i	Frequency f_i
2	0
3	7
4	9
5	19
6	16
7	13
8	11
9	4
10	6
11	11
12	4

Find the mean and variance of the numbers of dots obtained.

3.4 Simplified Computation of Mean and Standard Deviation.

For computational purposes it often saves a great deal of labor to calculate the mean and standard deviation by changing the measurements to a new scale with a new origin or to a new scale with a new unit and a new origin. First, we shall consider the simplest case: the use of a working origin.

3.41 Effect of adding a constant.

Suppose we convert the X measurements $X_1, X_2, \ldots, X_n$ to new measurements $Y_1, Y_2, \ldots, Y_n$ by adding any constant a, i.e., by using the following relation between any X value and its corresponding Y value:

$$Y_j = X_j + a.$$

The constant a may be either positive or negative. Let us consider the relation between the mean $\overline{Y}$ of the Y measurements and the mean $\overline{X}$ of the X measurements. We have

$$S(Y) = \sum_{j=1}^{n} (X_j + a) = \sum_{j=1}^{n} X_j + n\,a = S(X) + n\,a,$$

or

$$S(Y) = S(X) + n\,a.$$

Dividing by n and using (3.3) we find

(3.17) $$\overline{Y} = \overline{X} + a.$$

Hence, the effect of adding a constant a to each of a set of X measurements is to add a to the mean of the measurements. This statement also holds for grouped measurements.

Next we see that for any measurement

$$Y_j - \overline{Y} = (X_j + a) - (\overline{X} + a)$$

$$= X_j - \overline{X},$$

which means that the deviations from the sample mean are unchanged by adding a constant a to each measurement. Hence, the squares of these deviations remain unchanged, and therefore, we have

(3.18) $$\sum_{j=1}^{n} (X_j - \overline{X})^2 = \sum_{j=1}^{n} (Y_j - \overline{Y})^2,$$

or written another way

(3.19) $$S(X^2) - \frac{1}{n}[S(X)]^2 = S(Y^2) - \frac{1}{n}[S(Y)]^2 .$$

Expressed still another way, if s_X^2 is the variance of the X measurements and s_Y^2 is the variance of the Y measurements, we have

$$s_X^2 = s_Y^2 \text{ and } s_X = s_Y .$$

Therefore, adding a constant a to each X measurement does not change the variance or standard deviation of the measurements.

3.42 Examples of using a working origin.

As a simple example of the use of a working origin in finding the mean and standard deviation of a set of measurements, let us return to the example of the 5 weights mentioned early in Section 3.1. The five X measurements are 141, 136, 157, 143 and 138. Let us take the constant a to be - 140. Using the relation Y = X - 140, we find the five Y measurements to be + 1, - 4, + 17, + 3, - 2, respectively.

First, we consider the mean. We have,

$$S(Y) = + 15$$

and therefore from (3.17)

$$\frac{15}{5} = \overline{X} + (- 140) ,$$

or

$$\overline{X} = 143,$$

as found in Section 3.1.

Now consider the variance. To calculate the variance of the Y measurements we need S(Y) and $S(Y^2)$. Their values are S(Y) = 15, $S(Y^2)$ = 319. Hence

$$(4)\quad s_Y^2 = 319 - \frac{1}{5}(15)^2 ,$$

$$= 68.5 ,$$

or

$$s_Y = 8.27 = s_X ,$$

as found in Section 3.1.

A working origin may be used in a similar way for grouped measurements. Let us return to our example of zinc coatings, and use 1.50 ounces as our working origin. Table 3.3 shows the computation. (Ignore columns (f), (g) and (h) for the moment.)

To determine the value of $\overline{X}$, we note that

$$\overline{Y} = \frac{S(Y)}{n} = \frac{2.05}{75} = .0273 ,$$

and since

$$\overline{Y} = \overline{X} - 1.50 ,$$

we get

$$\overline{X} = 1.50 + .0273$$

$$= 1.527 ,$$

as found in Section 3.1 directly from the X measurements.

To determine the value of s_X^2 we have from (3.19)

$$(n-1)\, s_X^2 = (n-1)\, s_Y^2 = S(Y^2) - \frac{1}{n}[S(Y)]^2 .$$

Substituting values,

$$74\, s_Y^2 = .8125 - \frac{1}{75}[2.05]^2$$

$$= .7565.$$

Hence,

$$s_Y^2 = .01022$$

and

$$s_Y = .101 = s_X ,$$

as found in Section 3.3.

Columns (f), (g), (h) in Table 3.3 enable us to check our computations, for we have

$$\sum_{i=1}^{k} f_i (y_i + 1)^2 = \sum_{i=1}^{k} f_i y_i^2 + 2 \cdot \sum_{i=1}^{k} f_i y_i + n ,$$

which may be written more briefly as

(3.20) $$S[(Y + 1)^2] = S(Y^2) + 2 \cdot S(Y) + n.$$

Applying this to Table 3.3, we have

TABLE 3.3

Table Showing Computations Required in Computing Mean and Variance of a Frequency Distribution by Using a Working Origin

(a) Cell Midpoint x_i	(b) Frequency f_i	(c) $y_i = (x_i - 1.50)$	(d) $f_i y_i$	(e) $f_i y_i^2$	(f) $y_i + 1$	(g) $(y_i + 1)^2$	(h) $f_i(y_i + 1)^2$
1.30	1	-.20	-.20	.0400	.80	.6400	.6400
1.35	5	-.15	-.75	.1125	.85	.7225	3.6125
1.40	6	-.10	-.60	.0600	.90	.8100	4.8600
1.45	13	-.05	-.65	.0325	.95	.9025	11.7325
1.50	8	0	0	0	1.00	1.0000	8.0000
1.55	17	.05	.85	.0425	1.05	1.1025	18.7425
1.60	14	.10	1.40	.1400	1.10	1.2100	16.9400
1.65	7	.15	1.05	.1575	1.15	1.3225	9.2575
1.70	1	.20	.20	.0400	1.20	1.4400	1.4400
1.75	3	.25	.75	.1875	1.25	1.5625	4.6875
Total	75		$S(Y) = 2.05$	$S(Y^2) = .8125$			$S[(Y+1)^2] = 79.9125$

Check: $S[(Y+1)^2] = S(Y^2) + 2S(Y) + n$

$79.9125 = .8125 + 2(2.05) + 75$

$= 79.9125$

$$79.9125 = .8125 + 2(2.05) + 75$$
$$= 79.9125.$$

If (3.20) holds for any given table, we can be practically certain (although not absolutely certain) that our computations are correct. If it fails to hold, we can be sure that we have made an error.

3.43 Fully coded calculation of means, variances and standard deviations.

While we are at this job of trying to simplify the computation of the mean and standard deviation, there is one more step we may as well take -- at the expense of over-emphasizing the problem of calculating means and standard deviations. This step simplifies such calculation as much as possible.

To do this, we make a change in scale in addition to making a change in origin (as discussed in Sections 3.3 and 3.4), so that our deviations from the working origin are simple to square and multiply.

Suppose Z measurements are defined in terms of X measurements by the relation

$$X_j = a + b Z_j ,$$

where a and b are any constants. We may refer to the Z measurements as coded values of the X measurements.

Summing, we have

$$S(X) = n a + b S(Z),$$

and dividing by n, we have

(3.21) $$\bar{X} = a + b \bar{Z} .$$

Now consider the deviations between Z_j and its mean $\bar{Z}$. We have

$$X_j - \bar{X} = (a + b Z_j) - (a + b \bar{Z})$$

or

$$X_j - \bar{X} = b(Z_j - \bar{Z}).$$

Squaring and summing

(3.22) $$\sum_{j=1}^{n} (X_j - \bar{X})^2 = b^2 \cdot \sum_{j=1}^{n} (Z_j - \bar{Z})^2 .$$

But

$$\sum_{j=1}^{n} (Z_j - \bar{Z})^2 = (n-1)\, s_Z^2$$

and

$$\sum_{j=1}^{n} (X_j - \overline{X})^2 = (n-1)\, s_X^2 .$$

Therefore, from (3.22) we find

$$s_X^2 = b^2 \cdot s_Z^2 \tag{3.23}$$

or

$$s_X = b \cdot s_Z . \tag{3.24}$$

Hence we have the following rule expressed by (3.21) and (3.24):

If X and Z are measurements satisfying the relation $X_j = a + b\, Z_j$, then the mean of the X's is a plus b times the mean of the Z's, and the standard deviation of the X's is b times the standard deviation of the Z's.

This rule holds no matter whether we are talking about grouped or ungrouped measurements. However, it does not pay to code measurements unless there are enough of them to justify calculation as grouped data. When we have grouped data and wish to use coded values, it is natural to choose the value of the constant b equal to the cell-length, and to choose the value of a to be a cell midpoint near the middle of the grouped frequency distribution.

Example.

Returning to the example of the zinc coating measurements, the essential constituents for coded computation are shown in Table 3.4.

Note that we do not need a column for $(z_i + 1)$, since the values of $(z_i + 1)^2$ can be written down easily from sight.

In this example,

$$a = 1.50, \quad b = .05.$$

We have $S(Z) = 41$, and $S(Z^2) = 325$.

Therefore,

$$\overline{Z} = \frac{41}{75} = .55$$

and

$$74\, s_Z^2 = 325 - \frac{1}{75}(41)^2$$

$$= 303.5867$$

or

$$s_Z^2 = 4.0478$$

$$s_Z = 2.0119 .$$

Therefore, from (3.21) we find

$$\bar{X} = 1.50 + (.05)(.55) ,$$
$$= 1.528$$

and from (3.24),

$$s_X = (.05)(2.0119)$$
$$= .101 .$$

These values of $\bar{X}$ and s_X are very close to those found by the longer methods of Section 3.1 and 3.3.

TABLE 3.4

Table Showing Computations Needed for Fully Coded Computation of Mean and Variance of a Frequency Distribution

(a) Cell Midpoint x_i	(b) Frequency f_i	(c) z_i	(d) $f_i z_i$	(e) $f_i z_i^2$	(f) $(z_i+1)^2$	(g) $f_i(z_i+1)^2$
1.30	1	-4	-4	16	9	9
1.35	5	-3	-15	45	4	20
1.40	6	-2	-12	24	1	6
1.45	13	-1	-13	13	0	0
1.50	8	0	0	0	1	8
1.55	17	1	17	17	4	68
1.60	14	2	28	56	9	126
1.65	7	3	21	63	16	112
1.70	1	4	4	16	25	25
1.75	3	5	15	75	36	108
Total	n = 75		S(Z) = 41	$S(Z^2) = 325$		$S[(Z+1)^2]=482$

Check: $S[(Z+1)^2] = S(Z^2) + 2\,S(Z) + n$

$482 = 325 + 2(41) + 75$

Exercise 3.4.

(Every student is expected to do problem No. 9.)

1. Making use of a working origin, find the mean, variance and standard deviation of the following sample of measurements of copper content (in percent) in

10 bronze castings: 85.54, 85.54, 85.72, 85.48, 85.54, 85.72, 86.12, 85.47, 84.98, 85.12.

2. Suppose the mean and variance of a distribution of lengths expressed in inches are 28.3 and 16.0. What would the mean, variance and standard deviation of the distribution be if the lengths were expressed in feet?

3. The burning lives (in hours to the nearest 10 hours) for 10 incandescent light bulbs of a certain type were found to be as follows: 850, 900, 1370, 1080, 1060, 860, 1060, 1040, 1090, 1930. Find the mean, variance and standard deviation of this sample, making use of a working origin.

4. The Mathematical Aptitude Scores of the 24 Princeton Chemistry Department Seniors of the Class of 1938 were: 550, 569, 440, 608, 814, 695, 577, 595, 730, 698, 518, 705, 692, 563, 531, 582, 479, 505, 711, 539, 614, 614, 524, 653. Find the mean, variance and standard deviation of these scores.

5. A pair of dice can fall in 36 different "ways". One of these "ways" will yield a total of 2 dots, two will yield 3 dots, and so on. The frequencies of "ways" which will yield various total numbers of dots is as follows:

Total No. Dots x_i	Frequency f_i
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1

Find the mean, variance and standard deviation of total number of dots. (Use a working origin.)

6. Thirty two dice were thrown 100 times. The distributions of the total number of dots per throw was as follows:

Total no. of dots (cell midpoints) x_i	Frequency f_i
90	3
95	3
100	4
105	15
110	17
115	15
120	16
125	12
130	8
135	3
140	3
145	0
150	1

Find the mean, variance and standard deviation of the total number of dots per throw (by using either a working origin or full coding).

7. The following distribution of carbon content (percent) was obtained in 178 determinations on a certain mixed powder (Davies data):

Percent carbon (cell midpoint) x_i	Frequency f_i
4.145	1
4.245	2
4.345	7
4.445	20
4.545	24
4.645	31
4.745	38
4.845	24
4.945	21
5.045	7
5.145	3

Find the mean, variance and standard deviation of the carbon content (using working origin or full coding).

8. The grouped frequency distribution of thickness measurements (in inches) determined from 50 places on a coil of sheet metal were found to be as follows (Westman data):

Cell Midpoint x_i	Frequency f_i
.147	1
.148	1
.149	4
.150	6
.151	10
.152	6
.153	7
.154	6
.155	6
.156	3

Find the mean, variance and standard deviation of this distribution (by using either a working origin or full coding).

9. Find the mean, variance and standard deviation of the grouped frequency distribution obtained in whichever one of the following problems in Exercise 2.2 you have done: No. 8, 9, 10, 11, 12, 13, 14. Use either a working origin or full coding. (Keep a record of your work.)

10. Suppose $\overline{U}$ and s_U^2 are the mean and variance of a sample of m U measurements and $\overline{V}$ and s_V^2 are the mean and variance of a sample of n V measurements. Suppose all of these m + n measurements are put in one sample.

(a) Show that the mean of this sample is

$$\frac{m\,\overline{U} + n\,\overline{V}}{m + n}.$$

(b) Show that the variance of this sample is

$$\frac{1}{m+n-1}\left[(m-1)s_U^2+(n-1)s_V^2+\frac{m\cdot n}{m+n}(\overline{U}-\overline{V})^2\right].$$

CHAPTER 4. ELEMENTARY PROBABILITY

4.1 Preliminary Discussion and Definitions.

In the preceding chapters we have dealt with the problem of describing a given sample of quantitative measurements. We have shown how to describe such a sample graphically (by using a dot frequency diagram, a cumulative graph, a frequency histogram or a frequency polygon) and how to describe a sample numerically (by graphical calculation of the median, quartiles and other percentiles and by arithmetical calculation of the mean and standard deviation). But as pointed out in the introduction, we are usually interested in more than a mere description of a sample. We are interested in making inferences about the populations from which the samples come, i.e., we are interested in making statements about intervals within which population means, standard deviations and other population parameters are likely to lie. In order to be able to make such statements, we must know something about how much sample means, standard deviations and other sample statistics vary from sample to sample, when repeated samples of a stated size are drawn from the same population.

There are two approaches in the study of sample-to-sample fluctuations of sample statistics: experimental and mathematical.

In the experimental approach, we determine by repeated experiments (i.e., repeated drawing of samples) how a given sample statistic will be distributed. For example, if we want to know something about the distribution of means in samples of 30 measurements out of a given indefinitely large population, we draw a large number (say 100) of samples of 30 out of the population and calculate the mean of each sample and make a frequency distribution of these means and find its mean and standard deviation. In the case of sampling from a finite population, one would have to return the sample of objects to the population after each drawing. The main difficulty with the experimental method of determining sampling fluctuations is that it is very time-consuming and often costly. Even in such simple experiments as rolling dice, it takes a great deal of time to accumulate enough data to see how sampling laws work.

By taking the mathematical approach we can determine theoretical sampling laws of some of the simpler sampling statistics, particularly sample sums

and sample means, which apply pretty well to practical situations. It should be stated, however, that there are many sample statistics which have very complicated theoretical sampling laws, even for the simplest kinds of populations. In such cases one often has to resort to experimental sampling, using random numbers, to find out something about the sampling fluctuations. The mathematical approach to the study of sampling laws is based on the theory of probability. We shall now spend some time on the subject of probability.

The word probability or chance is used loosely in our everyday conversation and we know vaguely what it means. To give a few examples, we talk of the chance of winning a game of cards or dice or a football game; the probability of its raining tomorrow; the chance of a person living to be so many years old. In all these cases we are interested in a future event, of which the outcome is uncertain, and about which we want to make a kind of prediction. Sometimes we are content with a rough qualitative statement like: "It is very probable that it will rain tomorrow"; or "A man has little chance of living a hundred years". Sometimes we go further and tend to be numerical, as when we say "There is a fifty-fifty chance that we will win the game", or "I'll bet you two to one that it will rain today". In a mathematical discussion of probability we try to present conditions under which we can make sensible numerical statements about uncertainties, and to present methods of calculating numerical values of probabilities and expectations.

It must be agreed that when the term probability can be applied so loosely to so many diverse and complex phenomena as games, the weather, the span of human life, etc., it is hardly conceivable that one can give it a definite and precise meaning without some simplification. Familiar examples of application do not always lend themselves to simple analysis. For instance, it is well-known that the science of weather forecasting still is far from perfect; and for a layman to assign a probability to the weather of tomorrow in any doubtful case is hardly more than guesswork. Again, at a Princeton-Yale match, it will not be surprising to find Princeton and Yale men forming quite different estimates of the outcome of the game, even apart from sentimental reasons. In such a case, it may well be questioned if there is an unequivocal way of assigning a probability.

These and many other considerations one may think of should convince us that in order to be on solid ground we must confine ourselves at first to the simplest phenomena of chance. (But if the study of statistics is to be useful

in science, engineering, or even everyday life, we will have to study more and more complex cases and eventually stand on less solid ground.) An example of the simplest kind of chance phenomena is the toss of a coin. Here there are two possible outcomes: head or tail. The probability in question is whether the event "head" or the event "tail" will occur. Now although two men may well disagree on the weather of tomorrow or the result of a game, they would readily agree to call the chances even that one would get a head or a tail in tossing an ordinary coin. Indeed, agreement on this point is so universal that we say "It's a toss-up" when we want to say that "Its probability is 1/2".

Another simple case is the roll of a die. An "ideal" die is a perfect cube with six faces having 1 to 6 dots. If we roll it once there are six possible outcomes; the face turning uppermost may be any of the six faces and accordingly the number of dots we obtain may range from 1 to 6. Here again we would agree to assign equal probabilities to the six faces (of an "ideal" die) so that each gets its equal share (1/6th) of the total probability.

The case of a pack of playing cards is similar, though a little more complicated. There are 52 cards in a pack and if we pick a card "at random" from a well-shuffled pack, the chance of getting any particular card (named in advance) is 1/52. The chance of getting a spade is 1/4, that of getting a spade or a heart is 1/2. The chance of getting a face card is 12/52.

It is easy to see how these probabilities are obtained. In the case of the toss of a coin there are two possible cases, head and tail; and if the coin is "ideal", i.e., uniformly and symmetrically made, and the tossing is "fairly done", there is nothing in favor of the turning up of the one or the other. Or we may put it this way: any argument which may be advanced for one side of the coin applies equally well to the other side, so much so that in the end we have no reason to expect the one rather than the other. We then consider that the two possible alternatives head and tail are equally likely and the probability of each is one (case) in two (equally likely cases), i.e., 1/2. Similarly, the case of rolling a die admits six possible cases, i.e., the faces with 1, 2, 3, 4, 5, 6 dots. If the die is "well made" (in particular not loaded) and the rolling is "arbitrary", the six cases are considered equally likely. Thus each face, being one (case) in six (equally likely cases) gets the probability 1/6. The case of playing cards is entirely similar.

We are thus led to the following definition of probability in the simplest situations:

Definition I. If an event E can happen in m cases out of a total of n possible cases which are all considered by mutual agreement to be equally likely, then the probability of the event E is defined to be m/n, or more briefly Pr(E) = m/n.

Thus in the evaluation of the probability of an event we need to know two numbers: the number of possible cases and the number of favorable cases, i.e., those in which the event will have occurred. The ratio of the smaller number to the larger is the desired probability. Of course, it must be agreed in each instance that the possible cases are all considered equally likely. This can often be reached by plain common sense as in simple cases like those mentioned above.

This definition, as far as it goes, is perfectly clear-cut and agrees with our intuitive notion of chance. But, as the reader will easily see, it is difficult or impossible to apply as soon as we leave the field of coins, dice, cards and other simple games of chance. To return to one of the examples we had, how does it apply to the probability of rain? What are the possible cases? We might naïvely think that there are the two contingencies: "rain" and "no rain". But at any given locality it will not usually be agreed that they are equally likely. In fact, it is exactly the relative likelihood of these two cases that we are seeking -- we must beware of a vicious circle. (It should be noticed that in simple cases of coins, dice, cards, etc., this vicious circle is avoided by reasoning based on plain common sense.) In general, our first definition of probability cannot be applied whenever it is impossible to make a simple enumeration of cases which can be considered equally likely.

In order to assign probabilities to more complex phenomena we must appeal to a principle of agreement based on observational evidence, namely,

Definition II. If (a) whenever a series of many trials is made, the ratio of the number of times event E occurred to the total number of trials is nearly p, and if (b) the ratio is usually nearer to p when longer series of trials are made, then we agree in advance to define the probability of E as p, or more briefly Pr(E) = p.

For example, the probability of getting a head in a toss of an "ideal" coin is agreed to be 1/2 (under Definition I). In a large number of trials the percentage of heads is found experimentally to be nearly 50 percent, that is, heads come up in about half of the trials. Thus, in 1000 tosses we would get

"about" 500 heads, in 1200 rolls of a die we would get "about" 200 sixes, etc. Interpreted in statistical language, we can think of 1000 tosses of a coin as a sample of size 1000 from the indefinitely large population of tosses which could be made with that coin. Thus, if we should consider a number of samples of 1000 tosses, we might get some such sequence of heads as 488, 518, 508, 474, 497, etc. These numbers are all "nearly" equal to 50 percent of the total number of trials, that is, the relative frequencies of number of heads are distributed near .5 in a frequency distribution, with a certain standard deviation. If smaller sample sizes had been used, the relative frequencies would have been distributed around .5 with a larger scatter, i.e., with a larger standard deviation. If still larger samples, say of 10,000 were used, the standard deviation of the observed relative frequencies would have been less. Actually, there is a theoretical distribution of relative frequencies of heads in samples of n tosses, which will be discussed in Chapter 6 on the Binomial Distribution. The standard deviation of this theoretical distribution will be found to vary inversely as the square root of n. In actual experiments with coins, one will find that the standard deviations of distributions of relative frequencies in samples of different sizes behave in fairly close agreement with this inverse square root law.

Definition II gives us a way to "estimate" probabilities from experimental results in a simple way. For instance, from mortality tables compiled in past years, we find that for 100,000 new-born white American males, about 92,300 of them are alive at age 20. This allows us to estimate the probability that a new-born white American male will live to be 20 years old as 0.923. (Because of the way the data were accumulated, we believe this is "about" right! In fact it and many other similarly estimated probabilities of survival are used in calculating life insurance premiums.) Similarly, if we have examined 1000 electric light bulbs manufactured by a certain company, and find 20 defective, we estimate the probability for a bulb of this type ("taken at random") to be defective to be $20/1000 = .02 = 2\%$.

Exercise 4.1.

1. What are the greatest and least values a probability can have? What do these extreme values mean?

2. If the probability that an event occurs is m/n, what is the probability that it does not occur?

3. In a roll of one die, what is the probability that the number of dots obtained will not exceed 4? That the number of dots will be an even number?

4. If we draw a card from a pack of playing cards, what is the probability of not obtaining a spade? Of obtaining the nine of spades or the ten of spades? Of obtaining a nine or a ten of some single suit? Of obtaining a card with a nine or a ten on it?

5. How would you estimate the probability that a student picked at random from a list of all Princeton undergraduates is over 6 feet tall?

6. The following mortality table shows the number of survivors to various ages of 100,000 new-born white American males:

Age x	Survivors to Age x
0	100,000
10	93,601
20	92,293
30	90,092
40	86,880
50	80,521
60	67,787
70	46,739
80	19,360
90	2,812
100	65

(a) Estimate the probability of a new-born infant in this class living to be 60 years old.

(b) Estimate the probability of a 20-year-old in this class living until he is 50 years old.

(c) Make a cumulative polygon of "age at death" and from it estimate the probability of a 25-year-old living to be 75.

(d) From the cumulative polygon of (c) you would estimate the probability to be 1/2 that a 50-year-old would live how long?

7. How would you estimate experimentally the probability of getting less than

3 heads in throwing five coins? Carry out this experiment 50 times and record the results.

8. How would you estimate experimentally the probability of getting a busy signal between 4 and 5 P.M., on week-days from a telephone number picked at random from the telephone directory?

9. How would you estimate the probability that the first digit of a number picked at random out of the front page news text of a newspaper is a 1, 2, or 3? Estimate it from 50 such numbers.

4.2 Probabilities in Simple Repeated Trials.

We have discussed the probability involved in the toss of one coin. Now suppose that we toss two coins. The situation becomes a little more complicated. What, for example, is the probability of getting two heads in tossing two coins? In order to solve this problem mathematically we fall back on Definition I of probability. First, we have to find the number of possible cases which can happen when two coins are tossed. This is easily seen to be 4 if we actually write down all the possible combinations (H stands for head, T for tail):

		First Coin	Second Coin
Case 1	:	H	H
Case 2	:	H	T
Case 3	:	T	H
Case 4	:	T	T

More simply we can write these as

HH HT TH TT .

Hence, the total number of cases is 4. Since it is agreed that it is equally likely that either coin will turn up head or tail, it will (usually) be agreed that the four combinations in pairs will be equally likely. We must now find the number of cases in which the desired event, i.e., that of getting two heads, can occur. This is the case if, and only if, we have HH, i.e., in just one case. Therefore, applying Definition I, the probability we seek is 1/4. Note that the agreement about the four combinations of pairs being equally likely would not hold if the coins were glued together or even if they were magnetic.

It is instructive to work out the probability of getting one head and one tail. The only modification now is the number of cases in favor of this event. Now we see both HT and TH will do for our purpose, for all we care for is a head and a tail, no matter which coin is which. In other words, the first coin may turn up head and the second tail, or vice versa, and in either case we shall be satisfied. Hence, the number of cases in which the desired event can occur is two and the probability we seek is 2/4 or 1/2.

Of course, the probability of getting two tails is the same as that of getting two heads, that is, 1/4. Notice that

$$\frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1 ;$$

and also that the three events we have considered, i.e.,

(i) two heads
(ii) one head and one tail
(iii) two tails

exhaust all the possibilities of the outcome of tossing two coins. It is because of this fact that their respective probabilities add up to 1, as shown above.

Our next step will be to consider three coins. Here we can write down all the possibilities as follows:

HHH, HHT, HTH, THH, HTT, THT, TTH, TTT.

It is easy to see that the total number of possible cases is 2 x 2 x 2 = 8. For each coin has two possibilities, head or tail: and in combining the possibilities for all the coins we multiply them together.

What is the probability of getting exactly two heads? All we have to do is to pick from the above list those combinations with exactly 2 H's. They are

HHT, HTH, THH

and the number is three. Hence, the desired probability is 3/8.

It is easy to check the following results:

Pr (of getting 0 heads) = 1/8
Pr (of getting 1 head) = 3/8
Pr (of getting 2 heads) = 3/8
Pr (of getting 3 heads) = 1/8

Total = 1

Now let us take up the general case of n coins. The number of heads we may get is of course some number between 0 and n (both included); call this number x. The general problem we want to solve is this: what is the probability of getting exactly x heads if n coins are tossed? From our experience with 2 or 3 coins, it is easy to see that the total number of possible cases is $2.2. \ldots .2$ (n times) $= 2^n$. But how many cases show exactly x heads? Since the numbers n and x are general, we cannot write down all the cases as we did in the case of 2 or 3 coins. But even for particular numbers like n = 100 and x = 40 it will be too laborious to do so. We must, therefore, find them without actually listing them. Here a bit of mathematics comes in, which we shall discuss in Section 4.3.

We have considered a single toss of 1, 2, 3, ... n coins. Instead, we may consider 1, 2, 3, ... n successive tosses of a single coin. It is obvious that the situation is exactly the same if one coin is as good as another and also one toss as good as another.

Let us turn to dice. If we roll two dice, what is the number of all possible cases? It is easily seen to be 6 x 6 = 36, and we can actually write down all the combinations as follows (the first number of the pair refers to the first die, the second number to the second die -- the dice might be colored differently for example):

(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6).

This list gives us the complete information for all questions regarding two dice. For example, what is the probability that at least one die shows 6 dots? There are 11 favorable cases:

(1,6), (2,6), (3,6), (4,6), (5,6), (6,6),
(6,1), (6,2), (6,3), (6,4), (6,5).

Thus the answer is: Pr (getting at least one 6) = 11/36.

As another example, what is the probability of getting a total of 8 dots? We see that a total of 8 dots can be obtained in the following 5 ways:

(2,6), (3,5), (4,4), (6,2), (5,3).

Thus the answer is: Pr (getting a total of 8 dots) = 5/36.

Now consider the following problem. If we roll n dice (or, equivalently, if we roll a die n times) what is the probability of getting exactly x "sixes"? The problem is seen to be similar to the one posed for coins and will be solved later.

Before we leave this topic it is instructive to consider the case of biased coins and loaded dice. Up to now we have supposed that the coin is unbiased, for which reason we agreed that the probability of head (or tail) was 1/2. Now suppose that this is not so. Such a case can actually happen if, e.g., the coin is poorly made and is heavier on one side, or if the coin is worn more on one side than the other. No simple intuitive reasoning will tell us what the probability of head is now. A practical way of estimating the probability would be experimentally, i.e., to toss the coin a large number of times and apply Definition II to the relative frequency of number of heads obtained. Suppose that in 1000 throws of a coin the relative frequency of heads turns out to be .45, then by Definition II, .45 would be considered as an estimate of p which would then have to be used as though it were the true probability of a head, although it would be necessary to keep in mind that all we really know is that p has a value somewhere near the value .45. Suppose the probability of a head is p; then the probability of a tail is $1 - p$. With such a biased coin we can again ask similar questions to those already asked about unbiased coins. But it will be seen that we need a new method of computing probabilities of this kind. These will be discussed in Chapter 6 in the treatment of the binomial distribution.

One easy way of imitating a "biased coin" is the following: Take an ordinary "true" die and mark on two of the faces the letter H and the remaining four faces the letter T. Then this die will behave like a biased coin. The probability of H (head) is $2/6 = 1/3$ and that of T (tail) is $4/6 = 2/3$.

Exercise 4.2.

1. If a dime and a nickel are tossed, what is the probability that the dime shows head and the nickel tail? What is the probability that one coin shows head and the other coin shows tail?

2. If three coins are tossed, what is the probability of at least two heads?

Of at most two heads?

3. Construct a table showing all the possible cases when four coins are tossed. Find the probabilities of 0, 1, 2, 3, 4 heads, respectively.

4. What is wrong with the following argument: "If two coins are tossed, there are three possibilities: 2 heads, 1 head and 1 tail, two tails. Hence, the probability of 1 head and 1 tail is 1/3".

5. Two dice are rolled. What are the various total numbers of dots we may obtain? List their respective probabilities. What total number has the greatest probability? What is the probability of obtaining a total not exceeding 9?

6. Work problem 5, but using three dice instead of two.

7. Three dice are rolled. What is the probability of getting a pair, (i.e., exactly two faces alike)? The probability that none of the dice shows an ace?

8. Imagine we have dice made of the form of a regular tetrahedron (four faces) and marked with 1 to 4 dots on the faces. The number of dots on the bottom face is the number we get when we toss the die. Now if we toss two such dice, what is the probability of getting exactly one ace? At least one ace? Exactly two aces? What are the probabilities of getting the various possible total numbers of dots?

4.3 Permutations.

In Section 4.2 we ran into the following problem: If n coins are tossed, in how many ways can we get x heads? Let us denote head by H and tail by T as before. Now if we mark the coins 1 to n and think of them laid out in a row,

$$(1), (2), \ldots, (n),$$

then each coin may be an H or a T and the problem reduces to finding the number of possible arrangements of n symbols composed of x H's and (n - x) T's. For example, if n = 3, x = 2, we would have the following three arrangements:

HHT, HTH, THH.

For n = 5, x = 3, we have the following 10 arrangements:

HHHTT, HHTHT, HTHHT, THHHT, HHTTH,
HTHTH, THHTH, HTTHH, THTHH, TTHHH.

These arrangements are called permutations. More specifically, in the first case we have 3 permutations of 2 H's and 1 T; in the second case we have 10 permutations of 3 H's and 2 T's.

The general rule on the number of permutations in any given case may be stated as follows:

Rule on Number of Permutations: If there are n objects, of which i are alike, another j are alike, another k are alike, and so on, then the number $P^{(n)}(i, j, k, \ldots)$ of different permutations is given by the formula

$$P^{(n)}(i, j, k, \ldots) = \frac{n!}{i!\, j!\, k! \cdots} \tag{4.1}$$

where the symbol n! (read n factorial) means the product of all integers from n to 1: $n! = n(n-1) \cdots 3 \cdot 2 \cdot 1$.

For example, in the problem of n coins, we have n symbols, of which x are alike (all H) and another n-x are alike (all T). Hence the rule gives us the number of permutations

$$\frac{n!}{x!\,(n-x)!}\,. \tag{4.2}$$

If n = 3, x = 2, we get

$$\frac{3!}{2!\,1!} = 3;$$

if n = 5, x = 3, we get

$$\frac{5!}{3!\,2!} = 10.$$

We shall begin with the simple case where all n things are different. In this case i = j = k = ... = 1 and the rule reduces to the following:

Special Case: The number of permutations of n distinct objects is equal to n!, i.e.,

$$P^{(n)}(1, 1, 1, \ldots) = n!\,. \tag{4.3}$$

As an example of this special case, consider how many different arrangements 5 people can seat themselves on a bench. The answer is 5!=5x4x3x2x1 = 120.

To prove the rule in Special Case, we imagine n empty boxes marked 1, 2, ..., n lined up in a row, and n different objects to put into the boxes, one object in each box.

A box will be filled when it has one object in it. In the first box we can put any one of the n objects on hand; hence there are n ways of filling the first box. Suppose this has been done. Then we are left with n - 1 objects. We can put any one of them into the second box. Hence there n - 1 ways of filling the second box after the first one has been filled. But each way the first box can be filled can be combined with each way the second can be filled. Hence there are n(n - 1) ways the first two boxes can be filled. For the third box, there are n - 2 objects to choose from. Hence, there are n(n - 1)(n - 2) ways the first three boxes can be filled. We can continue this reasoning down to the last empty box. Then there is only one object to select, and hence only one way to fill the last box. Therefore, our reasoning leads us to the conclusion that there are

$$(n)(n - 1)(n - 2)\ldots(2)(1) = n!$$

different ways of filling the n boxes with the n different objects.

Now, if some of the things are alike (i.e., indistinguishable from each other) we shall have fewer ways of arranging them into distinguishable arrangements. For example, suppose we have 3 H's and 2 T's; for a moment let us mark them as H_1, H_2, H_3, T_1, T_2, although in reality H_1, H_2, and H_3 are indistinguishable among themselves, and so are T_1 and T_2. We have just seen that the total number of permutations, regarding the H's and T's as all distinct, is 5! Now take any permutation like

$$H_1T_1T_2H_2H_3 .$$

If we permute the three H's among themselves while leaving the T's alone, we get altogether 3! = 6 possible arrangements:

$$\begin{array}{ll} H_1\ T_1\ T_2\ H_2\ H_3 & H_2\ T_1\ T_2\ H_3\ H_1 \\ H_1\ T_1\ T_2\ H_3\ H_2 & H_3\ T_1\ T_2\ H_1\ H_2 \\ H_2\ T_1\ T_2\ H_1\ H_3 & H_3\ T_1\ T_2\ H_2\ H_1 . \end{array}$$

Now if the 3 H's were indistinguishable, i.e., if we erase the subscripts on the H's, all of these 6 permutations would be identical, and would be HT_1T_2HHH.

Thus the number of distinguishable permutations is only 1/6th of the original number of permutations when the H's were different. Similarly, owing to the fact that the T's are indistinguishable when the subscripts are dropped, the number of permutations is again reduced in the ratio of 2 to 1. Altogether, the number is reduced in the ratio of 3! x 2! or 6 x 2 = 12 to 1. Thus the reduced number of permutations (which are distinguishable) is

$$\frac{5!}{3!\ 2!}.$$

You should have a firm grasp of this simple case. Then the following general argument will be easy to understand.

The total number of permutations of n objects when regarded as all distinguishable is n!. Now suppose i of the objects are made indistinguishable. If we take any particular permutation of the n objects and permute i particular objects among themselves keeping the positions of the remaining n - i objects fixed, we get (by the previous reasoning) i! permutations. These i! permutations become indistinguishable when the i objects are made identical -- thus the i! permutations collapse into one distinguishable permutation when these i objects are made indistinguishable among themselves. Thus the number of permutations is reduced in the ratio of i! to 1. Similarly, if we make another j objects alike the number is further reduced in the ratio of j! to 1, and so on. Thus the final formula becomes, after the successive reductions:

$$\frac{n!}{i!\ j!\ k!\ \ldots}.$$

Consider an example. How many different permutations can we get from the word "statistics"? In other words, in how many ways can we scramble the letters in "statistics" and obtain arrangements which are distinguishable from one another? Let us break it down to the component letters:

a, c, i, i, s, s, s, t, t, t.

There are 10 letters, among which 1 is alike, another 1 is alike, another 2 are alike, another 3 are alike, another 3 are alike. By formula (4.1) the number of permutations is equal to

$$\frac{10!}{1!\ 1!\ 2!\ 3!\ 3!} = 50{,}400.$$

Another problem of permutation is the following. Suppose we have n

distinct objects. Instead of permuting all of them let us choose x out of them and then arrange those x objects in all possible orders. The number of different ways of doing this is given by the rule below.

Rule on Number of Permutations of x out of n: The number of ways of permuting x objects chosen from n distinct objects is given by

$$P_x^n = \frac{n!}{(n-x)!}. \tag{4.4}$$

The reasoning in establishing this rule is similar to that used for the previous rule on permutations. We imagine x boxes numbered 1 to x, and n objects, all distinct. We want to fill these boxes by x objects chosen from the n objects, one in each box. The first box can be filled in n ways, the second in n - 1 ways, and so on. Since there are x boxes we get x consecutive factors starting with n and going down to (n - x + 1), that is

$$n(n-1) \ldots (n-x+1).$$

Now let us multiply this number by $(n-x)(n-x-1) \ldots (2)(1)$, that is $(n-x)!$, and then divide by it. We obtain, finally

$$P_x^n = \frac{n(n-1)\ldots(n-x+1)(n-x)\ldots(2)(1)}{(n-x)\ldots(2)(1)} = \frac{n!}{(n-x)!}.$$

Example: The number of ways of arranging two different letters out of the word town (i.e., the number of permutations of 2 out of 4) is

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

The arrangements are:

to, ot, tw, wt, tn, nt, ow, wo, on, no, wn, nw.

Exercise 4.3.

1. In how many different ways can 10 men line up in a single line? How many if a specified man has to stand at the left end? How many if a specified man must stand at the left end and a specified man at the right end?

2. How many numbers can we obtain by rearranging the digits in the number 243402? How many if numbers beginning with 0 are excluded?

3. How many 4-letter code "words" can be made from the letters of the alphabet if no repetition is allowed? If any letter can be repeated any number of times? If the first letter has to be w and no repetition is allowed?

4. In how many ways can 20 students seat (arrange) themselves in a room which has 30 seats?

5. In how many different orders can four people seat themselves around a round table? In how many orders can they seat themselves along a bench?

4.4 Combinations.

Now we come to combinations which are more important in statistics than permutations. Suppose we have n (distinct) objects and we want to choose x from them without paying any attention to the arrangement of the x objects chosen. The number of ways of doing this is given by the following rule:

Rule on Number of Combinations: If there are n different objects, the number of ways C_x^n of selecting x out of the n objects is given by the formula

$$C_x^n = \frac{n!}{x!\,(n-x)!}; \tag{4.5}$$

C_x^n is called the number of combinations of x objects out of n.

If we were to consider the permutations of x out of n we would have $P_x^n = \frac{n!}{(n-x)!}$ such permutations. We know by the rule of permutation that each choice of x objects would give rise to x! different permutations. But all of these x! different permutations are made up of a single combination of x objects. Hence, one can break the $\frac{n!}{(n-x)!}$ permutations of x out of n into sets of x! permutations, the permutations in each of these sets being the same (one) combination of objects. Hence the number of different combinations of x objects out of n is

$$C_x^n = \frac{n!}{x!\,(n-x)!}$$

You will note that $P^{(n)}(x, n-x)$ is the same as C_x^n.

Example: Three men are to be chosen from 5 men; in how many ways can they be selected?

The answer is $$\frac{5!}{3!\,2!} = 10.$$

If we label the 5 men A, B, C, D, E, we can write down all the possibilities:

ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE.

Now we can solve the problem about n coins raised in Section 4.2, namely: If we toss a coin n times, what is the probability that we will get exactly x heads?

We have to find the number of ways of permuting n symbols, of which x are H and n - x are T. The number, given by the rule on the number of permutations, is

$$P^{(n)}_{(x,n-x)} = \frac{n!}{x!\,(n-x)!}.$$

This is the number of cases favorable to x heads. Since we know the total number of cases is 2^n, which we consider to be equally likely, the probability we seek is

$$(4.6)\qquad f(x) = \Pr(\text{of getting x heads in n tosses}) = \frac{n!}{x!(n-x)!}\Big/2^n = \frac{n!}{x!(n-x)!}\left(\frac{1}{2}\right)^n.$$

For n = 2, 3, 4 and 5, f(x) has the values given in Table 4.1.

TABLE 4.1

Probability of Getting x Heads in n Tosses of a Coin for n = 2, 3, 4, 5.

x	f(x)			
	n = 2	n = 3	n = 4	n = 5
0	$1\cdot(\frac{1}{2})^2$	$1\cdot(\frac{1}{2})^3$	$1\cdot(\frac{1}{2})^4$	$1\cdot(\frac{1}{2})^5$
1	$2\cdot(\frac{1}{2})^2$	$3\cdot(\frac{1}{2})^3$	$4\cdot(\frac{1}{2})^4$	$5\cdot(\frac{1}{2})^5$
2	$1\cdot(\frac{1}{2})^2$	$3\cdot(\frac{1}{2})^3$	$6\cdot(\frac{1}{2})^4$	$10\cdot(\frac{1}{2})^5$
3		$1\cdot(\frac{1}{2})^3$	$4\cdot(\frac{1}{2})^4$	$10\cdot(\frac{1}{2})^5$
4			$1\cdot(\frac{1}{2})^4$	$5\cdot(\frac{1}{2})^5$
5				$1\cdot(\frac{1}{2})^5$

4.41 Binomial coefficients.

The quantities C_x^n are also called binomial coefficients, for they may be obtained by expanding a binomial expression raised to the n-th power. Thus, consider the binomial expression $(T + H)^n$. We have

$$(T+H)^n = T^n + C_1^n H^1 T^{n-1} + C_2^n H^2 T^{n-2} + C_3^n H^3 T^{n-3} + \ldots + C_x^n H^x T^{n-x} + \ldots + H^n. \tag{4.7}$$

Note that the general term in this expansion is $C_x^n H^x T^{n-x}$; if we call H the head of a coin and T the tail, then we may regard this term as telling us symbolically that "there are C_x^n ways in which n coins can fall so that there are x heads and n-x tails". For x = 0 we have the first term T^n, which states that there is only one way the n coins can fall so that we get 0 heads and n tails; for x = 1, we have the second term $C_1^n H^1 T^{n-1}$ which states that the n coins can fall in C_1^n ways so that we get 1 head and n-1 tails; and so on.

If we put $H = \frac{1}{2}$, and $T = \frac{1}{2}$, we get

$$(\tfrac{1}{2} + \tfrac{1}{2})^n = (\tfrac{1}{2})^n + C_1^n \cdot (\tfrac{1}{2})^1 \cdot (\tfrac{1}{2})^{n-1} + C_2^n \cdot (\tfrac{1}{2})^2 \cdot (\tfrac{1}{2})^{n-2} + \ldots + C_x^n \cdot (\tfrac{1}{2})^x (\tfrac{1}{2})^{n-x} + \ldots + (\tfrac{1}{2})^n.$$

The general term $C_x^n (\frac{1}{2})^x (\frac{1}{2})^{n-x}$ now gives us the probability that x heads (and n-x tails) will appear in throwing n coins. The first term $(\frac{1}{2})^n$ is the probability of getting 0 heads and n tails; the second term $C_1^n (\frac{1}{2}) (\frac{1}{2})^{n-1}$ is the probability of getting 1 head and n-1 tails; and so on.

Table 4.2 shows the values of C_x^n for all values of n from 0 to 10: note that the entry for any given row can be constructed from the row immediately above it by adding the number above that entry to the number's left-hand neighbor. For example, the entry 126 (n = 9, x = 4) is found by adding 70 and 56.

TABLE 4.2

Binomial Coefficients C_x^n for n = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

	x											
n	0	1	2	3	4	5	6	7	8	9	10	Total
0	1											1
1	1	1										2
2	1	2	1									2^2
3	1	3	3	1								2^3
4	1	4	6	4	1							2^4
5	1	5	10	10	5	1						2^5
6	1	6	15	20	15	6	1					2^6
7	1	7	21	35	35	21	7	1				2^7
8	1	8	28	56	70	56	28	8	1			2^8
9	1	9	36	84	126	126	84	36	9	1		2^9
10	1	10	45	120	210	252	210	120	45	10	1	2^{10}

Exercise 4.4.

1. In how many ways can the members of a committee of three be chosen from 7 men?

2. In how many ways can a committee of three men and two women be selected from 5 men and 3 women? Solve this problem by formula first and then verify your results by enumeration.

3. If 5 coins are tossed, what is the probability of getting 0, 1, ..., 5 heads respectively? What is the probability of getting less than 3 heads? Of getting 3 or more? Of getting at least 2 heads and at least 2 tails?

4. If we want to pick r objects from n objects so that one particular object must always be included, in how many ways can this be done? Answer the same question replacing "included" by "excluded".

5. Generalize the previous problem: In how many ways can we pick r objects from

n objects so that $s (s < r)$ particular objects must always be included? So that s particular objects must always be excluded?

6. In how many ways can 17 coins fall so that exactly 10 heads show? So that at least 3 heads show?

7. Show that $C_x^n = C_{n-x}^n$.

8. Draw a frequency histogram for the number of ways of getting 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 heads in throwing 10 coins.

9. What number of heads can be obtained in the most ways when throwing 42 coins? 43 coins? n coins if n is even? n coins if n is odd?

10. By looking at Table 4.2, write down the values of C_x^{11} for x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11.

11. In how many different ways can a hand of 13 cards be selected from a pack of 52 playing cards?

12. How many different luncheons consisting of soup, an entrée, two vegetables, dessert and beverage can be ordered from a menu which lists 3 soups, 5 entrées, 4 vegetables, 7 desserts and 3 beverages?

4.5 Calculation of Probabilities.

There are several simple rules for calculating probabilities which will be stated and illustrated in this section.

If E, F, ... are events, we often want to know the probabilities of the following events derived from them:

the event "not E" (i.e., "E does not occur")
the event "E or F" (i.e., "either E or F or both occur")
the event "E and F" (i.e., "both E and F occur").

For example, if we roll a die, E may be the event that "we get a 6", F may be the event that "we get a 5". Then the event "not E" is simply that of not getting 6, i.e., any of the faces 1 to 5. The event "E or F" is that of

"getting either 5 or 6". In this example the event "E and F" is impossible, because we cannot get 5 and 6 at the same time. The following example will illustrate this case better. Suppose we roll a die twice. Let E now denote "the event that we get 6 in the first roll" and F that "we get 6 in the second roll". Then "E and F" is possible, and means we get 6 in both the first roll and the second roll.

4.51 Complementation.

Rule I: Complementation. Suppose E is an event. Then

(4.8) $$\Pr(\text{not } E) = 1 - \Pr(E).$$

For if there are n equally likely cases and E occurs in m cases, then the event "not E" occurs in the remaining n - m cases. Hence

$$\Pr(\text{not } E) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - \Pr(E).$$

To illustrate the application of Rule I, let us consider an

Example: Three coins are tossed. What is the probability of getting at least one head?

If we denote the event of "getting at least one head" by E, then "not E" is the event of "getting 0 heads" (i.e., 3 tails). Hence by Rule I

$$\begin{aligned}\Pr(\text{getting at least one head}) &= 1 - \Pr(\text{getting 0 heads}) \\ &= 1 - \Pr(\text{getting 3 tails}) \\ &= 1 - \frac{1}{8} = \frac{7}{8}.\end{aligned}$$

4.52 Addition of probabilities for mutually exclusive events.

Rule II: Addition. If two events E and F are mutually exclusive (i.e., they cannot occur together), then

(4.9) $$\Pr(E \text{ or } F) = \Pr(E) + \Pr(F).$$

For, suppose there are n equally likely cases such that E occurs in r of them, and F occurs in s of them. Then, since E and F do not occur together, there is no overlapping between these r + s cases. The event "E or F" occurs in these r + s cases and these only. Hence

$$\Pr(E \text{ or } F) = \frac{r+s}{n} = \frac{r}{n} + \frac{s}{n} = \Pr(E) + \Pr(F).$$

Similarly we see that if there are several events, E, F, G, ..., mutually exclusive, we have

(4.10) $$Pr(E \text{ or } F \text{ or } G \text{ or } \ldots) = Pr(E) + Pr(F) + Pr(G) + \ldots$$

Let us illustrate Rule II by the following

Example: Suppose a card is drawn at random from a pack of playing cards. What is the probability that it is either a "heart" or the "queen of spades"?

E is the event "heart" - F is the event "the queen of spades". They are mutually exclusive because a heart cannot be the queen of spades. Thus Rule II is applicable. Since there are 13 hearts

$$Pr(E) = \frac{13}{52} = \frac{1}{4}.$$

Since there is just one "queen of spades",

$$Pr(F) = \frac{1}{52} .$$

Hence

$$Pr(E \text{ or } F) = Pr(E) + Pr(F) = \frac{1}{4} + \frac{1}{52} = \frac{14}{52}.$$

4.53 Multiplication of probabilities for independent events.

Now suppose there are n equally likely cases and m favorable cases for the event E; N equally likely cases and M favorable cases for the event F, so that

$$Pr(E) = \frac{m}{n}$$

$$Pr(F) = \frac{M}{N} .$$

If each of the n cases when we consider E can be combined with each of the N cases when we consider F, so that it can be agreed that the nN combined cases for the joint event "E and F" can be considered equally likely, then we say that the two events E and F are independent.

Rule III: Multiplication. If E and F are independent in the sense just defined, then

(4.11) $$Pr(E \text{ and } F) = Pr(E) \cdot Pr(F) .$$

Since each of the n cases when considering E can be combined with each

of the N cases when considering F, we have altogether nN possible cases when we consider the joint event "E and F". Each favorable case for E combines with each favorable case for F to give a favorable case for the joint event "E and F". Thus, there are mM favorable cases for the event "E and F". Hence

$$\Pr(\text{E and F}) = \frac{mM}{nN} = (\frac{m}{n})\cdot(\frac{M}{N}) = \Pr(E)\cdot\Pr(F) .$$

Similarly, we see that if there are several events, E, F, G, ..., all independent, we have

$$\Pr(\text{E and F and G and } \ldots) = \Pr(E)\cdot\Pr(F)\cdot\Pr(G) \ldots . \tag{4.12}$$

Rule III can be illustrated by the following

Example: From a pack of playing cards two cards are drawn at random successively, the first being replaced before the second is drawn. What is the probability that the first is a "heart" and the second not a "king"?

E is the event: "The first card drawn is a heart"; F is the event: "The second is not a king". We have

$$\Pr(E) = \frac{13}{52} = \frac{1}{4} .$$

Notice that the event F is of the form "not king". Since there are 4 kings, the probability of "king" is $\frac{4}{52}$; by the rule of complementation,

$$\Pr(F) = \Pr(\text{not king}) = 1 - \Pr(\text{king}) = 1 - \frac{4}{52} = \frac{12}{13} .$$

Since the first card is replaced before the second drawing, any possibility for the first drawing can be combined with any possibility for the second. Thus, we can apply Rule III and obtain

$$\Pr(\text{E and F}) = \Pr(E)\cdot\Pr(F) = \frac{1}{4} \times \frac{12}{13} = \frac{3}{13} .$$

Sometimes we have to apply a rule similar to Rule III in evaluating the number of favorable cases in making a probability calculation; in other words, we apply a multiplication procedure in evaluating the number of favorable cases. An example will make this clear.

Example: In a game of bridge what is the probability of a hand of 13 cards consisting of 5 spades, 3 hearts, 3 diamonds and 2 clubs?

The total number of cases n is the number of ways of picking 13 cards from 52. This is a problem in combinations. We have

$$n = C_{13}^{52} .$$

The number of favorable cases m is obtained as follows: We want to pick 5 spades out of 13 spades; the number of ways of doing this is C_5^{13} by the Rule on number of combinations in Section 4.4. Similarly, the number of ways of picking 3 hearts out of 13 is C_3^{13}, and the number of ways of picking 2 clubs out of 13 is C_2^{13}. The combined number of ways of doing all these things jointly is obtained by multiplying the four numbers, i.e.,

$$m = C_5^{13} \cdot C_3^{13} \cdot C_3^{13} \cdot C_2^{13} .$$

This is the number of favorable cases m. Thus applying the definition of probability, we have

$$\Pr(5 \text{ spades, } 3 \text{ hearts, } 3 \text{ clubs, } 2 \text{ spades}) = \frac{m}{n} = \frac{C_5^{13} \cdot C_3^{13} \cdot C_3^{13} \cdot C_2^{13}}{C_{13}^{52}}$$

$$= \frac{\frac{13!}{5!\,8!} \cdot \frac{13!}{3!\,10!} \cdot \frac{13!}{3!\,10!} \cdot \frac{13!}{2!\,11!}}{\frac{52!}{39!\,13!}} = \frac{39!\,(13!)^5}{2!\,(3!)^2\,5!\,8!\,(10!)^2\,11!\,52!}$$

4.54 Multiplication of probabilities when events are not independent; conditional probabilities.

We often have situations in which we have to calculate the probability of the joint occurrence of two events E and F when they are not independent. If a trial can result in the occurrence of E or "not E" and can also result in the occurrence of F or "not F", then the result of a trial will belong to one and only one of the four classes: E and F; E and "not F"; "not E" and F; "not E" and "not F". If n_{11}, n_{12}, n_{21}, n_{22} are the numbers of cases favorable to these four classes respectively, where $n_{11} + n_{12} + n_{21} + n_{22} = n$, the total number of possible cases, we can display the situation in the four-fold table as shown in Table 4.3.

Now the probability of the occurrence of E and F is

$$\Pr(E \text{ and } F) = \frac{n_{11}}{n} .$$

But notice that this can be written as

$$\Pr(E \text{ and } F) = \left(\frac{n_{11} + n_{21}}{n}\right)\cdot\left(\frac{n_{11}}{n_{11} + n_{21}}\right) .$$

Now $n_{11}/(n_{11} + n_{21})$ is the probability of event E, assuming that event F has occurred. We write this as $\Pr(E|F)$, read "probability of E given F", and is called the conditional probability of the occurrence of E, given that F has occurred. The ratio $(n_{11} + n_{21})/n$ is clearly the probability of the occurrence of F, i.e., $\Pr(F)$. Therefore, we have a rule for multiplication when the events are not independent.

TABLE 4.3

Four-fold Table

	F	Not F	Total
E	n_{11}	n_{12}	$n_{11}+n_{12}$
Not E	n_{21}	n_{22}	$n_{21}+n_{22}$
Total	$n_{11}+n_{21}$	$n_{12}+n_{22}$	n

Rule IV: Multiplication when events are not independent.

If E and F are not independent, then the probability of the joint occurrence of E and F is given by the formula

(4.13) $$\Pr(E \text{ and } F) = \Pr(F)\cdot\Pr(E|F) .$$

It should be noted that if $\Pr(E|F) = \Pr(E|\text{"not F"})$ then each of these quantities will be equal to $\Pr(E)$ (i.e., to $(n_{11}+n_{12})/n$). In this case the numbers in the columns (and also rows) in Table 4.3 will be proportional to each other and we have a situation in which we have independence, and in which Rule IV reduces to Rule III.

Similarly, if we have three events E, F, G, which are not independent, we have the formula

(4.14) $$\Pr(E \text{ and } F \text{ and } G) = \Pr(G)\cdot\Pr(F|G)\cdot\Pr(E|F \text{ and } G) .$$

This can be extended to any number of events.

If we return to the formula given in Rule IV, we can write (assuming Pr(F) is not zero)

(4.15) $$\Pr(E|F) = \frac{\Pr(E \text{ and } F)}{\Pr(F)}$$

which gives us a way of calculating conditional probabilities. We shall consider an example to illustrate Rule IV.

Example: Suppose three bad light bulbs get mixed up with 12 good ones, and that you start testing the bulbs one by one until you have found all three defectives. What is the probability that you will find the last defective on the seventh testing?

Let F be the event of "finding 2 defectives among the first 6 tested" and E be the event of "finding the third defective on the seventh testing". Now finding Pr(F) is just a combinatorial problem, i.e.,

$$\Pr(F) = \frac{C_2^3 \cdot C_4^{12}}{C_6^{15}} .$$

Pr(E|F) is the probability of finding the third defective on the seventh testing after event F has happened. When F has occurred, we know that there are 9 bulbs left and that one is defective. The probability of picking it on the seventh testing (the first after F has occurred) is the desired probability Pr(E|F), i.e.,

$$\Pr(E|F) = \frac{1}{9} .$$

Hence

$$\Pr(E \text{ and } F) = \frac{C_2^3 \cdot C_4^{12}}{C_6^{15}} \times \frac{1}{9} = \frac{3}{91} ,$$

which is the desired probability.

4.55 Addition of probabilities when events are not mutually exclusive.

There are situations in which we need to find the probability of the occurrence of "E or F" when E and F are not mutually exclusive. In this case we have the following rule:

Rule V: Addition when events are not mutually exclusive.

If E and F are events which are not mutually exclusive, then the probability of the occurrence of E or F is given by the formula

$$\Pr(E \text{ or } F) = \Pr(E) + \Pr(F) - \Pr(E \text{ and } F). \tag{4.16}$$

The truth of this rule is clear if we look at Table 4.3. Remember that the event "E or F" means that "either E or F or both occur". The number of cases in Table 4.3 in which "either E or F or both occur" is seen to be $n_{11} + n_{12} + n_{21}$. Hence

$$\Pr(E \text{ or } F) = \frac{n_{11}+n_{12}+n_{21}}{n}$$

$$= \frac{n_{11}+n_{12}}{n} + \frac{n_{11}+n_{21}}{n} - \frac{n_{11}}{n}$$

$$= \Pr(E) + \Pr(F) - \Pr(E \text{ and } F).$$

Notice that n_{11} is the number of cases in which both E and F occur. Hence, if $\Pr(E \text{ and } F) = 0$, (i.e., $n_{11} = 0$), then E and F would be mutually exclusive. Therefore Rule V would reduce to

$$\Pr(E \text{ or } F) = \Pr(E) + \Pr(F)$$

which is Rule II.

In the case of three events E, F, G, which are not mutually exclusive, the event "E or F or G" means that "either E or F or G, any two of them or all three occur". In this case Rule V becomes

$$\begin{aligned}\Pr(E \text{ or } F \text{ or } G) = {} & \Pr(E) + \Pr(F) + \Pr(G) \\ & - \Pr(E \text{ and } F) - \Pr(E \text{ and } G) - \Pr(F \text{ and } G) \\ & + \Pr(E \text{ and } F \text{ and } G).\end{aligned} \tag{4.17}$$

To illustrate Rule V, let us consider the following

Example: If a card is dealt from a pack, what is the probability it will be an honor card or a spade?

Let E be the event "an honor card" and F the event "a spade", then "E or F" is the event "an honor card or a spade or both" and "E and F" is the event "an honor card and a spade". We have

$$\Pr(E) = \frac{20}{52}, \quad \Pr(F) = \frac{13}{52}, \quad \Pr(E \text{ and } F) = \frac{5}{52},$$

and therefore

$$\Pr(E \text{ or } F) = \frac{20}{52} + \frac{13}{52} - \frac{5}{52}$$

$$= \frac{28}{52} .$$

The numbers of cases involved in this problem may be represented in a table similar to Table 4.3 as follows:

TABLE 4.4

Table Showing Classification of Cards in a Pack with Respect to Honor Cards and Spades

	E (an honor card)	Not E (not an honor card)	Total
F (a spade)	5	8	13
Not F (not a spade)	15	24	39
Total	20	32	52

4.56 Euler diagrams.

An effective schematic representation of the various combinations of events and their complements, and the probabilities (or percentages of cases) associated with them in case the events are not mutually exclusive, is provided by Euler diagrams. For convenience let us use the following shortened notation:

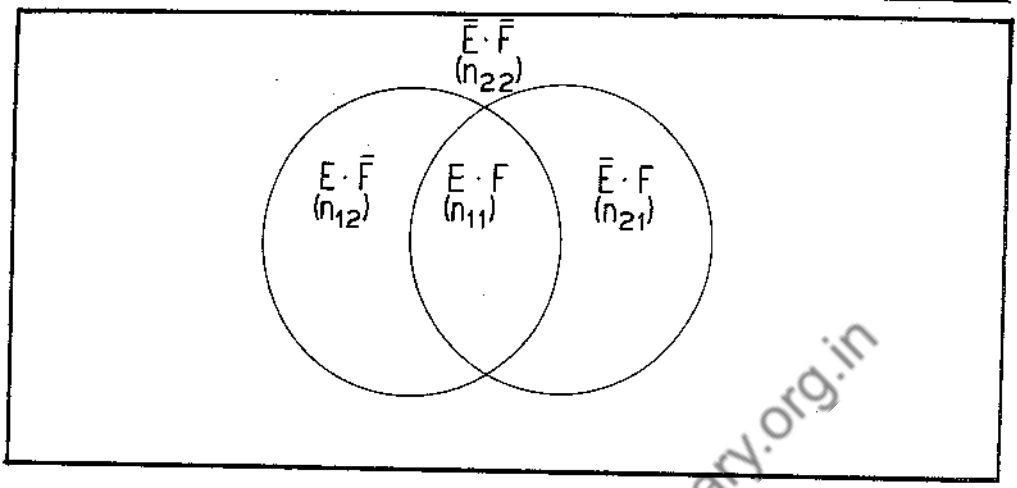
$\overline{E}$ = "not E"

E+F = "E or F"

E·F = "E and F"

with similar meanings for $\overline{F}$, E + $\overline{F}$, E·$\overline{F}$, etc. In this notation we sometimes refer to E, $\overline{E}$, E+F, E·F, etc., as classes as well as events.

We may then represent the various possible combinations of events given in Table 4.3 by the Euler diagram in Figure 4.1.



Euler Diagram for the Information in Table 4.3

Figure 4.1

The possible events are $E\cdot F$, $E\cdot\bar{F}$, $\bar{E}\cdot F$ and $\bar{E}\cdot\bar{F}$ and they are represented by the regions into which the rectangle is divided. The numbers of cases favorable to these various events are given in the parentheses. We may then write relations among the various events in an algebraic form:

$$E = E\cdot F + E\cdot\bar{F}$$
$$F = E\cdot F + \bar{E}\cdot F$$
$$E + F = E\cdot F + E\cdot\bar{F} + \bar{E}\cdot F$$

from which we can write down the probabilities:

$$\Pr(E) = \frac{n_{11} + n_{12}}{n}, \quad \Pr(E\cdot F) = \frac{n_{11}}{n}$$

$$\Pr(F) = \frac{n_{11} + n_{21}}{n}$$

$$\Pr(E+F) = \frac{n_{11} + n_{12} + n_{21}}{n}$$

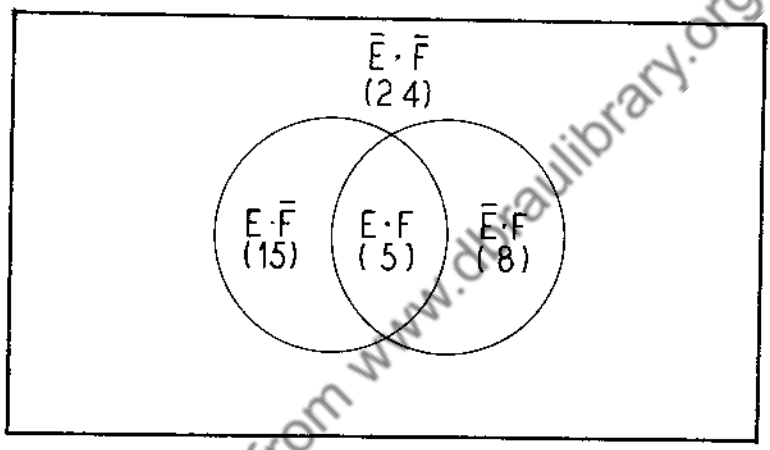
$$= \frac{n_{11} + n_{12}}{n} + \frac{n_{11} + n_{21}}{n} - \frac{n_{11}}{n},$$

or more briefly,

$$Pr(E+F) = Pr(E) + Pr(F) - Pr(E \cdot F) ,$$

which is simply another way of writing formula (4.16) using the shortened notation.

The Euler diagram for the example illustrating Rule V is shown in Figure 4.2. In the Diagram $E \cdot F$ is the class of cards in which each card is an E (honor card) and an F (spade). $E \cdot F$ also refers to the event of drawing a card which is an E and an F. Similar interpretations apply to $E \cdot \overline{F}$, $\overline{E} \cdot F$, and $\overline{E} \cdot \overline{F}$.



Euler Diagram for the Information in Table 4.4

Figure 4.2

The Euler diagram is more useful in the case of three events E, F, G and their complements $\overline{E}$, $\overline{F}$ and $\overline{G}$. For we have in this case the Euler diagram shown in Figure 4.3.

Algebraically, we may write

$$E = E \cdot F \cdot G + E \cdot F \cdot \overline{G} + E \cdot \overline{F} \cdot G + E \cdot \overline{F} \cdot \overline{G}$$

and

$$Pr(E) = \frac{n_{111} + n_{112} + n_{121} + n_{122}}{n} ,$$

with similar expressions for F, G, Pr(F) and Pr(G). (Note that event E is

represented by all regions in the circle marked (E), and similarly for event F and event G.)

For the event "E or F" (i.e., E+F) we take all regions falling in the two circles marked (E) and (F), i.e.,

$$E + F = E\cdot F\cdot G + E\cdot F\cdot \bar{G} + E\cdot \bar{F}\cdot G + E\cdot \bar{F}\cdot \bar{G}$$

$$+ \bar{E}\cdot F\cdot G + \bar{E}\cdot F\cdot \bar{G} ,$$

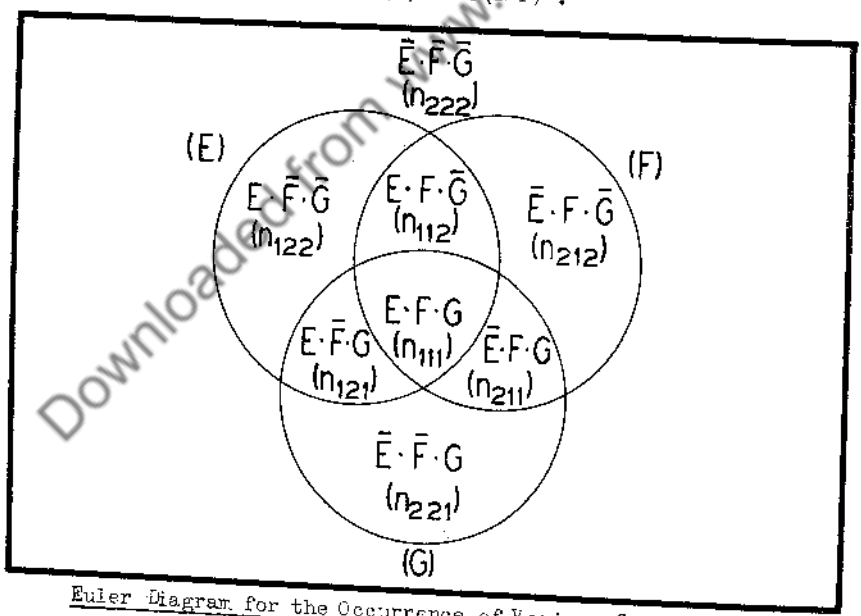
and we have

$$\Pr(E+F) = \frac{n_{111} + n_{112} + n_{121} + n_{122} + n_{211} + n_{212}}{n}$$

$$= \frac{(n_{111} + n_{112} + n_{121} + n_{122})}{n} + \frac{(n_{111} + n_{112} + n_{211} + n_{212})}{n}$$

$$-\frac{(n_{112} + n_{111})}{n}$$

$$= \Pr(E) + \Pr(F) - \Pr(E\cdot F) .$$



Euler Diagram for the Occurrence of Various Combinations of Events E, F, G and their Complements

Figure 4.3

Similarly, the event E + F + G is represented by the "sum" of all regions included anywhere in one or more of the three circles. By writing down the n's and grouping them properly we find that

$$\begin{aligned} Pr(E+F+G) = {} & Pr(E) + Pr(F) + Pr(G) \\ & - Pr(E\cdot F) - Pr(E\cdot G) - Pr(F\cdot G) \\ & + Pr(E\cdot F\cdot G) , \end{aligned}$$

which is formula (4.17) written in the shorter notation.

In applications, it is usually convenient to refer to the event or class $E\cdot\overline{F}\cdot\overline{G}$ as "E only" and $E\cdot F\cdot\overline{G}$ as "E and F only", and similarly for other events or classes of these types. The event or class E, it must be remembered, consists of all four classes inside the circle marked (E).

Consider the following example involving an Euler diagram with three classes and their complements:

Example: Among the children in a certain school

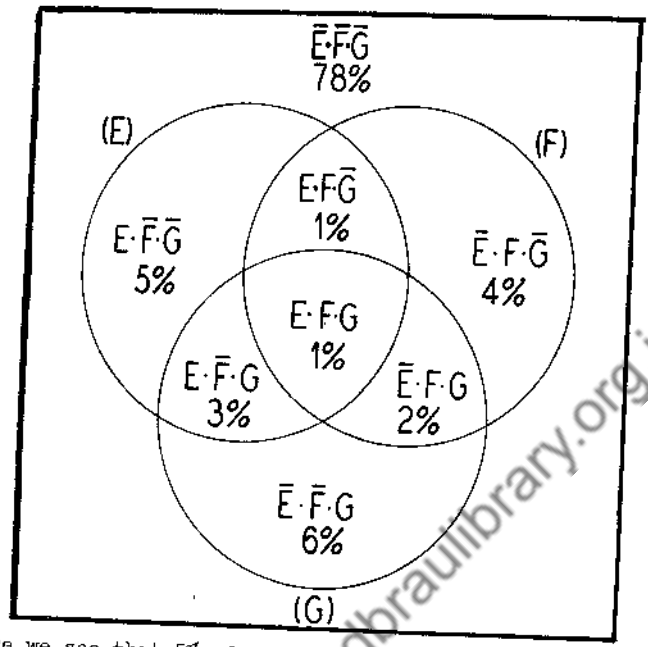
10% have defective eyes
8% have defective hearing
12% have defective teeth
4% have defective eyes and teeth
3% have defective hearing and teeth
2% have defective eyes and hearing
1% have defective eyes and hearing and teeth.

Construct an Euler diagram for these data.

Let us refer to defective eyes as class E, defective hearing as class F and defective teeth as class G. Then we have

10% belong to E
8% belong to F
12% belong to G
4% belong to E·G
3% belong to F·G
2% belong to E·F
1% belong to E·F·G

The Euler diagram is as follows :



From this figure we see that 5% of the children belong to $E\cdot\bar{F}\cdot\bar{G}$, i.e. have defective eyes only (do not have defective hearing or teeth). Similar interpretations hold for other classes.

4.57 General remarks about calculating probabilities.

In many probability problems such as those in poker, bridge and other card games, it is often simpler to go back to the definition of probability (Definition I) and compute the numbers of favorable and possible cases as in the original definition of probability, than to try to make numerous specific applications of the foregoing rules. Let us consider a typical example.

Example: What is the probability of getting "four of a kind" in a poker hand?

This means that of the five cards in a poker hand four are of the same kind and the remaining one arbitrary. There are 13 different "kinds", and we can pick any one of them. For the remaining card we have 52 - 4 = 48 possibilities. Hence

$$m = 13\cdot 48$$

and

$$n = C_5^{52}.$$

Hence the required probability is

$$\frac{m}{n} = \frac{13\ 48}{C_5^{52}} \quad \frac{13\ 48\ (5!)\ (47!)}{52!} = .00024\ .$$

Exercise 4.5.

(In a pack of playing cards, the honor cards are understood to be the ace, king, queen, jack and ten of each suit).

1. If a die is rolled twice, what is the probability that the first roll yields a 3 or 4, the second anything but 3?

2. A bag contains 6 black and 4 white balls. Three balls are drawn. What is the probability that 2 are black and 1 is white? That 1 is black and 2 are white?

3. If four people stand in a row, what is the chance that 2 particular persons designated in advance are next to each other? That they are not next to each other?

4. What is the probability that a bridge hand will consist of 5 spades, 4 clubs, 3 diamonds, 1 heart? Of 2 aces, a king, a queen, 2 tens and no other honor cards?

5. In a roll of 6 dice, what is the probability of getting exactly 5 faces alike?

6. If we draw two cards from a pack of playing cards, what is the probability that they form a pair (two cards with the same number)? If we draw two cards twice, the first two being replaced before the second two are drawn, what is the probability that we get a pair in the first drawing and exactly one ace in the second?

7. What is the probability of getting a "full house" (three cards of one kind and two of another) in a poker hand?

8. A bag contains 3 white and 2 black balls; another contains 2 white and 1 black ball. If we choose a bag at random and then draw a ball from it, what is the probability of getting a white ball? If all balls are poured into one bag and a ball is drawn, what is the probability of getting a white ball?

9. Six men and their wives play three tables of bridge. If the men are paired with the women by drawing score cards, what is the probability that each man will

draw his wife for his partner?

10. What is the probability that the birthdays of three particular students will fall on three different days of the year? That at least two of them will have birthdays on the same day of the year?

11. The probability of a man aged 60 dying within one year is .025, and the probability of a woman aged 55 dying within one year is .011. If a man and his wife are 60 and 55 respectively, what is the probability of their both living a year? Of at least one of them dying within a year? Of at least one of them living a year?

12. A buyer will accept a lot of 100 articles if a sample of 5 picked at random and inspected contains no defectives. What is the probability that he will accept the lot if it contains 10 defective articles?

13. A man has two Indian pennies and two Lincoln pennies in each of two pockets. Under which of the following conditions is the probability of getting two Lincoln pennies highest: (a) draw one penny from each pocket, (b) draw two pennies from one pocket, or (c) put all 8 pennies in one pocket and draw two pennies? Work out the probability for each case.

14. If we draw 3 cards from a pack, each time replacing the card drawn before the next drawing, what is the probability that at least one of the cards drawn is a spade?

15. If a person starts dealing off cards from a pack of playing cards, what is the probability that the 48th card dealt is the last red one? That the xth card dealt is the yth heart to be dealt?

16. If a card is drawn from a pack of playing cards, what is the probability it will be "an ace or a spade or an honor card"?

17. A bowl of 25 beads painted with red and green luminous and non-luminous paint has the following composition:

	Luminous	Non-Luminous	Total
Red	6	12	18
Green	1	6	7
Total	7	18	25

(a) If a bead is drawn, what is the probability it is green or luminous or both?

(b) If a bead is drawn at night and is seen to be non-luminous, what is the probability it is red?

(c) If a bead is drawn in daylight and is noted to be red, what is the probability it is non-luminous?

(d) Make an Euler diagram for the composition of beads in the bowl.

18. In a certain population, the percentages of persons reading magazines A, B, C and various combinations are as follows:

A : 9.8% A and B : 5.1%
B : 22.9% A and C : 3.7%
C : 12.1% B and C : 6.0%
A and B and C : 2.4%

(a) Make an Euler diagram from these data.

(b) What percent of the population read at least one of the 3 magazines?

(c) What is the probability that a person taken at random from this population would be a reader of A or B or both?

(d) Of those persons reading at least one of the magazines, what percentage read at least two magazines?

(e) What percentage of the population read:
A only; A and B only;
B only; A and C only;
C only; B and C only?

4.6 Mathematical Expectation.

Suppose $E_1, E_2, \ldots, E_k$ are mutually exclusive and exhaustive events, and that $P_1, P_2, \ldots, P_k$ are the respective probabilities of these events occurring. Suppose a person, A, receives an amount of money M_1 if E_1 happens, an amount M_2 if E_2 happens, ..., an amount M_k if E_k happens. Then A's mathematical expectation of gain or winnings E(M) is defined as

$$E(M) = M_1P_1 + M_2P_2 + \ldots + M_kP_k .$$

If A pays an amount of money equal to $M_1P_1 + M_2P_2 + \ldots + M_kP_k$, then we say A pays a fair price for this expectation. This concept of mathematical expectation and fair price enters into all gambling systems and insurance plans. More specifically, if any system of gambling is looked at from the point of view of an individual gambler, we find him paying a sum of money for the privilege of receiving a sum of money or nothing, depending on the outcome of the game. For example, an Irishman buys a lottery ticket for 10 shillings, knowing that the probability of getting a large prize of 30,000 pounds is very small and the probability of getting nothing is very large. A person aged 21 pays an $16 premium to a life insurance company for a one-year term insurance policy in return for a guarantee that the company will pay his beneficiary $1000 if he dies during the year (and nothing if he lives!).

Let us illustrate by the following

Example: Suppose A tosses two coins and receives 2 dollars from B if two heads appear, 1 dollar if one head appears and nothing if no heads appear. How much should A pay B in advance for the privilege of playing this game?

It will be seen that three events are involved: the event of 2 heads, the event of 1 head and the event of 0 heads. The probabilities of these events are 1/4, 1/2 and 1/4, respectively. If this game were played a large number of times, A would receive 2 dollars in about 1/4 of the trials, 1 dollar in about 1/2 of the trials and nothing in the remaining trials. Thus he would receive in the long run an average of $2 \times 1/4 = 1/2$ dollar per trial from the first kind of event (2 heads), an average of $1 \times 1/2 = 1/2$ dollar for the second kind of event (1 head), and nothing from the third. Or his average winnings per trial in the long run are

$$2 \times \frac{1}{4} + 1 \times \frac{1}{2} = 1 \text{ dollar.}$$

The mathematical expectation of A's gain or winnings is 1 dollar. A should pay B 1 dollar for the privilege of playing this game if it is to be a fair game.

Exercise 4.6.

1. If A received 1 cent for every dot that appears in throwing two dice, what would be the fair price for playing this game?

2. A throws four of his pennies. If he obtains more than two heads he receives a dime for each head and also keeps his pennies. Otherwise he forfeits his

pennies to B. How much should A pay B for every play in order to make this game fair?

3. A pays B one dollar and three dice are rolled. A receives 2 dollars if an ace appears, 4 dollars if 2 aces appear, and 8 dollars if 3 aces appear; otherwise he gets nothing. Is this a fair game? If not, how much should A receive for 3 aces to make it fair?

4. How much should a person pay to receive one dollar if he gets a face card at least once in cutting a deck of cards 3 times and nothing otherwise?

5. The probability that a man aged 50 will live another year is .988. How large a premium should he pay for a $1000 term life insurance policy for one year (ignoring insurance company charges for administration, etc.)?

6. Suppose A bets you 5 cents against x cents that two persons picked at random will have birthdays in the same month. Assuming a birthday to be equally likely to fall in any month, what should be the value of x for this to be a fair bet?

4.7 Geometric Probability.

There are elementary probability problems involving position of an object in a region which require an extension of Definition 1 of Section 4.1.

For example, suppose a box 12 inches square has a bottom made of 3-inch boards such that 3 cracks between boards are visible. If a penny is put in the box and shaken around, what is the probability that when the shaking stops, the penny will not overlap a crack?

In this problem, consider E as the event of not overlapping a crack. We consider the center of the penny as a point and find the areas of two regions: (1) the area C of the region R in which the center of the penny could fall, and (2) the area C_E of the region R_E in which the center of the penny must lie so that no part of the penny will overlap a crack. Then if it could be agreed that it is equally likely for the center of the penny to fall anywhere in C, we would define the probability of event E as the ratio of area C_E to area C. The radius of a penny is 3/8 ". The region R in which the center of the penny can fall is a square 11 1/4" on a side; its area C is 126 9/16 sq. inch. The region R_E consists of 4 rectangles $2\frac{1}{4}$ by $11\frac{1}{4}$ inches; its area is $101\frac{1}{4}$ sq. inch. The

probability of the occurrence of E is therefore $\frac{101.25}{126.5625} = .8$.

We could give other two-dimensional examples in geometric probability. Also one-dimensional and three-dimensional examples. These suggest that we can usefully make the following extension of Definition I to cover situations in geometric probability. In the definition we will consider the word content to mean length, area or volume, of a region depending on whether the definition is applied to problems involving one-, two-or three-dimensional regions.

Definition of Geometric Probability. If an event E can happen by the occurrence of a point in a region R_E within a region R, all points in R being considered by mutual agreement to be equally likely, then the probability of the event E is defined as C_E/C, or

$$Pr(E) = C_E/C$$

where C_E is the content of R_E and C is the content of R.

We can apply all of the rules of probability discussed in Section 4.4 to geometric probability.

Exercise 4.7.

1. Suppose you drop a penny on the tiled floor of a 10' by 10' room, the tiles being one-inch black and white squares, arranged in checkerboard pattern, with no mortar between tiles. What is the probability that the penny will

(a) lie completely within some tile?

(b) overlap parts of at least two tiles?

(c) overlap parts of four tiles?

2. If a needle two inches long is dropped on a floor and intersects a crack between two floor boards, what is the probability the needle would be "cut" by the crack less than one eighth of an inch from its center? If 10 throws of the needle are taken in which the needle intersects a crack, what is the probability that the needle is never "cut" within its middle third?

3. Suppose a bomb dropped from high altitude is equally likely to hit any part of a factory, if it hits it at all. What is the probability that if a bomb hits a factory covering 100,000 square feet of ground, it will hit the power plant which covers 2000 square feet of ground? If 10 bombs hit the factory, what is

the probability of at least one hit on the power plant? How many hits on the factory would be required in order to have a probability of 0.9 of getting at least one hit on the power plant?

4. Suppose 52 barrage balloons are moored at altitudes of 5000 feet by cables, and that they are arranged in a straight line, the balloons being spaced 1000 feet apart. If a plane having a wing span of 200 feet flies through this barrage at night, what is the probability it hits a cable

(a) if it flies through the plane of the cables at right angles?

(b) if it flies (level flight) through the plane of the cables at 30° (to the plane of cables)?

What is the probability that three planes will fly through the barrage as in (a) without striking cables? As in (b) without striking cables?

5. A petty gambling joint at a county fair operates as follows: the top of a table is solidly covered with 200 Lucky Strike cigarette packages in 10 rows of 20 packages each. Each package measures 2.25 by 2.75 inches and has a red circle 1.5 inches in diameter. A player pitches a penny onto these cigarette packages. If the penny falls inside a red circle, he wins a package of cigarettes and gets his penny back. Otherwise he loses his penny. Assuming that no pennies slide off the table or overlap its edges and that a penny is equally likely to stop anywhere on the mat of cigarette packages and that Lucky Strikes are worth 16 cents a package, is this a fair game? Who has the advantage and what is his expectation of winnings per 100 throws?

CHAPTER 5. PROBABILITY DISTRIBUTIONS.

5.1 Discrete Probability Distributions.

5.11 Probability tables and graphs.

In probability problems we often find that what we are really interested in are the probabilities for a set of events, where the events can be simply described by numbers. For example, in rolling a pair of dice, we are usually interested in the various total numbers of dots it is possible to get and in their probabilities. In other words, we are interested in the probabilities of getting a total of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 dots. These numbers can be considered as possible values of the chance quantity X, where X denotes the total number of dots obtained in rolling two dice. For any one of these values which X can take on, we would have a probability. In general, if we call $f(x)$ the probability that $X = x$ (i.e., of getting x dots), we can call $f(x)$ the probability distribution of X.

In case X can take on only certain isolated values on an interval rather than all values on an interval, we call X a discrete chance quantity. Thus, in the dice problem X can take on only the isolated values 2, 3, ..., 12 and not all values on the interval between 2 and 12. We use the word discrete in contrast to the word continuous in describing a chance quantity X in this section. A continuous chance quantity X (to be discussed in Section 5.2) is one which can take on any value in an interval. In probability problems it is usually unnecessary to use the adjectives discrete or continuous in referring to a chance quantity; the context of the problem makes it clear which case is under discussion.

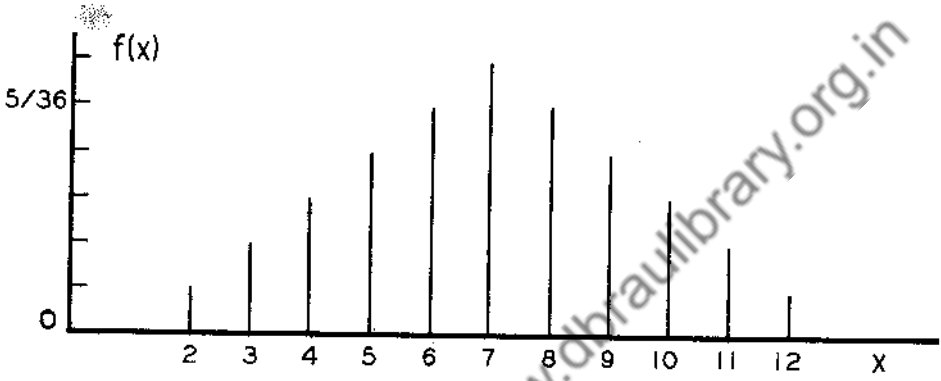
The values of x and $f(x)$ for the two-dice problem can be displayed in a probability table like this:

TABLE 5.1

Probability Table for Total Number of Dots Obtained in Throwing Two Dice

x	2	3	4	5	6	7	8	9	10	11	12
f(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

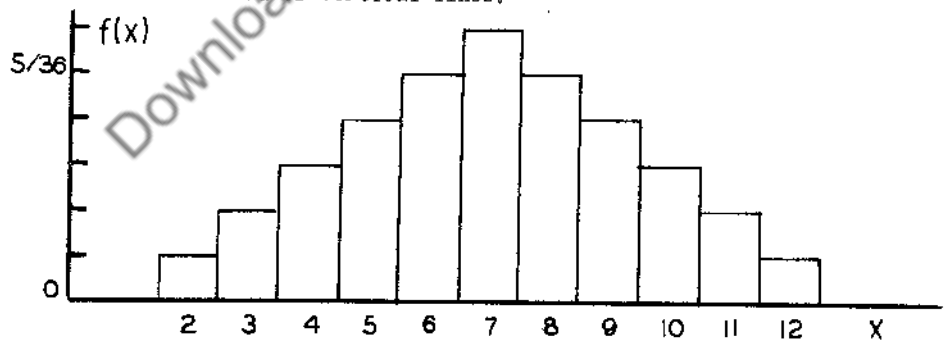
This table gives us the probability of each number of dots which can appear when throwing two dice. We can represent these probabilities graphically by a probability bar chart like Figure 5.1. The value of the probability corresponding to a given value of x is represented by the length of a vertical bar erected over that value of x. Note that the sum of the probabilities is 1.



Probability Bar Chart for Table 5.1

Figure 5.1

Alternatively, we can represent the probabilities by a probability histogram like that in Figure 5.2, where rectangles centered at the various possible values of x are erected instead of vertical lines.



Probability Histogram for Table 5.1

Figure 5.2

We can also make a cumulative probability table like this:

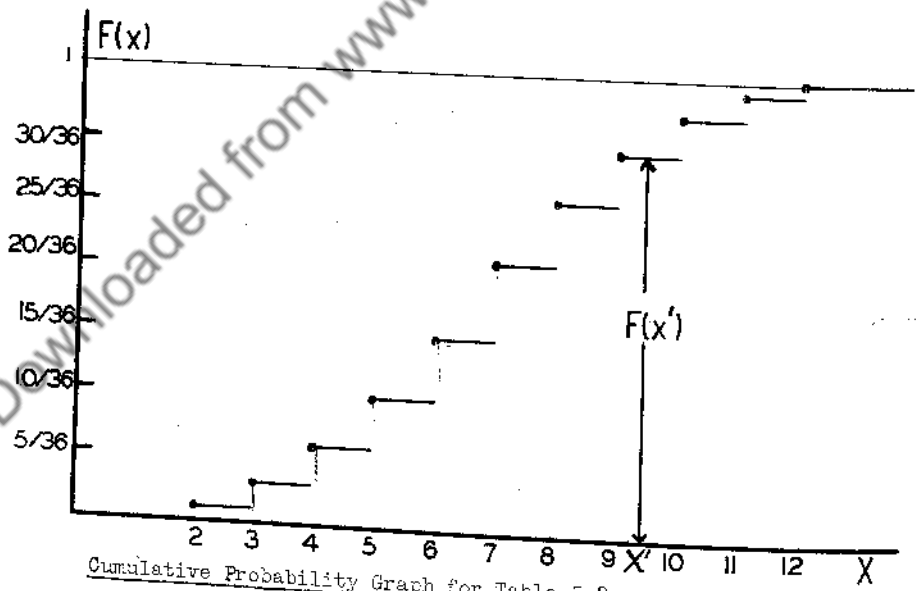
TABLE 5.2

Cumulative Probability Table for Table 5.1

x	2	3	4	5	6	7	8	9	10	11	12
F(x)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

and represent it graphically by a cumulative probability graph as shown in Figure 5.3 (where the coordinates of each large plotted dot are from the cumulative probability table). You should particularly notice that f(x) represents probabilities and F(x) represents cumulative probabilities for a population, more or less as f_i represents frequencies and F_i represents cumulative frequencies in a sample of grouped data (Section 2.4).

If in Figure 5.3 we pick any value of x we please, say x', the probability that $X \leq x'$ (i.e., that the number of dots will be less than or equal to x') is F(x'), or more briefly $Pr(X \leq x') = F(x')$.



Cumulative Probability Graph for Table 5.2

Figure 5.3

For example, if $x' = 5$, then $Pr(X \leq 5) = \frac{10}{36}$, or if $x' = 9.3$, then $Pr(X \leq 9.3) = \frac{30}{36}$. For any two values x' and x'' $(x' < x'')$, we can also find the probability that X will exceed x' and be less than or equal to x'', i.e., we can find $Pr(x' < X \leq x'')$ by taking the difference between $F(x'')$ and $F(x')$, i.e., we have $Pr(x' < X \leq x'') = F(x'') - F(x')$. For example, if $x' = 3$ and $x'' = 6$, we have $Pr(3 < X \leq 6) = F(6) - F(3) = \frac{15}{36} - \frac{3}{36} = \frac{12}{36}$, or to take another example, $Pr(4.5 < X \leq 7.5) = F(7.5) - F(4.5) = \frac{21}{36} - \frac{6}{36} = \frac{15}{36}$ respectively.

These ideas, illustrated for the case of two dice, extend to more general situations. In the general case, we would have a chance quantity X which could take on the possible values $x_1, x_2, \ldots, x_k$ with probabilities $f(x_1)$, $f(x_2)$, $\ldots$, $f(x_k)$ respectively, (where $f(x_1) + f(x_2) + \ldots + f(x_k) = 1$), and the probability table would be like this:

TABLE 5.3

Probability Table for a General Discrete Chance Quantity X

x	x_1	x_2	...	x_k
f(x)	$f(x_1)$	$f(x_2)$	...	$f(x_k)$

From this table one could construct general graphs similar to those in Figures 5.1, 5.2, 5.3 for the two dice problem. The probability distribution f(x) is called a discrete probability distribution, although we usually omit the word discrete in any specific problem if it is clear from the problem that the chance quantity X involved is discrete.

5.12 Remarks on the statistical interpretation of a discrete probability distribution.

Working with a discrete probability distribution is similar to working with a relative frequency distribution. Whatever manipulation (tables, graphs, etc.) can be done with one, can also be done with the other. In fact, we will regard a probability distribution as a relative frequency distribution for an

indefinitely large sample in which each measurement can take on one of a discrete set of values. Since we regard an indefinitely large sample as an indefinitely large population, we will therefore regard a discrete probability distribution as a relative frequency distribution for an indefinitely large population in which each measurement can take on one of a discrete set of values. This is a theoretical concept but a very useful one. Thus, for example, the probability distribution shown in Table 5.1 will be regarded as the relative frequency distribution of the total number of dots obtained on a pair of dice in an indefinitely large number of "perfect" throws of a pair of "perfect" dice - i.e., the population of "perfect" throws of a pair of "perfect" dice.

We can calculate means, variances, standard deviations and other quantities from a probability distribution very much as in the case of a relative frequency distribution.

5.13 Means, variances and standard deviations of discrete chance quantities.

The mean of a discrete chance quantity X having a probability distribution $f(x)$, is given by the following formula:

(5.1) $$\mu = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k)$$

or

(5.1a) $$\mu = \sum_{i=1}^{k} x_i f(x_i) .$$

Or more briefly we may write

(5.1b) $$\mu = E(X) .$$

We usually shorten our statement and say "μ is the mean of the probability distribution $f(x)$".

It is customary to refer to μ as the mean of the distribution $f(x)$ or more briefly, the mean of X and also the mathematical expectation of X. The expression $E(X)$, which is a shorthand expression for $\sum_{i=1}^{k} x_i f(x_i)$, is read "expectation of X". Note that $E(X)$ and $\frac{1}{n}S(X)$ are shorthand symbols playing similar roles on probability distributions and sample frequency distributions respectively.

Example: Find the mean μ of the number of dots obtained in throwing two dice.

We have, by applying formula (5.1) to Table 5.1

$$\mu = 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + 4\left(\frac{3}{36}\right) + 5\left(\frac{4}{36}\right) + 6\left(\frac{5}{36}\right) + 7\left(\frac{6}{36}\right) + 8\left(\frac{5}{36}\right)$$

$$+ 9(\tfrac{4}{36}) + 10(\tfrac{3}{36}) + 11(\tfrac{2}{36}) + 12(\tfrac{1}{36}) = 7 ,$$

or

$$\mu = E(X) = 7 .$$

The statistical interpretation of this mean is simply this: if two "perfect" dice are thrown indefinitely many times, the average number of dots per throw would be 7.

Note that if we rewrite the mean of a sample of measurements on X in a grouped frequency distribution as given by the familiar formula in the form

$$\bar{X} = \sum_{i=1}^{k} x_i(\tfrac{f_i}{n}) , \tag{5.2}$$

then the similarity of the formulas for μ and $\bar{X}$ shows up at once. The relative frequency $(\frac{f_i}{n})$ plays the same role in the definition of $\bar{X}$ that the probability $f(x_i)$ plays in the definition of μ.

The variance σ^2 of a discrete chance quantity X having probability distribution $f(x)$ (or more briefly, the variance σ^2 of the probability distribution $f(x)$) is defined as

$$\sigma^2 = (x_1 - \mu)^2 \cdot f(x_1) + (x_2 - \mu)^2 \cdot f(x_2) + \ldots + (x_k - \mu)^2 \cdot f(x_k) \tag{5.3}$$

or

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot f(x_i) . \tag{5.3a}$$

A more convenient form for σ^2 is obtained by writing (5.3a) as follows:

$$\sigma^2 = \sum_{i=1}^{k} (x_i^2 - 2x_i\,\mu + \mu^2) \cdot f(x_i) = \sum_{i=1}^{k} x_i^2 \cdot f(x_i) - 2\mu \sum_{i=1}^{k} x_i f(x_i) + \mu^2 \sum_{i=1}^{k} f(x_i) .$$

But

$$\sum_{i=1}^{k} x_i f(x_i) = \mu , \text{ and } \sum_{i=1}^{k} f(x_i) = 1 ;$$

hence we have

$$\sigma^2 = \sum_{i=1}^{k} x_i^2 f(x_i) - \mu^2 , \tag{5.3b}$$

or it may be written more briefly as

$$\sigma^2 = E(X^2) - [E(X)]^2 \quad . \tag{5.3c}$$

Compare this with the shorthand expression for s_X^2 in a sample, namely formula (3.11a

The standard deviation σ is defined as the square root of the expression for σ^2.

Example: Find the variance of the number of dots obtained in throwing two dice. We have, by applying formula (5.3b),

$$\sigma^2 = 2^2 \cdot \frac{1}{36} + 3^2 \cdot \frac{2}{36} + 4^2 \cdot \frac{3}{36} + 5^2 \cdot \frac{4}{36} + 6^2 \cdot \frac{5}{36}$$
$$+ 7^2 \cdot \frac{6}{36} + 8^2 \cdot \frac{5}{36} + 9^2 \cdot \frac{4}{36} + 10^2 \cdot \frac{3}{36} + 11^2 \cdot \frac{2}{36}$$
$$+ 12^2 \cdot \frac{1}{36} - 7^2 = \frac{35}{6} \quad .$$

The standard deviation is

$$\sigma = \sqrt{35/6} = 2.415 \quad .$$

The variance s_X^2 of a sample of grouped measurements as given by formula (3.16) may be written as follows:

$$s_X^2 = \left(\frac{n}{n-1}\right) \sum_{i=1}^{k} (x_i - \overline{X})^2 \frac{f_i}{n} \quad .$$

The similarity between this formula and formula (5.3a) for the variance of a probability distribution is clear. As n, the sample size, becomes indefinitely large, the factor $\frac{n}{n-1}$ approaches 1; the sample mean $\overline{X}$ approaches μ ; $(x_i - \overline{X})^2$ approaches $(x_i - \mu)^2$; and $\frac{f_i}{n}$ approaches $f(x_i)$. In other words, as the sample size increases indefinitely, the formula for the sample variance changes into the formula for the variance of a probability distribution; this is as one would expect, since a probability distribution is regarded as a relative frequency distribution for an indefinitely large sample.

Exercise 5.1.

1. Three coins are tossed. Let X be the chance quantity denoting the number of heads obtained. Write down the probability distribution of X in table form. Draw a probability bar chart and a cumulative probability graph for the distribution.

Also, find the mean and variance of X.

2. Five chips are marked 1, 2, 3, 4, 5 respectively. Let X be the chance quantity denoting the sum of the numbers on two chips drawn at random. Write down the probability distribution of X in table form. Draw the probability bar chart and the cumulative probability graph of the distribution. Also find the mean and variance of X.

3. Suppose there are 3 defective articles in a lot of 12. A sample of four articles is taken at random out of the lot. Let X be the chance quantity denoting the number of defective articles in the sample. Write down the expression for f(x), the probability distribution of X. Write down the probability distribution of X in table form. Construct the probability bar chart and cumulative probability graph. Find the mean and variance of X.

4. Suppose a hand of 13 cards is dealt from a deck of 52 playing cards. Let X be the chance quantity denoting the number of aces obtained. Write down the expression for f(x), the probability distribution of X. Write down the probability distribution in table form, and construct its probability bar chart and cumulative probability graph. Find the mean and variance of X.

5. One cigarette from each of four brands, A, B, C, D is partially smoked by a blindfolded person. As soon as he has taken a few puffs on a cigarette, he states the letter of the brand to which he considers it to belong. (Of course he can use each letter only once.) Let X be the chance quantity denoting the number of cigarettes correctly identified. If the identification is done at random (i.e., he is equally likely to assign any letter to any cigarette), write down the probability distribution of X in table form. Draw the probability bar chart and cumulative probability graph for the distribution. Find the mean and variance of X.

6. There is a box of 10 articles known to contain 2 defective articles. A person looks for the defectives by taking one article out at a time and testing it. What is the probability f(x) that the x-th article tested will be the last defective in the box? (In this problem the chance quantity X is the number of articles tested by the time the last defective is found; the possible values of X are 2, 3, 4, 5, 6, 7, 8, 9, 10.) Construct the probability bar chart and the cumulative probability graph for this case. Also find the mean and variance of X.

5.2 Continuous Probability Distributions.

In Section 5.1 we discussed a type of probability distribution called a discrete probability distribution, i.e., one for which there is a chance quantity X which can take on only certain isolated or discrete values. For example, if X denotes the number of heads appearing in tossing 4 coins, the only possible values X can take on are 0, 1, 2, 3, 4. If X denotes the number of aces in a hand of bridge, the only possible values X can take on are 0, 1, 2, 3, 4.

But there are situations in which a chance quantity X can take on (ideally) any value in an interval. This kind of a chance quantity is called a continuous chance quantity and probability distributions associated with them are called continuous probability distributions. As mentioned in Section 5.1 we shall, when no ambiguity arises, omit the word continuous. Our problem here is to describe continuous probability distributions.

5.21 A simple continuous probability distribution.

We can fix the ideas by a simple example. Suppose we have a 360° circular scale one "unit" long, and that we have a balanced pointer pivoted at the center as shown in Figure 5.4.

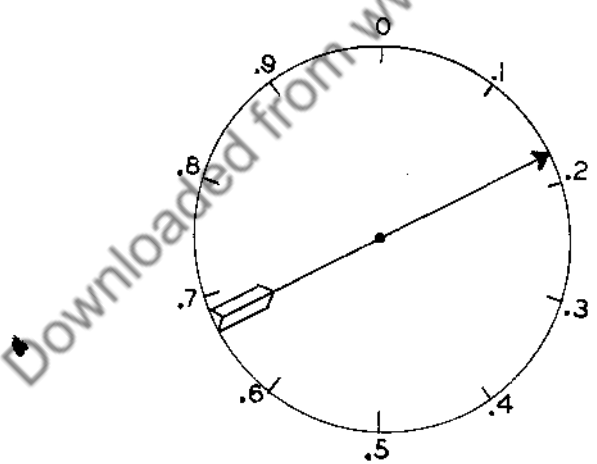


Figure 5.4

Suppose that when the pointer is whirled and allowed to stop, it is "equally likely" to stop anywhere. Then it follows from the Definition of Geometric Probability that the probability of the pointer falling into any interval is

equal to the length of that interval divided by the length of the whole scale. But we have chosen the length of the scale to be 1 unit. Thus, the probability, for example, of the pointer falling between 0.2 and 0.6 is 0.4/1 = 0.4.

If we let X be the distance along the scale measured from the zero-point to the point at which the pointer stops, then X is a continuous chance quantity. It can take on any value between 0 and 1. Its value for any single whirl of the pointer depends on where the pointer stops. It is natural therefore for us to try to set up a probability distribution which would give us any probability dependent on the chance quantity X in the pointer problem which we may wish to evaluate -- just as Table 5.2 gives us any probability we may wish to know about the chance quantity X representing the number of dots appearing in a throw of two dice. The main question is this: How can we describe the probability behavior of the pointer in mathematical terms?

The direct answer is this: Just as in the case of a discrete chance quantity (See Section 5.1 and Figure 5.3), we construct a cumulative distribution function $F(x)$ which shows the probability that the chance quantity X is less than or equal to any particular value x we may want to consider.

For the pointer example, the graph of $F(x)$ is shown in Figure 5.5. It is to be emphasized that for any particular x, say x', $F(x')$ is the probability that $X \leq x'$ (i.e., that the pointer will stop between 0 and x'), or written more briefly

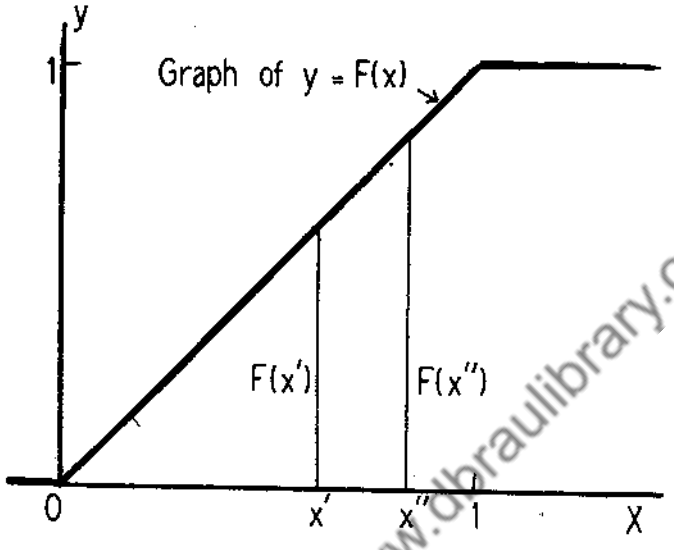
$$\Pr(X \leq x') = F(x'). \tag{5.4}$$

The graph of $F(x)$ should be compared with that for the two dice problem in Figure 5.3. Notice that the graph in that case rises by jumps, while it rises smoothly or continuously in the case of the pointer. This is what one would expect, for in the two dice problem, the probability is partitioned into eleven "pieces" and concentrated at 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 as shown in Table 5.1, whereas in the pointer problem the probability is continuously spread over the interval between 0 and 1 in a uniform manner.

As you will see from Figure 5.5, $F(x)$ is zero for any x less than 0, since the chance quantity X in the pointer problem can never be less than 0; $F(x)$ is 1 for any x greater than 1, since X is certain never to exceed 1. For any x between 0 and 1, we agreed earlier in this section (and it follows from the Definition of Geometrical Probability) that the probability of $X \leq x$ is x,

or more briefly

(5.5) $$\Pr(X \leq x) = x \ .$$



Cumulative Probability Graph When Pointer is Equally Likely to Stop Anywhere on the Scale

Figure 5.5

But $\Pr(X \leq x)$ is what we mean by $F(x)$. Thus for any value of x between 0 and 1 we have $F(x) = x$, as you can see from the graph in Figure 5.5. Summarizing the values of $F(x)$ for the pointer problem, we may say:

(5.6)
$$F(x) = 0 \text{ when } x \leq 0 \ ,$$
$$F(x) = x \text{ when } 0 < x \leq 1 \ ,$$
$$F(x) = 1 \text{ when } 1 < x \ .$$

In the case of a continuous chance quantity $\Pr(X \leq x) = \Pr(X < x)$, i.e., it does not matter whether we use $\leq$ or $<$. The probability that $X = x$ is 0, i.e., the probability that a continuous chance quantity has any single value specified in advance is zero.

To find probabilities of events more complicated than $X \leq x$ in the pointer problem, we can apply the rules of probability in Chapter 4. Suppose x' and x'' are any two particular numbers (where $x' < x''$), and that we want to

find $Pr(x' < X \leq x'')$. This probability is simply the probability of $X \leq x''$ minus the probability of $X \leq x'$, i.e.,

$$Pr(x' < X \leq x'') = Pr(X \leq x'') - Pr(X \leq x') .$$

But from (5.4) we may write this as

$$Pr(x' < X \leq x'') = F(x'') - F(x') . \tag{5.7}$$

But in our problem $F(x) = x$. Therefore,

$$Pr(x' < X \leq x'') = x'' - x' . \tag{5.8}$$

For example, $Pr(0.2 < X \leq 0.3) = 0.1$ and $Pr(0.75 < X \leq 0.87) = 0.12$. The probability that X falls in either the interval (0.2, 0.3) or the interval (0.75, 0.87) is 0.22, since the two intervals have no point in common (the events $0.2 < X \leq 0.3$ and $0.75 < X \leq 0.87$ are mutually exclusive).

If the pointer is whirled twice, the probability that it will stop between 0.3 and 0.6 both times is, by the multiplication rule for probability, the square of the probability of its stopping in this interval in one whirl, i.e., $(0.3)^2 = 0.09$. And so on, for other special probability calculations.

5.22 More general continuous probability distributions.

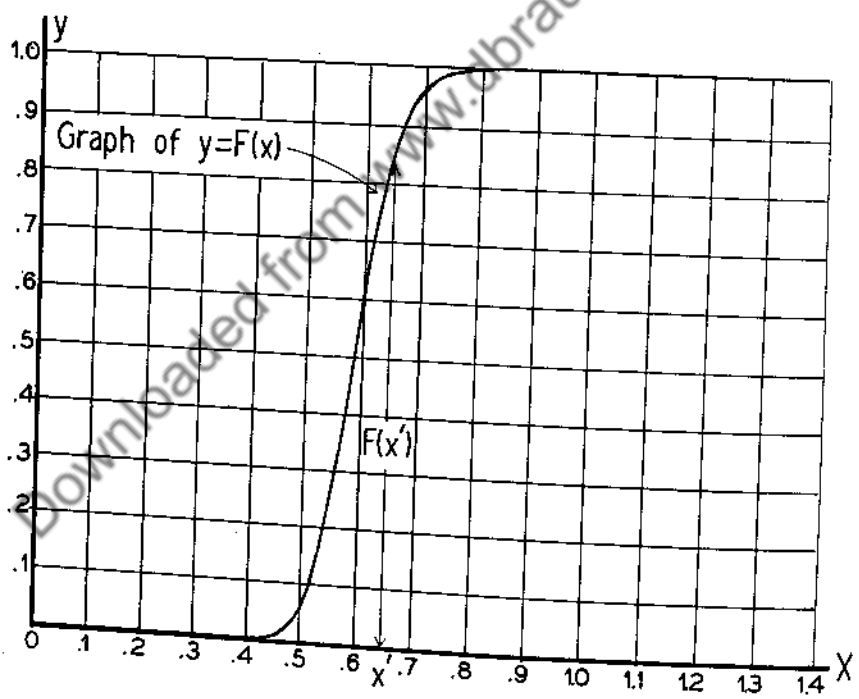
Let us generalize the pointer problem a little. Suppose the pointer is the pointer of a one pound scale weighing "half-pound" packages as they come off a production line. We may consider this flow of packages as constituting a population. They will not all weigh exactly the same. In practice, a few of them may weigh less than a half-pound, and most of them more than a half-pound (since the manufacturer must guarantee a half-pound of material in each package). The chance quantity X here is still the position on the scale (measured from the zero point) at which the pointer will stop when weighing a package. For any specified weight x' there will be a fraction F(x') of the packages weighing less than x' pounds. The graph of F(x) may look something like that shown in Figure 5.6.

In Figure 5.6 it will be seen that F(x) is 0 when $x < 0$, and F(x) = 1 when $x \geq 1$, just as in the simple case of uniformly distributed probability (Figure 5.5). But between 0 and 1, F(x) is some S-shaped curve (or ogive) instead of a straight line inclined at 45° to the X-axis. In such a problem as

that of weighing packages, there would not exist, in general, a mathematical formula for $F(x)$; the function $F(x)$ would simply be given graphically by a smooth curve. In order to be able to plot $F(x)$ we would actually need a large sample of measurements from which we could construct a frequency polygon -- and then "smooth" the polygon into a curve by drafting curves or by freehand drawing. From a carefully drawn graph we could read off the value of $F(x)$ at any value of x to a practical degree of accuracy. The curve would then provide us with a way of obtaining "estimated" probabilities. For example, reading from the graph we make these "estimates"

$$\Pr(X \leq 0.5) = F(0.5) = 0.07$$
$$\Pr(0.5 < X \leq 0.9) = F(0.9) - F(0.5)$$
$$= 1.00 - 0.07 = 0.93 .$$

And so on. The median of X is that value x for which $F(x) = 0.5$, or median = 0.58.



Cumulative Probability Graph for Pointer When Weighing "half-pound" Packages

Figure 5.6

Exercise 5.2.

1. Consider the pointer problem as having the cumulative probability distribution F(x) graphed in Figure 5.5. What is the probability that the first digit of X (after the decimal point) is even? That the second digit is ≤ 5? That the pointer stops within a distance of 0.2 of the middle of the interval?

2. Suppose a pointer is equally likely to stop at any point on a scale between 0 and 10. Graph the cumulative probability function F(x) for this problem.

3. In problem No. 1, what is the probability that $0.3 < X \leq 0.6$ or $0.4 < X \leq 0.8$? If the pointer is spun five times, what is the probability that it will stop within the middle tenth of the interval at least once?

4. If the pointer in problem No. 1 is spun three times, what is the probability the pointer will stop between 0 and x every time? If you call this probability G(x), what formula does G(x) have for $x < 0$, for $x \geq 1$ and for $0 < x \leq 1$? Graph G(x). What is the interpretation of G(x)? What is the chance quantity X in this problem?

5. Reading from the graph in Figure 5.6, the probability is about 0.9 that $X \leq$ what value? That $X \geq$ what value? Estimate the probabilities that X will fall in each of the following intervals: (0,0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), (0.4, 0.5), (0.5, 0.6), (0.6, 0.7), (0.7, 0.8), (0.8, 0.9) and (0.9, 1.0), and arrange them in table form.

5.3 Mathematical Manipulation of Continuous Probability Distributions.

The simplest way to express the distribution of a continuous chance quantity is by means of a cumulative probability distribution. Tables of special distributions like the normal distribution to be discussed in Chapter 8 are almost always tables of values of cumulative probability functions computed for closely spaced values of x.

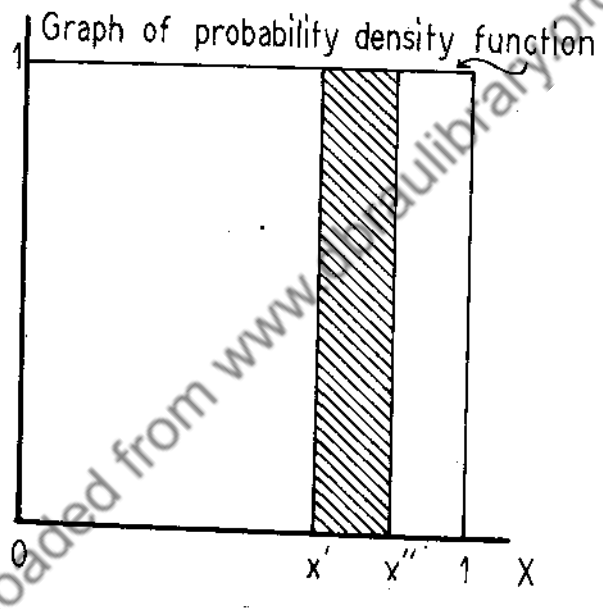
For easy mathematical handling, we often want to represent the probability distribution of a continuous chance quantity in another way. In simple cases this will allow us to find means and variances by integration (without numerical computation from a table or otherwise).

5.31 Probability density functions - a simple case.

In Section 5.2 we showed how the probability of X falling less than or

equal to a value x' is represented by the ordinate at x' (Figures 5.5 and 5.6); or the probability of X falling between two values, say x' and x'', by the difference between the ordinates at x' and x''. Now can we find a curve in each case so that the probability of X falling between x' and x'' is represented by the area under the curve between x' and x'' (and above the X-axis)?

For the case of the uniformly distributed probability (Figure 5.5) the answer is easy. The curve in this case is a straight line parallel to the X-axis and one unit above it, as shown in Figure 5.7.



Graph of Probability Density Function for Cumulative Probability Function graphed in Figure 5.5

Figure 5.7

If we take two values of x, say x' and x'', the area under the graph of y = 1 between x = x' and x = x'' is (x'' - x')·1. But you will remember that this is the probability that X will fall between x' and x'', and is equal to F(x'') - F(x') in Figure 5.5. This is true for any two points x' and x''. Therefore, we have found a "curve" (a straight line here) such that the area between it and the X-axis which lies between any two values of x (x' and x'') has the same

numerical value as the difference between the ordinates of a cumulative probability graph at x' and x''.

We shall call the function having the horizontal line in Figure 5.7 as its graph the probability density function for this problem, and its graph the probability density graph. We can think of the graph in Figure 5.7 as the histogram of relative frequencies in an indefinitely large sample of spins of the pointer. As the sample becomes larger and larger, the histogram of relative frequencies (based on any number of cells into which the range is divided) "becomes" the probability density graph.

5.32 Probability density functions - a more general case.

What will the probability density graph for the cumulative probability graph of Figure 5.6 look like? Since we are looking for a curve such that the area under it between x = x' and x = x'' is equal to F(x'') - F(x') in Figure 5.6, for any two values x' and x'', we can use the following table of figures for constructing the desired curve (read from the graph in Figure 5.6).

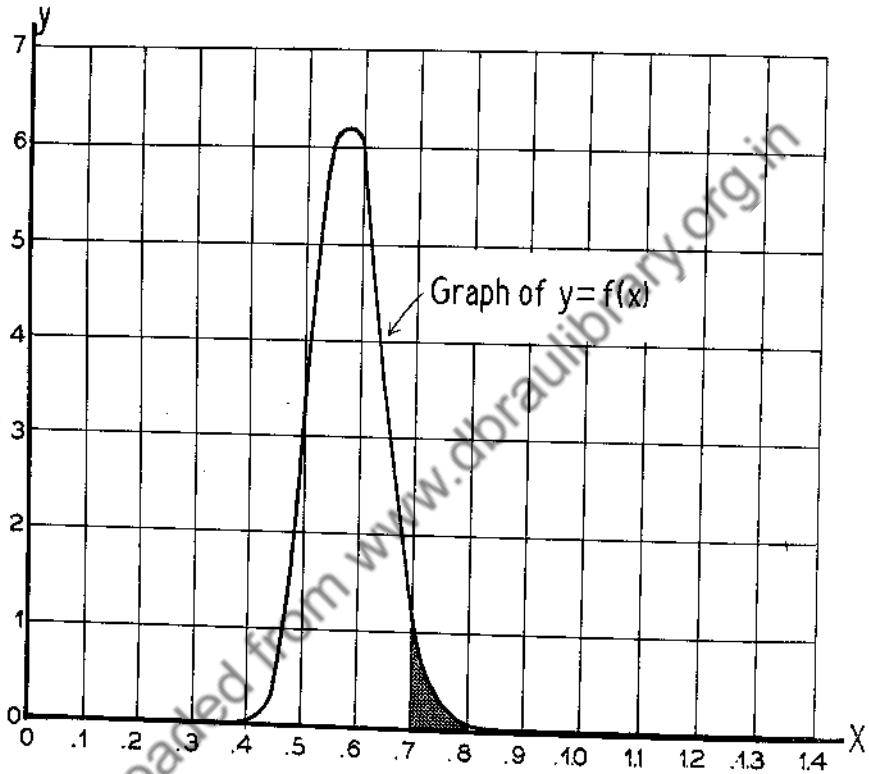
TABLE 5.4

Ordinate Differences from Figure 5.6

x'	x''	Ordinate Differences: F(x'') - F(x') (Area under desired curve between x' and x'')
0	0.1	0 - 0 = 0
0.1	0.2	0 - 0 = 0
0.2	0.3	0 - 0 = 0
0.3	0.4	0 - 0 = 0
0.4	0.5	.07 - 0 = .07
0.5	0.6	.62 - .07 = .55
0.6	0.7	.96 - .62 = .34
0.7	0.8	1.00 - .96 = .04
0.9	1.0	1.00 - 1.00 = 0
1.0	1.1	1.00 - 1.00 = 0
1.1	1.2	1.00 - 1.00 = 0
1.2	1.3	1.00 - 1.00 = 0

Considering the 10 sets of values of x' and x'' in Table 5.4 as boundaries of 10 cells and the ten "areas" as relative frequencies, we can construct a histogram so that the total area under the histogram is 1. If we imagine making

histograms with finer and finer cells these histograms would become more and more like a smooth curve as shown in Figure 5.8. This smooth curve is the graph of the probability density function f(x).



Graph of the Probability Density Function f(x) for the Cumulative Probability Function graphed in Figure 5.6

Figure 5.8

The area under the curve in Figure 5.8 between any two values of x would then be equal to the difference between the values of F(x), (the function graphed in Figure 5.6), corresponding to the two values of x. For example, the shaded area shown in Figure 5.8 is equal to 0.04, which is the value of F(0.8) - F(0.7), the difference between the ordinates at x = 0.7

and 0.8 in Figure 5.6.

We now have two ways of expressing the values of $Pr(x' < X \leq x'')$, the probability of X falling between any two values x' and x''. First, we can express this probability as the difference between the ordinates of the cumulative probability curve F(x) of Figure 5.6 at x' and x'', i.e.,

$$Pr(x' < X \leq x'') = F(x'') - F(x'); \tag{5.9}$$

and secondly, we can express the probability as the area under the graph of the probability density function f(x) of Figure 5.8 between x = x' and x = x'', i.e.,

$$Pr(x' < X \leq x'') = \int_{x'}^{x''} f(x)dx \ . \tag{5.10}$$

The expression on the right of (5.10) is the definite integral of f(x)dx between x = x' and x = x'', and yields the desired area. Hence, from (5.9) and (5.10),

$$F(x'') - F(x') = \int_{x'}^{x''} f(x)dx \ ; \tag{5.11}$$

putting x' = 0, and dropping the '' on x'', we have

$$F(x) = \int_{0}^{x} f(x)dx \ , \tag{5.12}$$

which gives us a way of finding the cumulative probability function F(x) from the probability density function f(x) by integration. Conversely, we can find f(x) from F(x) by differentiation, i.e.,

$$f(x) = \frac{dF(x)}{dx} \ . \tag{5.13}$$

You should get the difference between f(x) and F(x) firmly fixed in mind and then keep the distinction clear.

Returning to the pointer problem in which the probability is uniformly distributed, we have

$$f(x) = 1$$

and

$$F(x) = \int_0^x 1 \cdot dx = x \ , \quad 0 < x \leq 1 \ ,$$

as we saw earlier and as was graphed in Figure 5.5.

It should be noted that we are using the notation $f(x)$ for probability distribution in both the discrete and continuous cases. There will be no confusion, however, because it will always be clear from the context of any problem or situation whether we are talking about the discrete case or the continuous case (i.e., whether the chance quantity X is discrete or continuous).

5.33 Continuous probability distributions - general case.

We have discussed the setting up of a continuous probability distribution for the stopping point of a pointer on a continuous scale reading from 0 to 1. The ideas introduced extend to more general situations involving continuously distributed probabilities. More generally, we would have an interval (α, β) (where α could be any finite number or $-\infty$, and β could be any finite number or $+\infty$) within which it is certain that a chance quantity X will fall. We would have a probability density function $f(x)$ defined over the interval (α, β) which could be represented as some kind of a smooth curve above the X-axis. In the general case, the total area under the curve would represent the probability of X falling between α and β and hence would be 1. The probability of X falling between any two values x' and x'' would be represented by a shaded area similar to that shown in Figure 5.8. We would have a smooth cumulative probability distribution $F(x)$; its graph would be similar to that shown in Figure 5.6, except that the curve would extend over the interval (α, β) instead of $(0,1)$. $F(x)$ would be determined from $f(x)$ by formula (5.12) with 0 replaced by α and $f(x)$ would be determined from $F(x)$ by formula (5.13).

5.34 The mean and variance of a continuous probability distribution.

As in the discrete case discussed in Section 5.1, we may calculate the mean μ and variance σ^2 of a continuous probability density function $f(x)$ defined over the interval (α, β). For the mean μ of the distribution $f(x)$ we have

$$\mu = \int_\alpha^\beta x f(x) dx \ , \tag{5.14}$$

which is analogous to the formula (5.1a) for the discrete case. We may also briefly write (just as was done in (5.1b))

$$\mu = E(X) \ , \tag{5.14a}$$

where it is understood that E(X) is a shorthand expression for the integral in (5.14).

For the variance σ^2 of the distribution f(x) we have

$$\sigma^2 = \int_{\alpha}^{\beta} (x-\mu)^2 \, f(x)dx \ ; \tag{5.15}$$

a more convenient form is

$$\sigma^2 = \int_{\alpha}^{\beta} x^2 f(x)dx - \mu^2, \tag{5.15a}$$

or more briefly

$$\sigma^2 = E(X^2) - [E(X)]^2. \tag{5.15b}$$

Formulas (5.15), (5.15a) and (5.15b) are analogous to formulas (5.3a), (5.3b) and (5.3c) in the discrete case.

Returning to the example of the pointer, consider the pointer equally likely to stop at any point on the scale. Then f(x) = 1 as shown in Figure 5.7. To get the mean, we use (5.14) with $\alpha = 0$, and $\beta = 1$,

$$\mu = E(X) = \int_0^1 x \cdot 1 \cdot dx = \frac{x^2}{2}\Big]_0^1 = \frac{1}{2} - 0 = \frac{1}{2} \ ,$$

or

$$\mu = \frac{1}{2} \ .$$

To get the variance we use (5.15a) with $\alpha = 0$, and $\beta = 1$,

$$\sigma^2 = \int_0^1 x^2 \cdot 1 \cdot dx - (\tfrac{1}{2})^2$$

$$= \frac{x^3}{3}\Big]_0^1 - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12},$$

or

$$\sigma^2 = \frac{1}{12}.$$

5.35 Remarks on the statistical interpretation of continuous probability distributions.

As in the case of discrete probability distributions, we will use a continuous probability distribution as a population distribution "model". More specifically, we will regard a continuous probability distribution as a relative frequency distribution for an indefinitely large sample (i.e., for a population) when the measurements can take on any value within a given interval. The reasonableness of this "model" is clear if you consider taking larger and larger samples from a population (in which measurements can "ideally" take on any value in a certain interval), and make histograms of relative frequencies with smaller and smaller cell lengths. Under such conditions these histograms become increasingly more like a smooth curve such as that shown in Figure 5.8, and their cumulative polygons become increasingly more like a smooth curve such as that shown in Figure 5.7. (In general, we would have other numbers than 0 and 1 for the end points of the interval.) Therefore, if we arbitrarily introduce a smooth curve to represent the distribution of relative frequencies for such a population, we have, at least in some cases, a fairly simple and fairly accurate model to use in calculating frequencies, means and other quantities dependent on the population distribution.

Exercise 5.3.

1. Suppose X is a chance quantity with a continuous probability distribution $f(x) = \frac{1}{10}$ on the interval (0,10). Find the expression for the cumulative probability distribution F(x) and graph it. Find the median (from F(x), not from its graph). Find the mean μ and variance σ^2 of the distribution.

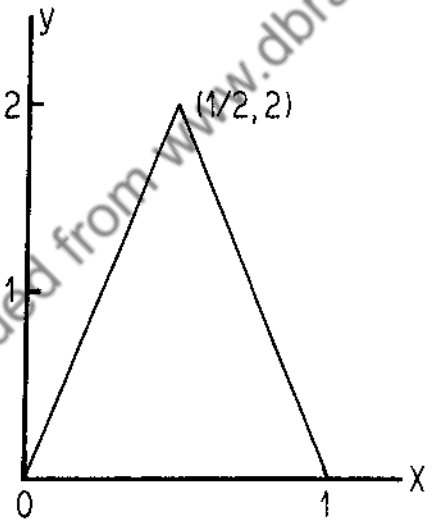
2. Suppose X is a chance quantity with a continuous probability distribution $f(x) = 2x$ on the interval (0, 1). Find F(x) and graph it. Find μ and σ^2. Find the median, the lower and upper quartiles of the distribution. Find the

value of $Pr(\frac{1}{2} < X \leq \frac{3}{4})$; of $Pr(X \leq \frac{1}{4})$.

3. A chance quantity X has the continuous probability distribution $f(x)=6x(1-x)$ on the interval (0, 1). Find $F(x)$ and graph it. Find μ and σ^2 .

4. Since $G(x)$ in problem No. 4 of Exercise 5.2 is a cumulative probability distribution, find the probability density function from it. Find the mean and the variance of X.

5. A point is taken "at random" from the interval (0, 1), all points being equally likely. A second point is then taken in the same way. Let X be the coordinate of the point half way between these points. X is a continuous chance quantity with a probability density function having an inverted V graph as shown in the following figure:



Find the values of $Pr(X \leq .25)$, $Pr(.1 < X \leq .9)$, $Pr(X > .8)$. Write down the formula for $f(x)$. Find the mean and variance of X. Find the formula for $F(x)$ and graph $F(x)$.

6. Suppose a continuous cumulative distribution $F(x)$ is defined as follows:

$$F(x) = 0 \qquad x \leq 1$$

$$F(x) = \frac{(x-1)^4}{16} \qquad 1 < x \leq 3$$

$$F(x) = 1 \qquad x > 3 .$$

Graph $F(x)$. Find the probability density function $f(x)$ and graph it. Find the mean and variance of X. Find $Pr(2 < X \leq 3)$. Find the median.

CHAPTER 6. THE BINOMIAL DISTRIBUTION.

6.1 Derivation of the Binomial Distribution.

Let us return to the coin-tossing problem discussed in Section 4.4. There it was shown that if n "true" coins (i.e., coins for which heads and tails are equally likely) are tossed once (or one "true" coin is tossed n times) then the probability f(x) of getting x heads is given by formula (4.6) i.e.,

$$f(x) = \frac{n!}{x!(n-x)!} \left(\frac{1}{2}\right)^n .$$

Now suppose we have a biased coin for which the probability of a head is p and the probability of a tail is q (= 1 - p). What is the probability of getting x heads in throwing such a coin n times?

First consider throwing the coin twice. The four possible events are

TT, TH, HT, HH .

Considering the results of the two throws as being independent events, then it follows by Rule III (multiplication of probabilities) that the probabilities for these four events are

qq, qp, pq, pp .

The two middle events each result in one head (and one tail), and the probability of getting one head is, by Rule II (addition of probabilities), the sum of the two probabilities, i.e., 2pq. Therefore, the probabilities of getting 0, 1, 2 heads in throwing the biased coin twice are

$$q^2, \ 2pq, \ p^2$$

respectively.

In the case of tossing the coin three times, the possible events are

TTT, TTH, THT, HTT, THH, HTH, HHT, HHH,

and their probabilities are

qqq, qqp, qpq, pqq, qpp, pqp, ppq, ppp .

The second, third and fourth events result in one head, the fifth, sixth and seventh result in two heads. Thus the probability of getting one head is $qqp + qpq + pqq = 3pq^2$; similarly the probability of two heads is $3p^2q$. Hence the probabilities of getting 0, 1, 2, 3 heads in throwing the biased coin three times are

$$q^3,\ 3pq^2,\ 3p^2q,\ p^3$$

respectively.

In general, suppose we ask for the probability of x heads (and n - x tails) in n tosses of the coin. Any particular order in which x heads and n - x tails appear is simply an arrangement (permutation) of x H's and n - x T's, and the probability of this particular arrangement is $p^x q^{n-x}$. We know from Section 4.3 (formula (4.2)) that there are $\frac{n!}{x!\,(n-x)!}$ (i.e., C_x^n) possible permutations of x H's and n-x T's. Hence by Rule II (addition) the probability f(x) of getting x heads and n-x tails is $p^x q^{n-x} + p^x q^{n-x} + \ldots + p^x q^{n-x}$ (the number of terms being $\frac{n!}{x!\,(n-x)!}$), that is,

$$f(x) = C_x^n p^x q^{n-x} . \tag{6.1}$$

For example, suppose we imitate a biased coin by marking two of the faces on a die H and four of the faces T. In this case, the probability of a "head" is $\frac{1}{3}$ and that of a "tail" is $\frac{2}{3}$. If this "coin" is "tossed" 4 times, the probability of getting 3 "heads" is

$$C_3^4 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right) = \frac{8}{81} .$$

Note that the expression for f(x) in (6.1) is <u>simply</u> the general term in the expansion of the binomial $(q + p)^n$. In fact, we have

$$(q+p)^n = q^n + C_1^n pq^{n-1} + C_2^n p^2 q^{n-2} + \ldots + C_x^n p^x q^{n-x} + \ldots + p^n . \tag{6.2}$$

The first term on the right is the probability of 0 "heads" in n "tosses", the second is the probability of 1 "head" in n "tosses", and so on. We may list these terms in the form of a probability distribution table as follows:

TABLE 6.1

Probability Table for General Binomial Distribution

x	f(x)
0	q^n
1	$C_1^n\, pq^{n-1}$
2	$C_2^n\, p^2 q^{n-2}$
.	.
.	.
.	.
x	$C_x^n\, p^x q^{n-x}$
.	.
.	.
.	.
n	p^n

So far, we have talked about "heads" and "tails" in throwing "coins". We lose nothing by being a little more general and talking about event "E" and event "not E". We can then summarize our discussion in the following important result:

If an event E has probability p of occurring on each of n independent trials, then the probability f(x) that it will occur exactly x times in n trials is given by

$$f(x) = C_x^n\, p^x q^{n-x} \tag{6.3}$$

where $q = 1 - p$.

The probability distribution (6.3) is naturally called the binomial probability distribution, or simply the binomial distribution.

By putting $p = \frac{1}{2}$ into formula (6.3) we get the formula for the probability of obtaining x heads in tossing n unbiased coins which we have already discussed in Section 4.4 (see formula (4.6)).

For small values of n, the individual probabilities in any binomial distribution problem can be conveniently calculated by means of a recursion formula, i.e., a formula relating the values of f(x) for two successive values of x. For we have

$$f(x) = C_x^n \, p^x q^{n-x}$$

$$f(x+1) = C_{x+1}^n \, p^{x+1} \, q^{n-x-1} .$$

Taking ratios we have

$$\frac{f(x+1)}{f(x)} = \frac{C_{x+1}^n \, p^{x+1} \, q^{n-x-1}}{C_x^n \, p^x q^{n-x}} = \frac{C_{x+1}^n}{C_x^n} \cdot \frac{p}{q} = \frac{n-x}{x+1} \cdot \frac{p}{q}$$

or

$$f(x+1) = \frac{n-x}{x+1} \cdot \frac{p}{q} f(x) . \tag{6.4}$$

To use this recursion formula on any given problem we first calculate the value of f(0), which is

$$f(0) = q^n ;$$

then substituting x = 0 in (6.4) we get the value of f(1), i.e.,

$$f(1) = \frac{n}{1} \cdot \frac{p}{q} \cdot f(0) .$$

Similarly

$$f(2) = \frac{n-1}{2} \cdot \frac{p}{q} \cdot f(1) ,$$

and so on for all terms.

Example: If a die is rolled four times, what is the probability distribution of x, where x is the number of times the "six" occurs? The possible values of x are 0, 1, 2, 3, 4.

The probability of x "sixes" is

$$f(x) = C_x^4 \, (\tfrac{1}{6})^x \, (\tfrac{5}{6})^{4-x} .$$

The recursion formula is

$$f(x+1) = \frac{4-x}{x+1} \cdot \frac{1}{5} \cdot f(x) ,$$

and $f(0) = (\frac{5}{6})^4 = 0.482$. For the other values we have

$$f(1) = 4 \cdot \frac{1}{5} \cdot f(0) = 0.386 \qquad f(3) = \frac{2}{3} \cdot \frac{1}{5} \cdot f(2) = 0.015$$

$$f(2) = \frac{3}{2} \cdot \frac{1}{5} \cdot f(1) = 0.116 \qquad f(4) = \frac{1}{4} \cdot \frac{1}{5} \cdot f(3) = 0.001.$$

Arranged in table form we have:

TABLE 6.2

Probabilities of Getting 0, 1, 2, 3, 4 Sixes in Rolling a Die Four Times.

x (No. "sixes")	f(x)	f(x) (to 3 decimal places)
0	$1(\frac{5}{6})^4$	0.482
1	$4(\frac{1}{6})(\frac{5}{6})^3$	0.386
2	$6(\frac{1}{6})^2(\frac{5}{6})^2$	0.116
3	$4(\frac{1}{6})^3(\frac{5}{6})^1$	0.015
4	$1(\frac{1}{6})^4$	0.001

In the case of large values of n, we find that a law known as the normal or Gaussian probability distribution gives a good approximation to the binomial distribution probabilities. The normal distribution and its applications will be discussed in Chapter 8.

It should be made clear that it is not necessary to calculate the probabilities in Table 6.2 by recursion. They could be calculated directly from the formula for f(x). But you will find that if n is very much larger than x, recursion computation simplifies the calculations.

6.2 The Mean and Standard Deviation of the Binomial Distribution.

In any given example where the values of n and p are given, we could

calculate arithmetically the mean and standard deviation of the binomial distribution. For example, suppose we want to know the mean μ and standard deviation σ of X in the die problem mentioned above. We have for the mean μ, applying formula (5.1) to the table of values given in Table 6.2:

$$\mu = 0 \cdot \left(\frac{5}{6}\right)^4 + 1 \cdot \left[4\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^3\right] + 2 \cdot \left[6\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2\right]$$

$$+ 3 \cdot \left[4\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)\right] + 4 \cdot \left(\frac{1}{6}\right)^4$$

$$= 4 \cdot \left(\frac{1}{6}\right) \cdot \left[\left(\frac{5}{6}\right)^3 + 3\left(\frac{1}{6}\right)^1 \cdot \left(\frac{5}{6}\right)^2 + 3\left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^1 + \left(\frac{1}{6}\right)^3\right]$$

$$= 4\left(\frac{1}{6}\right) \cdot \left(\frac{5}{6} + \frac{1}{6}\right)^3 .$$

But, since $\left(\frac{5}{6} + \frac{1}{6}\right)^3 = 1$, we finally have

$$\mu = 4\left(\frac{1}{6}\right) = \frac{2}{3} .$$

Similarly, if we apply formula (5.3b) to this example, we find the variance σ^2 to be

$$\sigma^2 = 4\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = \frac{5}{9} .$$

In the case of the general binomial distribution (6.3):

The mean μ is given by the formula

(6.5) $$\mu = np .$$

The variance is given by the formula

(6.6) $$\sigma^2 = npq ,$$

The standard deviation is given by the formula

(6.7) $$\sigma = \sqrt{npq} .$$

These formulas are important and we shall make repeated use of them. The simplest derivations of them are based on theoretical sampling principles for sampling from an indefinitely large population; these will be discussed in Chapter 9. The direct derivation of the formulas is a little cumbersome, but straightforward, and will be given here for the benefit of those who want to see how they are directly derived.

The mean μ is given by applying formula (5.1) to the binomial distribution (6.3). This gives

$$\mu = \sum_{x=0}^{n} x \left[C_x^n p^x q^{n-x}\right] . \tag{6.8}$$

Similarly, applying formula (5.3b) to the binomial distribution (6.3), we have

$$\sigma^2 = \sum_{x=0}^{n} x^2 \left[C_x^n p^x q^{n-x}\right] - \mu^2 . \tag{6.9}$$

Now we must find a simple expression for each of the two sums involved. In order to do this, we introduce a device that may appear a little puzzling at first, but provides an easy way of evaluating these sums. Let us write down the following binomial expression and its expansion

$$(q + tp)^n = C_0^n (tp)^0 q^n + C_1^n (tp)^1 q^{n-1} + \ldots + C_x^n (tp)^x q^{n-x} + \ldots + C_n^n (tp)^n q^0 , \tag{6.10}$$

where t is a letter just arbitrarily inserted. If we differentiate both sides with respect to t, we have

$$np(q + tp)^{n-1} = 1 \cdot \left[C_1^n p^1 q^{n-1}\right] t^0 + 2 \cdot \left[C_2^n p^2 q^{n-2}\right] t^1 + \ldots + x \cdot \left[C_x^n p^x q^{n-x}\right] t^{x-1} + \ldots + n \cdot \left[C_n^n p^n q^0\right] t^{n-1} . \tag{6.11}$$

But if we put t = 1, and remember that q + p = 1, we find that the left side of (6.11) reduces to np, and the right-hand side is simply the sum indicated on the right-hand side of (6.8) all written out. Hence, the mean of the binomial

distribution (6.3) is

$$\mu = np \,, \tag{6.12}$$

as stated in formula (6.5). Now to find a simple expression for σ^2, let us return to (6.11) and multiply both sides of the equation by t. We get

$$np\,t(q + tp)^{n-1} = 1 \cdot [C_1^n p^1 q^{n-1}]t + 2 \cdot [C_2^n p^2 q^{n-2}]t^2 + \ldots + x \cdot [C_x^n p^x q^{n-x}]t^x + \ldots + n \cdot [C_n^n p^n q^0]t^n \,. \tag{6.13}$$

Now differentiate both sides with respect to t. We get

$$np(q + tp)^{n-1} + n(n-1)\,p^2\,t\,(q + tp)^{n-2} = 1^2 \cdot [C_1^n p^1 q^{n-1}]t^0 + 2^2 \cdot [C_2^n p^2 q^{n-2}]t^1 + \ldots + x^2 \cdot [C_x^n p^x q^{n-x}]t^{x-1} + \ldots + n^2 \cdot [C_n^n p^n q^0]t^{n-1} \,. \tag{6.14}$$

Putting t = 1, and remembering that q + p = 1, we see that the right-hand side of (6.14) becomes the written out form of the sum indicated on the right-hand side of (6.9). The left-hand side of (6.14) reduces to $np + n(n-1)\,p^2$ when t = 1. Hence, substituting this quantity in (6.9) and also the value of μ (i.e., np), we find the variance σ^2 to be

$$\sigma^2 = np + n(n-1)p^2 - n^2 \cdot p^2 = np - np^2 = npq \,.$$

Therefore, the variance of X in the binomial distribution (6.3) is

$$\sigma^2 = npq \,, \tag{6.15}$$

as stated in formula (6.6), and the standard deviation is

$$\sigma = \sqrt{npq} \,. \tag{6.16}$$

6.3 "Fitting" a Binomial Distribution to a Sample Frequency Distribution.

In the "true" coin and "true" dice problems mentioned earlier, the numerical value of p was agreed to in advance, and theoretical binomial

distributions of probabilities of numbers of heads (or numbers of "sixes") in n trials were obtained. But there are problems in which one cannot arrive at such an agreement or determine in advance what value p has. In such cases p will usually be estimated experimentally in accordance with Definition II in Section 4.1. To take a simple example, what is the probability p that an ordinary celluloid-headed thumb tack, when thrown on the floor, will fall "point up"? One cannot tell by examining the tack. But he can estimate it by throwing it a large number of times or throwing several tacks like it a large number of times. For instance, in 100 throws of 5 tacks the observed frequency distribution of values of x (number of tacks falling point up) in Table 6.3 was actually obtained.

TABLE 6.3

Frequency Distribution of Number of Tacks Falling Point up in Throwing 5 Tacks 100 Times

x	0	1	2	3	4	5
f	2	14	20	34	22	8

Now the question is: How do we estimate p, the probability of a tack falling "point up"? If one were to set up a probability distribution for x (the number of tacks in 5 falling "point up"), using the unknown probability p, he would obtain

$$f(x) = C_x^5 \, p^x \, q^{5-x}$$

for x = 0, 1, 2, 3, 4, 5.

To estimate p, we set the mean of the (theoretical) probability distribution equal to the mean of the (observed) frequency distribution. Since n = 5, the mean of the theoretical distribution is (by formula (6.5))

$$\mu = 5p \ ,$$

and the mean of the observed distribution is

$$\overline{X} = \sum_{i=0}^{5} \frac{f_i x_i}{100} = \frac{284}{100} = 2.84 \ .$$

Hence, equating the values of μ and $\bar{X}$, from

$$5p = 2.84$$

we get the following estimate of p based on the experimental results in Table 6.3:

$$p = .568 .$$

It should be noted that if we consider our data as coming from the results of throwing one tack 500 times, we would have 284 trials in which the tack fell point up, thus giving us $p = \frac{284}{500} = .568$ directly by probability Definition II. It is to be emphasized that .568 is only an estimate of p based on one experiment. If another experiment were made, a value would be obtained which would probably be slightly different but near .56 or .57.

Using this value of p in $C_x^5 p^x q^{5-x}$ for x = 0, 1, 2, 3, 4, 5, we would have the "fitted" binomial distribution

$$C_x^5 (.568)^x (.432)^{5-x},$$

which can be expected to "approximate" the observed relative frequency distribution of the number of tacks falling point up. The results are shown in Table 6.4. The entries in the column headed "expected" frequency are obtained by simply multiplying the "fitted" probabilities by 100 (the value of n).

It will be noticed that there is a fairly close agreement between relative frequencies and "fitted" binomial probabilities and similarly between observed frequencies and "expected" frequencies. The cumulative observed frequencies and cumulative "expected" frequencies are also in good agreement, as you will see.

TABLE 6.4

Fitting of a Binomial Distribution to the Data in Table 6.3

x_i	Observed Frequency	Relative Frequency	"Fitted" Binomial Probability Distribution	"Expected" Frequency	Cumulative Observed Frequency	Cumulative "Expected" Frequency
0	2	.02	$C_0^5(.568)^0(.432)^5=.015$	1.5	2	1.5
1	14	.14	$C_1^5(.568)^1(.432)^4=.099$	9.9	16	11.4
2	20	.20	$C_2^5(.568)^2(.432)^3=.260$	26.0	36	37.4
3	34	.34	$C_3^5(.568)^3(.432)^2=.342$	34.2	70	71.6
4	22	.22	$C_4^5(.568)^4(.432)^1=.225$	22.5	92	94.1
5	8	.08	$C_5^5(.568)^5(.432)^0=.059$	5.9	100	100
	100	1.00	1.000	100		

Exercise 6.

1. If ten "true" dice are tossed once, what is the probability that x aces will turn up? What is the probability that at least two aces will turn up? That not more than one ace will turn up?

2. Assume that on the average one telephone number out of five called between four and five p. m. on weekdays in a certain city is busy. What is the probability that if ten randomly selected telephone numbers are called not more than two of them will be busy?

3. Consider all families of five children, in which there are no twins. Assuming the probabilities of a child being a boy or a girl to be equal, what fraction of the families would you estimate to have at least one son and at least one daughter? What fraction would have five sons or five daughters?

4. If a hand of 13 cards is dealt from a deck of 52 bridge cards, the probabilities are approximately .30, .44, and .26 of getting no aces, getting one ace and getting more than one ace respectively. What is the probability that

a person will play four hands of bridge and never receive an ace? That he will play four hands and never fail to get at least one ace? That he will play four hands and get at least two aces every time?

5. In problem No. 4, let X be the number of hands in which a player gets no aces in playing four hands. Write down the expression for $f(x)$, the probability that $X = x$. Write down the probability distribution of X in table form. Find the mean and variance of X.

6. In inspecting 1450 welded joints produced by a certain type of welding machine, 148 defective joints were found. In welding five joints, what is the probability of getting 0, 1, 2, 3, 4, 5 defective joints?

7. Two coins are tossed together five times. Let X be a chance quantity denoting the number of pairs of heads obtained. Write down the expression for $f(x)$, the probability that $X = x$. Write down the probability distribution of X in table form. Find the mean and variance of X.

8. In throwing a single die, suppose an event E is defined as the appearance of a "five or six". Three dice were thrown 50 times and yielded the following distribution of X (the number of E's per throw):

x	0	1	2	3
f	18	20	11	1

(a) Estimate from the data the probability of getting an E in a single throw of one die.

(b) "Fit" a binomial distribution to the observed distribution, and calculate "fitted" binomial probabilities and "expected" frequencies.

(c) If the dice are true, the probability of an E is what? Use this value of p in the binomial distribution and "fit" this binomial distribution to the observed distribution.

CHAPTER 7. THE POISSON DISTRIBUTION

7.1 The Poisson Distribution as a Limiting Case of the Binomial Distribution

In using the binomial distribution (6.1) one frequently encounters situations in which p is very small (less than .1) and n large (greater than 50) so that the mean np is some "moderate" number (between 0 and 10, say). In such cases there is an approximation for formula (6.1) which is simpler to deal with than (6.1) itself. If we put $np = m$, then the approximate distribution is

$$f(x) = \frac{m^x e^{-m}}{x!} , \tag{7.1}$$

called the Poisson probability distribution, or more briefly, the Poisson distribution, where the possible values of x are 0, 1, 2, 3, 4, ... (to infinity), and where e = 2.71828 ..., the base of natural logarithms. (Actually, e is the limiting value of $(1 + \frac{1}{k})^k$ as k is allowed to become indefinitely large and can be shown to be given by the formula

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots .$$

The value of e can be computed to any desired degree of accuracy by taking the sum of this series to a sufficiently large number of terms.)

7.2 Derivation of the Poisson Distribution.

The argument which gives (7.1) as the approximate distribution runs as follows. Remembering that C_x^n may be written as $\frac{n(n-1)\ldots(n-x+1)}{x!}$, we can write the binomial distribution (6.1) as follows:

$$f(x) = \frac{n(n-1)\ldots(n-x+1)}{x!}\, p^x (1-p)^{n-x} . \tag{7.2}$$

Putting $p = \frac{m}{n}$, we may rewrite the right side of (7.2) as

$$(\tfrac{n-1}{n})\ (\tfrac{n-2}{n})\ \cdots\ (\tfrac{n-x+1}{n})\ (\tfrac{m^x}{x!})\ (1-\tfrac{m}{n})^n\ (1-\tfrac{m}{n})^{-x} . \tag{7.3}$$

Now let us hold x and m fixed, let n become indefinitely large, and see what

happens to each term in this expression. Each of the terms $(\frac{n-1}{n})$, $(\frac{n-2}{n})$, ..., $(\frac{n-x+1}{n})$ and $(1-\frac{m}{n})^{-x}$ has the value 1 as its limiting value. The term $(1-\frac{m}{n})^{n}$ can be written as

$$[(1-\frac{m}{n})^{\frac{n}{m}}]^{m} .$$

If, for the moment, we write $\frac{n}{m}$ as k, this expression can be written as

$$[(1-\frac{1}{k})^{k}]^{m} .$$

If n increases indefinitely, so does k. Hence, the expression inside the square brackets has e^{-1} as its limiting value. Making use of the fact that the limiting value of a product of terms (such as those considered here) is equal to the product of the limiting values of the separate terms, we find that the limiting value of (7.2) is

$$1 \cdot 1 \ldots 1 \cdot \frac{m^{x}}{x!} (e^{-1})^{m} \cdot 1 = \frac{m^{x} e^{-m}}{x!} ,$$

which is the Poisson distribution (7.1).

It should be noted that if n is allowed to approach infinity so that np = m is fixed, then p must approach zero. Hence, the fact that the binomial distribution has the Poisson distribution (7.1) as a limit when n approaches infinity and p approaches zero (so that np is a constant, m) means that the Poisson distribution is an approximation to the binomial distribution for "large" n and "small" p.

To see that the sum of the probabilities in the Poisson distribution is unity, we write

$$(7.4) \qquad \frac{m^{0}}{0!} e^{-m} + \frac{m^{1}}{1!} e^{-m} + \frac{m^{2}}{2!} e^{-m} + \frac{m^{3}}{3!} e^{-m} + \ldots$$

$$= e^{-m} (1 + m + \frac{m^{2}}{2!} + \frac{m^{3}}{3!} + \ldots).$$

But it can be shown that

$$(7.5) \qquad 1 + m + \frac{m^{2}}{2!} + \frac{m^{3}}{3!} + \ldots = e^{m} .$$

Hence the sum of the probabilities is equal to $e^{-m}(e^{m}) = e^{0} = 1$.

7.3 The Mean and Variance of a Poisson Distribution.

The Poisson distribution (7.1) has the following simple property:

The mean and variance are each equal to m.

This may be seen if we proceed as we did in the case of the binomial distribution. We first consider the expansion

$$e^{tm} = 1 + tm + \frac{t^2m^2}{2!} + \frac{t^3m^3}{3!} + \cdots , \tag{7.6}$$

then multiply by e^{-m} and get

$$e^{-m}e^{tm} = e^{-m} + tme^{-m} + \frac{t^2m^2}{2!}e^{-m} + \frac{t^3m^3}{3!}e^{-m} + \cdots . \tag{7.7}$$

Now differentiating with respect to t we find

$$me^{-m}e^{tm} = me^{-m} + \frac{2tm^2}{2!}e^{-m} + \frac{3t^2}{3!}m^3e^{-m} + \cdots . \tag{7.8}$$

By putting t = 1, and using summation notation, we see that the right-hand side of (7.8) is

$$\sum_{x=0}^{\infty} x\,\frac{m^x}{x!}e^{-m} ,$$

which is, by definition, the mean of the Poisson distribution (7.1). But its value must be equal to the left-hand side of (7.8) when t = 1, which is m. Hence, the mean of the Poisson distribution (7.1) is m, i.e.,

$$\mu = m .$$

To find the variance, we first multiply both sides of (7.8) by t, getting

$$mte^{-m}e^{tm} = tme^{-m} + \frac{2t^2m^2}{2!}e^{-m} + \frac{3t^3}{3!}m^3e^{-m} + \cdots . \tag{7.9}$$

Then differentiating with respect to t,

$$(m + m^2t)e^{-m}e^{tm} = me^{-m} + 2\cdot\frac{2tm^2}{2!}e^{-m} + 3\cdot\frac{3t^2m^3}{3!}e^{-m} + \cdots . \tag{7.10}$$

Now putting t = 1, the expression on the right-hand side of (7.10) is

$$\sum_{x=0}^{\infty} \frac{x^2 m^x}{x!} e^{-m} , \tag{7.11}$$

which has a value equal to the left-hand side (after putting t = 1), namely, $m + m^2$. But we know from the general formula for the variance (5.3b) that the variance of the Poisson distribution is (7.11) (which has the value $m + m^2$) minus the square of the mean, i.e.,

$$\sigma^2 = (m + m^2) - (m)^2 = m ,$$

thus establishing the fact that the variance of a Poisson distribution is equal to m.

As a simple example of the use of the Poisson distribution in approximating a binomial distribution for small values of p and large values of n, consider the following

Example: Five coins are tossed 64 times. What are the probabilities of getting five heads 0, 1, 2, 3, 4, 5, ..., 64 times?

The exact values of these probabilities are given by the binomial probability distribution (6.1) for $p = \frac{1}{32}$ (probability of five heads in throwing 5 coins) and n = 64 (the number of throws of five coins), i.e.,

$$f(x) = C_x^{64} \left(\frac{1}{32}\right)^x \left(\frac{31}{32}\right)^{64-x} \tag{7.12}$$

for x = 0, 1, 2, 3, ..., 64.

Approximate values of these probabilities are given by the Poisson distribution, with $m = np = 64 \left(\frac{1}{32}\right) = 2$, i.e.,

$$f(x) = \frac{2^x e^{-2}}{x!} \tag{7.13}$$

for x = 0, 1, 2, 3, ... (to infinity).

The fact that x goes from 0 to 64 in the binomial case and from 0 to ∞ in the Poisson case will probably seem puzzling. But the point to be emphasized is that the Poisson distribution is only an approximation (involving a simple formula) to the binomial distribution; the probabilities are all nearly zero anyway beyond x = 10 in this example. The accuracy of the Poisson approximation and the rapidity with which the probabilities become small as x becomes large are shown in Table 7.1.

TABLE 7.1

Exact Probabilities (from Binomial Distribution) and Approximate Probabilities (from Poisson Distribution) for Obtaining 5 Heads 0, 1, 2, etc., Times in Throwing 5 Coins 64 Times.

x	Exact Probability (from Binomial Distribution (7.12))	Approximate Probability (from Poisson Distribution (7.13))
0	.131	.135
1	.271	.271
2	.275	.271
3	.183	.180
4	.090	.090
5	.035	.036
6	.011	.012
7	.003	.004
8	.001	.001
9	.000	.000

7.4 "Fitting" a Poisson Distribution to a Sample Frequency Distribution.

In the example just given, we had a situation in which we are given the values of n and p and hence the product np (=m) in advance. But there are problems in which we do not have enough information to know m in advance, in which cases it has to be "estimated" experimentally. In such cases, we obtain a frequency distribution of values of X from a sample of "measurements" and "fit" a Poisson probability distribution to the relative frequency distribution, by equating the mean of the sample distribution to the mean m of the Poisson distribution.

As an illustration we shall consider an experiment conducted by Rutherford and Geiger. They counted the number of alpha-particles emitted from a disc in 2608 periods of time, each period of 7.5 seconds duration. The frequencies f of periods in which x particles (x = 0, 1, 2, 3, ...) were counted are shown in column (b) of Table 7.2.

The mean of the distribution of the observed frequencies is

$$\bar{x} = \frac{1}{2608} \sum_{i=0}^{14} x_i f_i = 3.870 ,$$

which is the average number of alpha-particles per 7.5-second interval. The mean of the Poisson distribution is m. Hence, equating m to the observed mean, we have

$$m = 3.870 .$$

The "fitted" Poisson distribution is, therefore, obtained by putting m = 3.87 in formula (7.1), thus giving us

$$f(x) = \frac{(3.87)^x}{x!} e^{-3.87} , \tag{7.14}$$

for x = 0, 1, 2, 3, ..., as the "fitted" distribution. Substituting successively x = 0, 1, 2, 3, 4, ...,14 (probabilities for values of x beyond 14 are too small to be significant in this problem) into formula (7.14), we obtain the "fitted" values in column (d) of Table 7.2.

TABLE 7.2

Fitting of a Poisson Distribution to the Rutherford-Geiger Data on Number of α-particles Emitted per 7.5-second Interval

(a) x_i	(b) Observed Frequency (no. α-particles per 7.5-second interval)	(c) Relative Frequency	(d) "Fitted" Poisson Probability Distribution	(e) "Expected" Frequency (no. α-particles per 7.5-second interval)	(f) Cumulative Observed Frequency	(g) Cumulative "Expected" Frequency
0	57	.0219	.0209	54.4	57	54.4
1	203	.0778	.0807	210.5	260	264.9
2	383	.1468	.1562	407.4	643	672.3
3	525	.2013	.2015	525.5	1168	1197.8
4	532	.2040	.1949	508.4	1700	1706.2
5	408	.1564	.1509	393.5	2108	2099.7
6	273	.1047	.0973	253.8	2381	2353.5
7	139	.0533	.0538	140.3	2520	2493.8
8	45	.0173	.0260	67.9	2565	2561.7
9	27	.0104	.0112	29.2	2592	2590.9
10	10	.0038	.0043	11.3	2602	2602.2
11	4	.0015	.0015	4.0	2606	2606.2
12	2	.0008	.0005	1.3	2608	2607.5
13	0	.0000	.0001	.4	2608	2607.9
14	0	.0000	.0000	.1	2608	2608.0
Total	2608	1.0000	1.0000	2608.0		

The closeness of the fit of the Poisson distribution to the (observed) relative frequencies is clear from a comparison of the figures in column (d) with those in column (c), or from a comparison of those in column (e) with those in column (b), or from a comparison of those in column (g) with those in column (f).

A question which one would normally ask is this: Since $m = np$, what are n and p in this alpha-particle counting problem? The answer is this: Think of the number of atoms in the radio-active material as being n and the probability of an atom emitting an alpha particle in a 7.5-second interval as being p. Hence, for a single 7.5-second interval we can think of there being a very large number n of trials for each of which there is a very small probability p of an alpha-particle being emitted. Hence, the mean number of occurrences of alpha-particles in a 7.5-second interval is np. But an experimental value of np has been determined, namely 3.87, i.e., $np = 3.87$. Thus

$$p = 3.87/n \ .$$

The probability that x of the atoms will emit alpha particles in a specified 7.5-second interval is given by the binomial distribution

$$(7.15) \qquad C_x^n \left(\frac{3.87}{n}\right)^x \left(1 - \frac{3.87}{n}\right)^{n-x} \ .$$

Since n, the number of atoms, is very large, it follows (by the same argument used earlier in this section to establish (7.1) as the limiting value of the binomial distribution (6.1)) that the limiting value of (7.15) is the Poisson distribution (7.14).

It should be noted that in fitting the Poisson distribution to the Rutherford-Geiger data, it is not necessary to know the value of n or p individually; it is sufficient to know only np (= m), and this is determined experimentally.

This is true in many situations, i.e., it is sufficient to know only np (= m) and not n and p separately. For instance, suppose that on the average 3 nails out of 100 manufactured by an automatic nail machine are defective and that the nails are packaged in boxes of 200. Then the average number of defective nails per box of 200 is 6 (i.e., $m = np = 6$). Hence, we can say that the probability of a box containing x defective nails is

$$\frac{6^x e^{-6}}{x!}$$

for $x = 0, 1, 2, 3, \ldots$. In writing down this distribution note that it was not necessary to use n and p individually; we used only the product np. Actually $n = 200$ and $p = \frac{3}{100}$. To take another example, suppose a wool loom leaves an average of 1 defect of a certain type per 100 square yards of woolen cloth. What is the probability that a piece of woolen cloth consisting of 10 square yards has x defects in it? We must first find the mean number of defects per 10 square yards. This is clearly .1, i.e., $m = .1$. Hence the required probability is

$$\frac{(.1)^x e^{-.1}}{x!} .$$

In this case we may think of dividing the material into a large number n of small areas, and think of p as the probability of a defect in a given small area. Evidently n would be large and p small. But the important thing is that the mean number of defects per unit area be known. It is 1 defect per 100 square yards, or .1 defect per 10 square yards.

The Poisson distribution occurs in many situations involving events occurring in time intervals of fixed length, space of fixed volume, areas of fixed size, line segments of fixed length, etc. Such examples are as follows:

(a) Distribution of numbers of telephone calls received at a given switchboard per minute (for a large number of minutes) for a given part of the day.

(b) Distribution of numbers of automobiles passing a given point on a highway per minute (for a large number of minutes) at a given time of the day.

(c) Distribution of numbers of bacterial colonies in a given culture per .01 square millimeter (for a large number of units of .01 sq. mm.) on a microscope slide.

(d) Distribution of numbers of deaths per day (for a large number of days) by heart attack in a large city.

(e) Distribution of numbers of typographical errors per page (for a large number of pages) in typed material.

(f) Distribution of numbers of fragments per square foot (for many square foot units) received by a test surface exposed to a fragmentation bomb at a given distance from the detonated bomb.

(g) Distribution of numbers of times one received four aces per 75 hands of bridge (for a large number of sets of 75 hands).

(h) Distribution of number of defective screws per box of 100 screws (for a large number of boxes).

Exercise 7.

The following table of values of e^{-m} is given for convenience in working the problems in this Exercise:

TABLE 7.3

Table of Values of e^{-m}

m	e^{-m}	m	e^{-m}
.1	.9048	1	.3679
.2	.8187	2	.1353
.3	.7408	3	.0498
.4	.6703	4	.0183
.5	.6065	5	.0067
.6	.5488	6	.0025
.7	.4966	7	.0009
.8	.4493	8	.0003
.9	.4066	9	.0001

1. Three dice are rolled 216 times. Using the Poisson distribution as an approximation to the binomial distribution, write down the approximate probability of getting 3 aces x times. Work out the probability distribution of X to two decimal places. What is the chance variable X here?

2. Using the Poisson distribution, what is the probability that if a person plays 76 hands of bridge he will get four aces x times? Make a probability distribution of X in table form. What is the chance variable X?

3. Suppose a certain automatic screw machine produces one slotless screw on the average out of every 100 screws. The screws are packaged in boxes of 100. Using the Poisson distribution as an approximation to the binomial distribution, what is the probability that a specified box will have x slotless screws? What is the chance quantity X? Using the Poisson distribution, write down the probability distribution of X (to three decimal places). What fraction of boxes (of 100 screws) would you estimate to have no slotless screws? Not more

than 2 slotless screws?

4. A direct mail advertising firm finds that on the average one person out of 100 persons in small middle-western towns will send in orders for a certain article from a mail advertisement. If the firm sends 50 letters to persons in each of 200 such towns, from what percentages of the towns can the firm expect to receive 0, 1, 2, 3, 4, 5, ... orders?

5. Suppose there is an average of one typographical error per ten pages of page proof of a certain book. What percentages of pages would you estimate to have 0, 1, 2, 3, 4, ... errors? What is the probability that a 20-page chapter will contain no errors?

6. Suppose that in normal summer driving in New Jersey a driver has an average of one puncture per 2,000 miles. What is the probability that the driver will have x punctures in making a 1,000 mile trip? Write down the probability distribution of X in table form.

7. The probability that a man aged 35 will die before reaching the age of 40 is .018. What is the probability that x of the 50 alumni, 35 years old, of the class of 1935 of college Z will die within five years? Write down the probability distribution of X in table form.

8. The following table shows the distribution of numbers of vacancies occuring per year in the U. S. Supreme Court by years from 1837 to 1932 (data compiled by Wallis):

No. vacancies per year	Frequency
0	59
1	27
2	9
3	1

Fit a Poisson distribution to this observed distribution.

9. The following table shows the distribution of number of articles turned in per day to the lost and found bureau of a large office building for a period of

423 days, excluding summer months, Sundays and holidays, (data compiled by Thorndike):

No. of articles per day	Frequency
0	169
1	134
2	74
3	32
4	11
5	2
6	0
7	1

Fit a Poisson distribution to these data.

10. The following table shows the distribution of number of deaths of soldiers in individual Prussian cavalry corps due to kicks from horses in 200 corps-years (data compiled by Bortkiewicz):

No. deaths per corps-year	Frequency
0	109
1	65
2	22
3	3
4	1

Fit a Poisson distribution to these data.

CHAPTER 8. THE NORMAL DISTRIBUTION

8.1 General Properties of the Normal Distribution.

The most important continuous probability distribution is the normal or Gaussian distribution which has been referred to in several previous sections, particularly in Section 3.2. The cumulative distribution function $F_N(x)$ for a normal distribution having mean μ and standard deviation σ is given by the formula

$$F_N(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx . \tag{8.1}$$

We usually refer to $F_N(x)$ as given by (8.1) as the cumulative normal distribution with mean μ and standard deviation σ. If X is a (continuous) chance quantity having a normal distribution with mean μ and standard deviation σ, then for any specified value of x, say x', $F_N(x')$ is simply the probability that $X \leq x'$ or more briefly

$$\Pr(X \leq x') = F_N(x') . \tag{8.2}$$

Since the normal distribution is the distribution of a continuous chance quantity, we can replace $\leq$ by $<$ without affecting (8.2). If x' and x'' (x' < x'') are any two specified values of x, then (formula (5.9))

$$\Pr(x' < X \leq x'') = F_N(x'') - F_N(x') . \tag{8.3}$$

The probability density function $f_N(x)$ which is obtained by differentiating $F_N(x)$ with respect to x (see formula (5.13)), is

$$f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} . \tag{8.4}$$

For $x = -\infty$, $F_N(x) = 0$, and for $x = +\infty$, $F_N(x) = +1$. But for all practical purposes, the value of $F_N(x)$ ranges from "nearly" 0 to "nearly" 1 as

TABLE 8.1

Values of the Cumulative Normal Distribution $F_N(x)$

z	x	$F_N(x)$	z	x	$F_N(x)$
-3.0	$\mu - 3.0\sigma$	.0013	.1	$\mu + .1\sigma$	.5398
-2.9	$\mu - 2.9\sigma$	.0019	.2	$\mu + .2\sigma$	.5793
-2.8	$\mu - 2.8\sigma$	.0026	.3	$\mu + .3\sigma$	.6179
-2.7	$\mu - 2.7\sigma$	.0035	.4	$\mu + .4\sigma$	.6554
-2.6	$\mu - 2.6\sigma$	.0047	.5	$\mu + .5\sigma$	.6915
-2.5	$\mu - 2.5\sigma$	.0062	.6	$\mu + .6\sigma$	.7256
-2.4	$\mu - 2.4\sigma$	.0082	.7	$\mu + .7\sigma$	.7580
-2.3	$\mu - 2.3\sigma$	.0107	.8	$\mu + .8\sigma$	.7881
-2.2	$\mu - 2.2\sigma$	.0139	.9	$\mu + .9\sigma$	.8159
-2.1	$\mu - 2.1\sigma$	.0179	1.0	$\mu + 1.0\sigma$	.8413
-2.0	$\mu - 2.0\sigma$	.0227	1.1	$\mu + 1.1\sigma$	.8643
-1.9	$\mu - 1.9\sigma$	.0287	1.2	$\mu + 1.2\sigma$	.8849
-1.8	$\mu - 1.8\sigma$	.0359	1.3	$\mu + 1.3\sigma$	.9032
-1.7	$\mu - 1.7\sigma$	.0446	1.4	$\mu + 1.4\sigma$	.9192
-1.6	$\mu - 1.6\sigma$	.0548	1.5	$\mu + 1.5\sigma$	.9332
-1.5	$\mu - 1.5\sigma$	.0668	1.6	$\mu + 1.6\sigma$	.9452
-1.4	$\mu - 1.4\sigma$	.0808	1.7	$\mu + 1.7\sigma$	.9554
-1.3	$\mu - 1.3\sigma$	.0968	1.8	$\mu + 1.8\sigma$	.9641
-1.2	$\mu - 1.2\sigma$	.1151	1.9	$\mu + 1.9\sigma$	.9713
-1.1	$\mu - 1.1\sigma$	.1357	2.0	$\mu + 2.0\sigma$	.9773
-1.0	$\mu - 1.0\sigma$	.1587	2.1	$\mu + 2.1\sigma$	.9821
-.9	$\mu - .9\sigma$	.1841	2.2	$\mu + 2.2\sigma$	.9861
-.8	$\mu - .8\sigma$	.2119	2.3	$\mu + 2.3\sigma$	.9893
-.7	$\mu - .7\sigma$	.2420	2.4	$\mu + 2.4\sigma$	.9918
-.6	$\mu - .6\sigma$	.2742	2.5	$\mu + 2.5\sigma$	.9938
-.5	$\mu - .5\sigma$	.3085	2.6	$\mu + 2.6\sigma$	.9953
-.4	$\mu - .4\sigma$	.3446	2.7	$\mu + 2.7\sigma$	.9965
-.3	$\mu - .3\sigma$	.3821	2.8	$\mu + 2.8\sigma$	.9974
-.2	$\mu - .2\sigma$	.4207	2.9	$\mu + 2.9\sigma$	.9981
-.1	$\mu - .1\sigma$	.4602	3.0	$\mu + 3.0\sigma$	.9987
0	μ	.5000			

x ranges from $\mu-3\sigma$ to $\mu+3\sigma$. Values of $F_N(x)$ for values of x ranging from $\mu-3\sigma$ to $\mu+3\sigma$ by intervals of $.1\sigma$ are given in Table 8.1.

It will be convenient for applications of the normal distribution to write $\frac{x-\mu}{\sigma} = z$ (or $x = \mu + z\sigma$) and consider values of z corresponding to any chosen value of x and vice versa. A column of values of z, for the various values of x considered, is given in Table 8.1.

If we think of X as a chance quantity with mean μ and standard deviation σ, then Z (where $Z = \frac{X-\mu}{\sigma}$) is a chance quantity having mean 0 and standard deviation 1. As a matter of fact, this statement is true whether X (and Z) have normal distributions or not. For example, suppose X is a chance quantity denoting the total number of heads obtained in throwing 10 coins. Then we know from formula (6.5) that the mean of X is 5 (i.e., $\mu = 5$), and from formula (6.7) that the standard deviation of X is $\sqrt{10\cdot\frac{1}{2}\cdot\frac{1}{2}} = 1.581$ (i.e., $\sigma = 1.581$). If we set $Z = \frac{X-5}{1.581}$, then Z will be a chance quantity having mean 0 and standard deviation 1. In this example, X (and also Z) is a discrete chance quantity, but its cumulative probability graph can be closely approximated by a cumulative normal distribution, as we shall see in Section 8.22.

It should be noticed from Table 8.1 that the values of $F_N(x)$ corresponding to $\mu \pm z\sigma$ are symmetrical with respect to .5000 and the sum of the two values of $F_N(x)$ is 1. For example, for $x = \mu \pm 1.5\sigma$ (i.e., for $z = \pm 1.5$) we have $F_N(x) = .5000 \pm .4332$. The sum of these two values of $F_N(x)$ is clearly 1.

If a chance quantity X is known to have a normal distribution with a specified mean μ and standard deviation σ, then one can determine the probability of X falling into a given interval from Table 8.1.

Example: Suppose X is a chance quantity having a normal distribution with mean 30 and standard deviation 5. What is the probability that $26 < X \leq 40$?

We have

$$\Pr(26 < X \leq 40) = F_N(40) - F_N(26) .$$

To find the values of $F_N(40)$ and $F_N(26)$, we must make use of the relationship between x and z for this problem. Since $\mu = 30$ and $\sigma = 5$, we have

$$z = \frac{x-30}{5} .$$

The values of z corresponding to 40 and 26 are $\frac{40-30}{5} = 2.0$ and $\frac{26-30}{5} = -.8$. The values of $F_N(40)$ and $F_N(26)$ are given by entering Table 8.1 for $z = 2.0$ and $z = -.8$, respectively. We find, therefore, that

$$F_N(40) = .9773, \quad F_N(26) = .2119 .$$

Hence we have

$$\Pr(26 < X \leq 40) = \Pr(-.8 < Z \leq 2.0) = .9773 - .2119 = .7654 .$$

In applying the normal distribution, it is often convenient to talk about the probability of X departing not more than (or departing more than) by a specified multiple of σ from the mean μ. For example, what is the probability that X will differ from μ by not more than σ ? By this we mean: What is the value of $\Pr(\mu - \sigma \leq X \leq \mu + \sigma)$? Or more briefly expressed: What is the value of $\Pr(|X-\mu| \leq \sigma)$? We have

$$\Pr(\mu - \sigma \leq X \leq \mu + \sigma) = \Pr(-1 < Z \leq 1)$$

$$= F_N(\mu + \sigma) - F_N(\mu - \sigma)$$

$$= .8413 - .1587 = .6826 ,$$

or more briefly

$$\Pr(|X - \mu| \leq \sigma) = .6826 .$$

Similarly, we see from Table 8.1

$$\Pr(|X - \mu| \leq 2\sigma) = \Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \Pr(-2 < Z \leq 2) = .9546$$

and

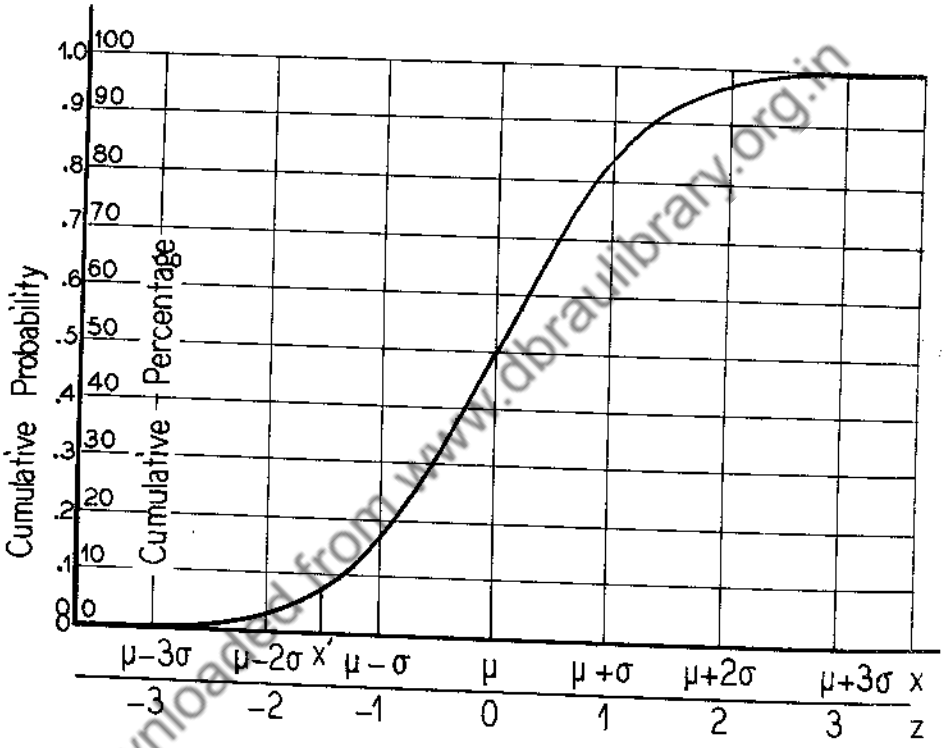
$$\Pr(|X - \mu| \leq 3\sigma) = \Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = \Pr(-3 < Z \leq 3) = .9974 .$$

Expressed verbally and statistically we may say this:
If the measurements in a population are normally distributed with mean μ and standard deviation σ, then

68.26% of the measurements deviate less than 1σ from μ,
95.46% of the measurements deviate less than 2σ from μ,
99.74% of the measurements deviate less than 3σ from μ.

This is a more precise form of the statements we made in Section 3.2.

The graph of $F_N(x)$ is shown in Figure 8.1, with the scales of both x and z indicated. For convenience, we place the z scale on a line slightly below the x scale. In applications, the x values which are actually marked off on the x axis are conveniently chosen values which are not necessarily those shown in Figure 8.1, but it is usually convenient to show at least the following values of z: -3, -2, -1, 0, 1, 2, 3.



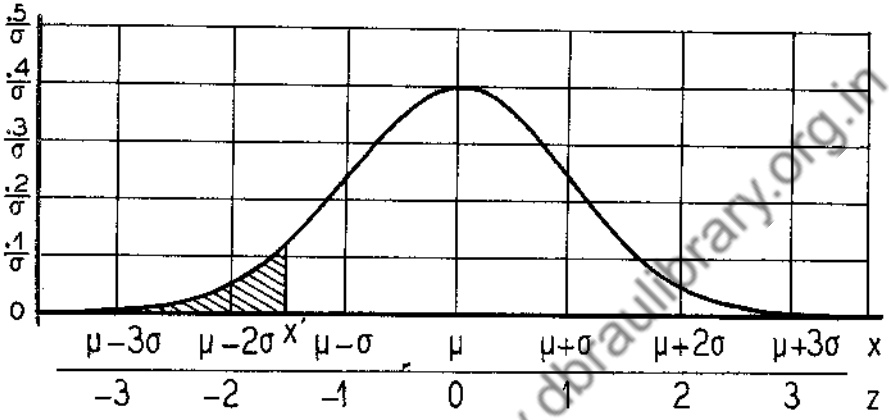
Graph of the Cumulative Normal Distribution $F_N(x)$

Figure 8.1

The graph of the probability density function $f_X(x)$ as given by (8.4) is shown in Figure 8.2. The relationship between the graph of $f_N(x)$ in Figure 8.2 and that of $F_N(x)$ in Figure 8.1 is exactly the same as that between the graph in Figure 5.8 and the graph in Figure 5.6. To repeat: the numerical value of the ordinate of the graph in Figure 8.1 at any value, say x', is equal to the numerical value of the area under the curve in Figure 8.2 to the left

of x' (the shaded area).

Actually, we shall make very little direct use of the probability density function $f_N(x)$. Figure 8.2 is given merely to show how the graph of $f_N(x)$ looks. In practical applications of the normal distribution it is always more convenient to work with the cumulative normal distribution $F_N(x)$.



Graph of the Normal Probability Density Function $f_N(x)$

Figure 8.2

8.2 Some Applications of the Normal Distribution.

The normal probability distribution is very widely used in probability and statistics problems. It is used mostly for the following purposes:

(1) To approximate or "fit" a distribution of measurements in a sample under certain conditions.

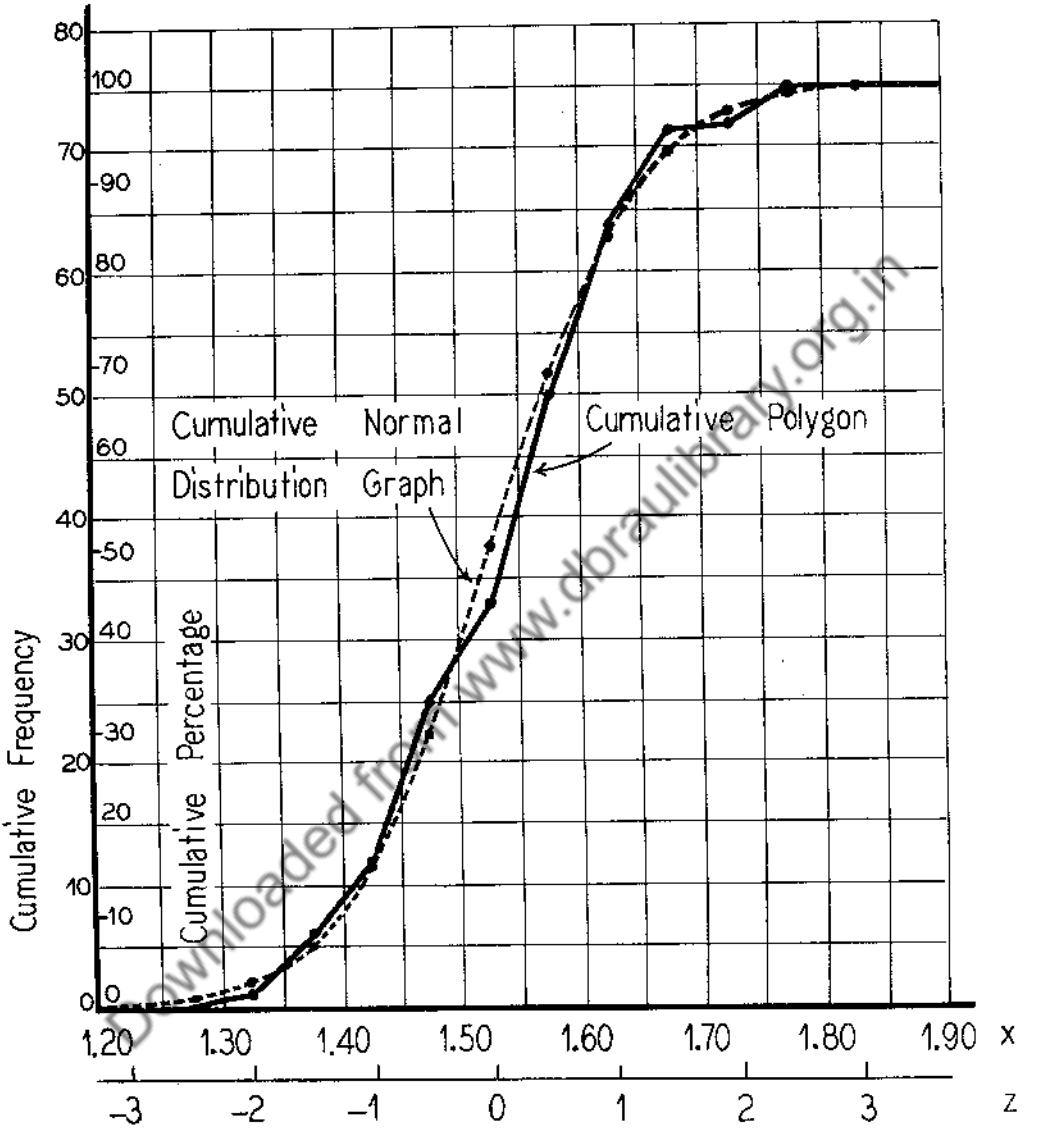
(2) To approximate the binomial distribution and other discrete or continuous probability distributions under suitable conditions.

(3) To approximate the distribution of means and certain other quantities calculated from samples, especially large samples.

In the present section we shall discuss (1) and (2). But (3) must be reserved for discussion in Chapter 9.

8.21 "Fitting" a cumulative distribution of measurements in a sample by a cumulative normal distribution.

It is very often true that samples of measurements are such that their cumulative frequency polygons (in case of grouped data) or cumulative graphs (for



"Fitted" Cumulative Normal Distribution Graph for the Cumulative Polygon of Figure 2.4

Figure 3.3

ungrouped data) have a shape which can be fairly closely approximated by the graph of the cumulative normal distribution one obtains by replacing μ in (8.1) by the sample mean $\overline{X}$ and σ in (8.1) by the sample standard deviation s_x. For example, let us consider the cumulative polygon shown in Figure 2.4, which is plotted from the cumulative frequencies in Table 2.2. You will recall from Section 3.31 that

$$\overline{X} = 1.527$$

$$s_x = .101 .$$

Using these values for μ and σ respectively, we have the following relationship between z and x:

$$z = \frac{x - 1.527}{.101} . \tag{8.5}$$

In "fitting" the cumulative normal distribution, it will be sufficient for our purpose to consider the following values of z: -3, -2, -1, 0, 1, 2, 3. The values of x corresponding to those values of z as determined from formula (8.5) and the values of $F_N(x)$ are given in Table 8.2. You will note that x is measured in ounces, but that z is a "dimensionless" number.

TABLE 8.2

Values of the Cumulative Normal Distribution $F_N(x)$ when Fitted to the Cumulative Polygon in Figure 2.4

z	x	$F_N(x)$
-3	1.224	.0013
-2	1.325	.0227
-1	1.426	.1587
0	1.527	.5000
1	1.628	.8413
2	1.729	.9773
3	1.830	.9987

Plotting the values of $F_N(x)$ shown in Table 8.2 and drawing a smooth curve through them, we obtain the "fitted" cumulative normal distribution graph shown in Figure 8.3. The cumulative polygon is also shown.

If one wanted to do a more accurate job of constructing the "fitted" graph, he would use more closely spaced values of z and, of course, more of them; e.g., z = -3, -2.5, -2, -1.5, ..., 2.5, 3.

The degree of "goodness of fit" depends, of course, on how near the "fitted" curve and cumulative polygon are to each other. By reading ordinates off the "fitted" curve at upper boundaries of cells, one obtains "fitted" cumulative frequencies. For example, we find from Figure 8.3 the "observed" cumulative frequencies are those shown in Table 8.3 (reading to the nearest .5).

TABLE 8.3

Comparison of Observed and Fitted Frequencies from Figure 8.3

cell midpoint	observed cumulative frequency	"fitted" cumulative frequency	observed frequency	"fitted" frequency
1.25	0	.5	0	.5
1.30	1	1.5	1	1.0
1.35	6	5.0	5	3.5
1.40	12	11.5	7	6.5
1.45	25	23.5	13	12.0
1.50	33	37.5	8	14.0
1.55	50	50.5	17	13.0
1.60	64	63.0	14	12.5
1.65	71	70.0	7	7.0
1.70	72	73.5	1	3.5
1.75	75	74.5	3	1.0
1.80	75	75.0	0	.5
Total			75	75

8.22 "Fitting" a cumulative binomial distribution by a cumulative normal distributio

In Chapter 7 we discussed the Poisson distribution as a simplified approximation to the binomial distribution when n is "large" and p is "small". We can use the normal distribution to very good advantage in approximating the binomial distribution under other conditions, particularly when p is not too "close" to 0 or 1, and when n is "large" or even "moderately large", or roughly

speaking, when np is at least 5. It turns out that for a specified value of p, no matter how small, the cumulative normal distribution provides an approximation to the cumulative binomial distribution which gets better and better as n increases, the approximation becoming perfect in the limit as n increases indefinitely.

Now how do we actually use the normal distribution to approximate the binomial distribution? As mentioned earlier, we approximate the cumulative binomial distribution by means of the cumulative normal distribution. Or graphically expressed: we approximate the cumulative binomial probability graph (a step-like graph of the type shown in Figure 5.3) by the cumulative normal distribution graph, (a smooth curve of the type shown in Figure 8.1).

The binomial probability distribution is (6.3), and its cumulative distribution $F_B(x)$ for any value x' is given by

$$F_B(x') = C_0^n p^0 q^n + C_1^n p^1 q^{n-1} + \ldots + C_{[x']}^n p^{[x']} q^{n-[x']} \tag{8.6}$$

where $[x']$ is the largest integer which does not exceed x' .

You will remember from Chapter 6 that the mean and standard deviation of the binomial distribution (6.3) are np and $\sqrt{npq}$ respectively. The cumulative normal distribution $F_N(x)$ which approximates the cumulative binomial distribution given by (8.6) is given by making the following substitutions in (8.1): $\mu = np$, and $\sigma = \sqrt{npq}$. More explicitly, the approximating cumulative normal distribution function $F_N(x)$ is such that, for a given value x' ,

$$F_N(x') = \frac{1}{\sqrt{2\pi}\sqrt{npq}} \int_{-\infty}^{x'} e^{-\frac{1}{2npq}(x-np)^2} dx \ . \tag{8.7}$$

In other words, $F_N(x')$ is an approximation to $F_B(x')$. But the real question is this: How good is this approximation? To give a full mathematical discussion to this question is a difficult matter which is beyond the scope of this course. At this point it may be sufficient to give two examples and to make this statement: For any specified value of p and x', the difference between $F_N(x')$ and $F_B(x')$ approaches zero as n increases indefinitely.

Example 1: Let us fit a cumulative normal distribution to the cumulative binomial distribution in the case in which n = 10, $p = \frac{1}{2}$.

We calculate the binomial probabilities from the formula (6.3) by putting $n = 10$, $p = \frac{1}{2}$, i.e., from

$$f(x) = C_x^{10}(\tfrac{1}{2})^{10} \tag{8.8}$$

for $x = 0, 1, 2, 3, 4, \ldots, 10$. You will remember that these are the probabilities of getting 0, 1, 2, 3, 4, ... 10 heads in tossing 10 coins. We get the probability distribution $f_B(x)$ and cumulative probability distribution $F_B(x)$ shown in Table 8.4. Now we need to calculate values of the cumulative normal distribution for several values of z. In this problem we shall consider all values of z from -3 to +3 by intervals of .5. For values of μ and σ we have

$$\mu = np = 5$$

$$\sigma = \sqrt{npq} = \sqrt{10 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 1.581.$$

The relationship between x and z is given by

$$z = \frac{x-5}{1.581}.$$

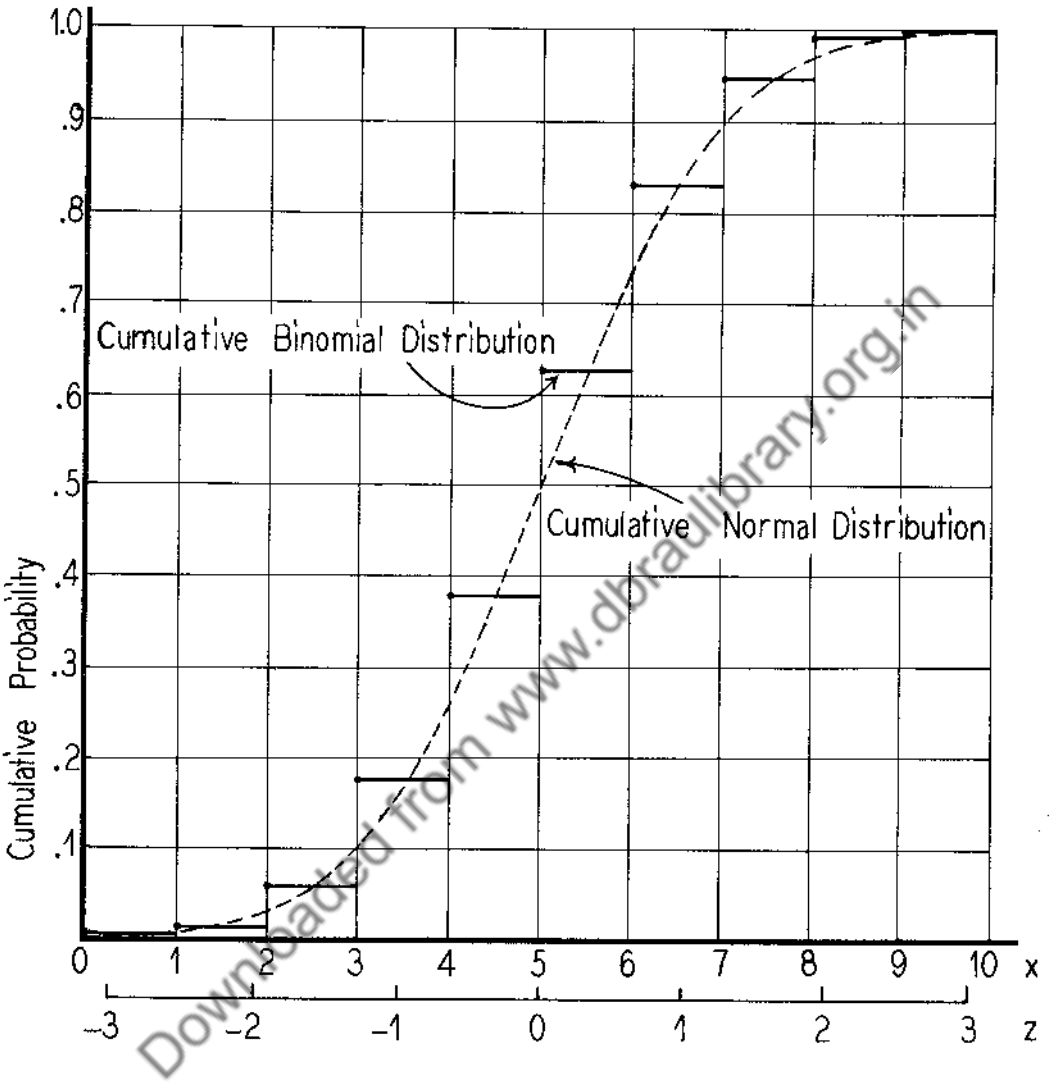
Values of z, x and $F_N(x)$ are given in Table 8.5.

The graphs of the cumulative binomial distribution $F_B(x)$ and the cumulative normal distribution $F_N(x)$ are shown in Figure(8.1)for the case $n = 10$, $p = \frac{1}{2}$.

TABLE 8.4

Binomial Distribution and Cumulative Binomial Distribution for $n = 10$, $p = \frac{1}{2}$

x	$f_B(x)$	$F_B(x)$
0	.001	.001
1	.010	.011
2	.044	.055
3	.117	.172
4	.205	.377
5	.246	.623
6	.205	.828
7	.117	.945
8	.044	.989
9	.010	.999
10	.001	1.000
Total	1	



Graphs of the Cumulative Binomial Distribution $F_B(x)$ for $n = 10$, $p = \frac{1}{2}$ (Table 8.4) and of the Approximating Cumulative Normal Distribution $F_N(x)$ (Table 8.5)

Figure 8.4

TABLE 8.5

Fitted Cumulative Normal Distribution for the Cumulative Binomial Distribution for $n = 10$, $p = \frac{1}{2}$

z	x	$F_N(x)$
-3.0	.257	.0013
-2.5	1.048	.0062
-2.0	1.838	.0227
-1.5	2.628	.0668
-1.0	3.419	.1587
-.5	4.210	.3085
0	5.000	.5000
.5	5.791	.6915
1.0	6.581	.8413
1.5	7.372	.9332
2.0	8.162	.9773
2.5	8.953	.9938
3.0	9.743	.9987

To help us assess how well the cumulative normal distribution approximates the cumulative binomial distribution for other values of p, let us consider the following example.

Example 2: Fit a cumulative normal distribution to the cumulative binomial distribution for $n = 50$, $p = .1$.

In this case

$$\mu = np = 50(.1) = 5$$

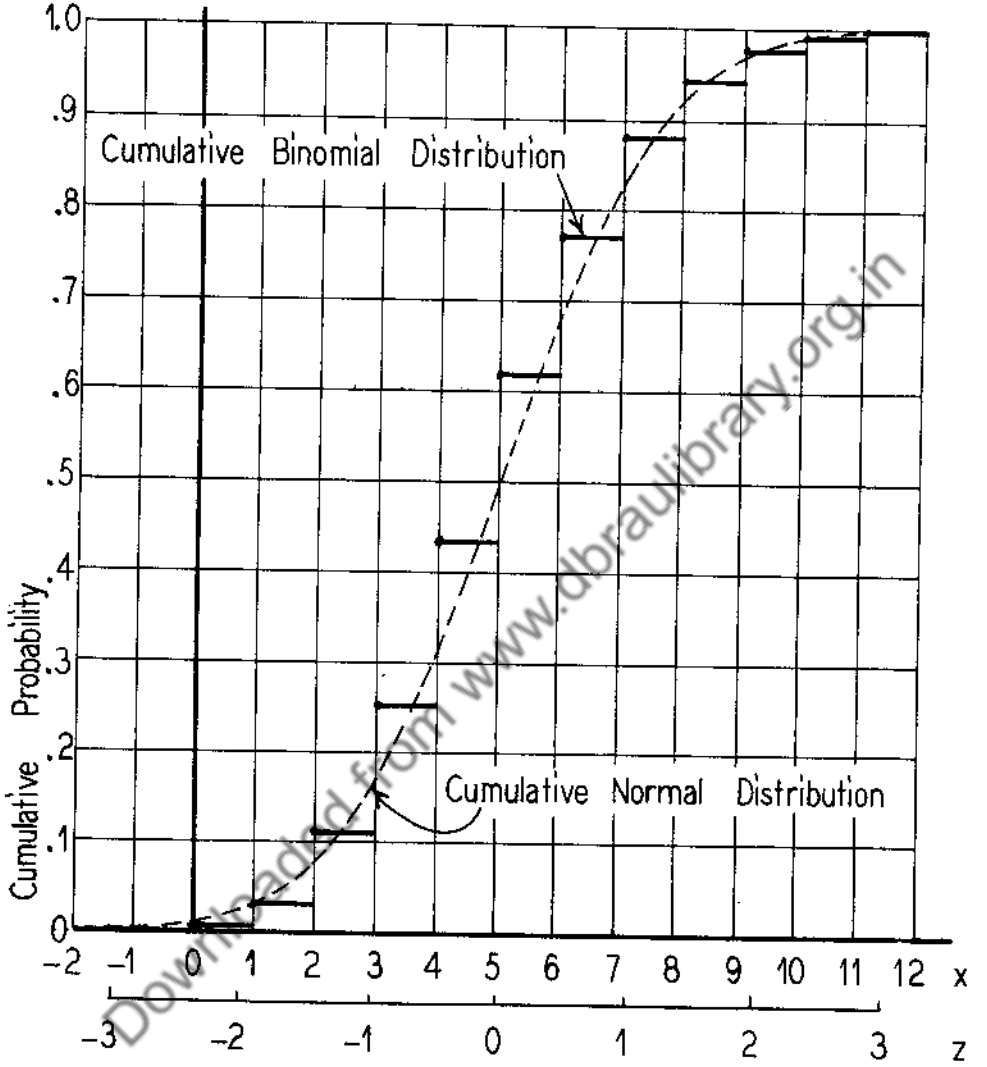
$$\sigma = \sqrt{npq} = \sqrt{50(.1)(.9)} = 2.1213$$

and

$$z = \frac{x-5}{2.1213}.$$

To actually carry out the arithmetical work of "fitting" we would need tables similar to Table 8.4 and Table 8.5. To avoid unnecessary repetition, we shall not write down the tables here. We simply present the results graphically in Figure 8.5 (which is similar to Table 8.4). The "fit" is seen to be very satisfactory.

One may easily ask what point there is in fitting a cumulative binomial distribution by means of a cumulative normal distribution. The answer



Graphs of $F_B(x)$ and of $F_N(x)$ for the case $n = 50$, $p = .1$

Figure 8.5

is that it is easier to calculate probabilities by using a cumulative normal distribution than by using a binomial distribution. For instance, to approximate the value of

$$f_B(x) = C_x^{10}\left(\frac{1}{2}\right)^{10} \tag{8.8}$$

for a single value of x, say x = 6, by using the normal approximation, we would proceed as follows: referring to Figure 8.4, we note that $f_B(6)$, which is the size of the jump in the step-like graph of the cumulative binomial distribution, is approximately equal to the difference $F_N(6.5) - F_N(5.5)$, i.e., the difference between the ordinates of the fitted cumulative normal distribution at x = 5.5 and 6.5. This gives us the approximate result $f_B(6) \cong F_N(6.5) - F_N(5.5)$.

To approximate the value of a sum of terms of the binomial distribution (8.8), e.g., for x = 4, 5, 6, we would have

$$f_B(4) + f_B(5) + f_B(6)$$

approximately equals

$$[F_N(4.5) - F_N(3.5)] + [F_N(5.5) - F_N(4.5)] + [F_N(6.5) - F_N(5.5)]$$

$$= F_N(6.5) - F_N(3.5) .$$

In this example $n = 10$, $p = \frac{1}{2}$, $\mu = 5$, $\sigma = 1.581$. The values of z corresponding to x = 6.5 and x = 3.5 are $\frac{6.5-5}{1.581} = .95$, and $\frac{3.5-5}{1.581} = -.95$, respectively. Since Table 8.1 gives values of z to only one decimal place, we interpolate to obtain approximate values of $F_N(6.5)$ and $F_N(3.5)$. We find $F_N(3.5) \cong .1714$ (actually $F_N(3.5) = .1711$) and $F_N(6.5) \cong .8286$, and hence we obtain $F_N(6.5) - F_N(3.5) = .657$. The exact value of $f_B(4) + f_B(5) + f_B(6)$ from Table 8.4 is seen to be .205 + .246 + .205 = .656, which is close to the approximate value .657, obtained by taking the difference $F_N(6.5) - F_N(3.5)$. With sufficiently extensive tables of values of z, one can easily and rapidly find the approximate value of the sum of any number of consecutive terms of a binomial distribution by simply taking a difference of the fitted cumulative normal distribution for two values of x, assuming, of course, that the conditions of satisfactory approximation are satisfied.

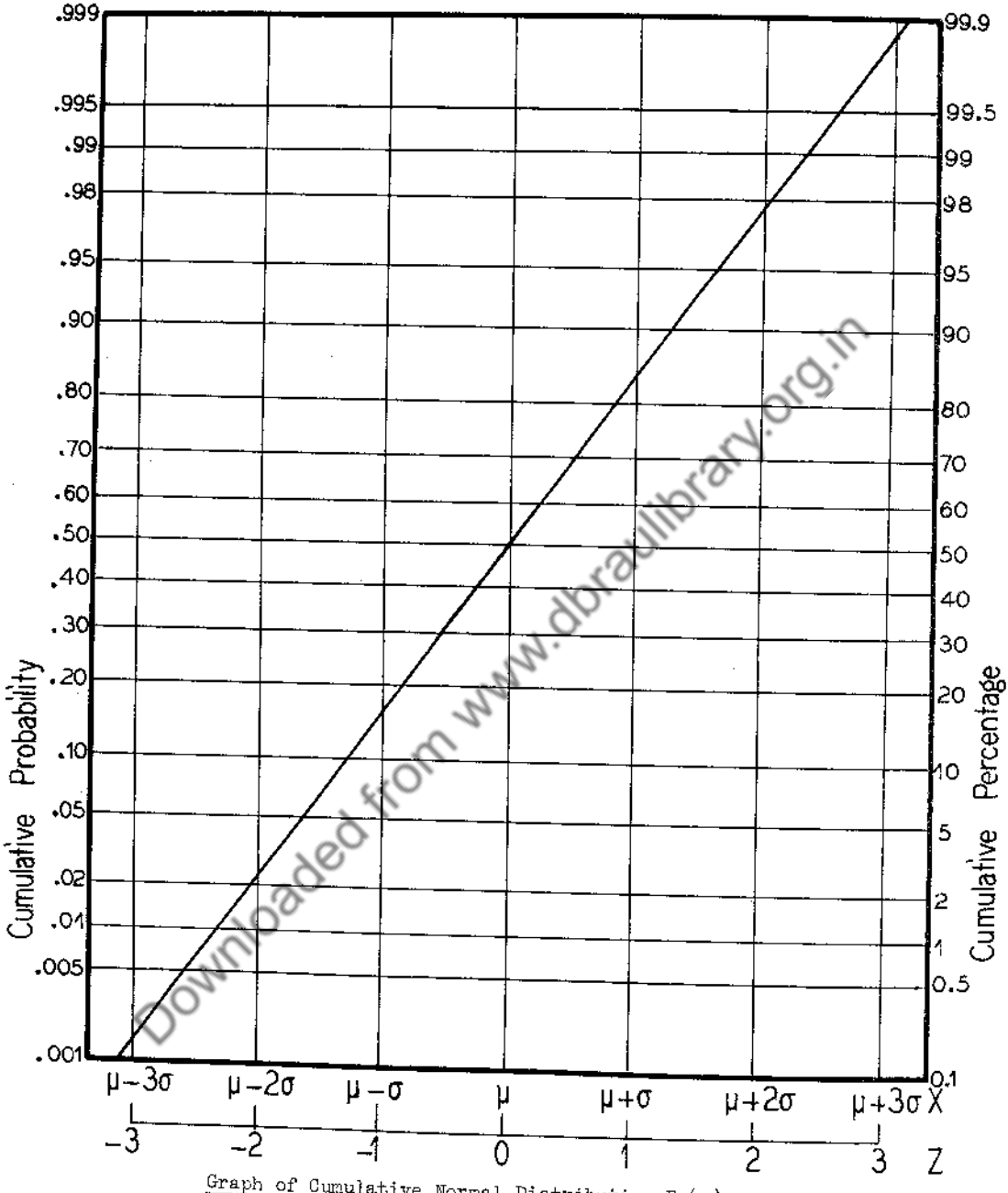
To get some notion of the gain in accuracy one obtains by evaluating

$F_N(x)$ at points halfway between integers, let us consider the case of $n = 400$ and $p = \frac{1}{2}$. For instance, if one wanted to obtain an approximation to the probability that if a coin is tossed 400 times, the number of heads will lie between 196 and 205 inclusive, we might use $F_N(205) - F_N(195)$ as an approximation. Here $\mu = np = 200$, $\sigma = \sqrt{400 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 10$, $z = \frac{x-200}{10}$, and the two values of z corresponding to $x = 195$ and 205 are $-.5$ and $+.5$, respectively. Hence, the approximate value of the desired probability is $.6915 - .3085 = .3830$. The closer approximation, using points halfway between integers, would be given by $F_N(205.5) - F_N(194.5)$. The values of z would be $-.55$ and $+.55$ and the approximate probability would be $.7088 - .2912 = .4176$. The difference between the two approximations is $.4176 - .3830 = .035$. The difference would be smaller if the approximations involved evaluations of $F_N(x)$ for values of x further away from the mean 200 than those just considered, i.e., further from 200 than 195 or 205. In general, unless n is larger than about 400, the accuracy gained by using half-integer positions is worthwhile.

8.3 The Cumulative Normal Distribution on Probability Graph Paper.

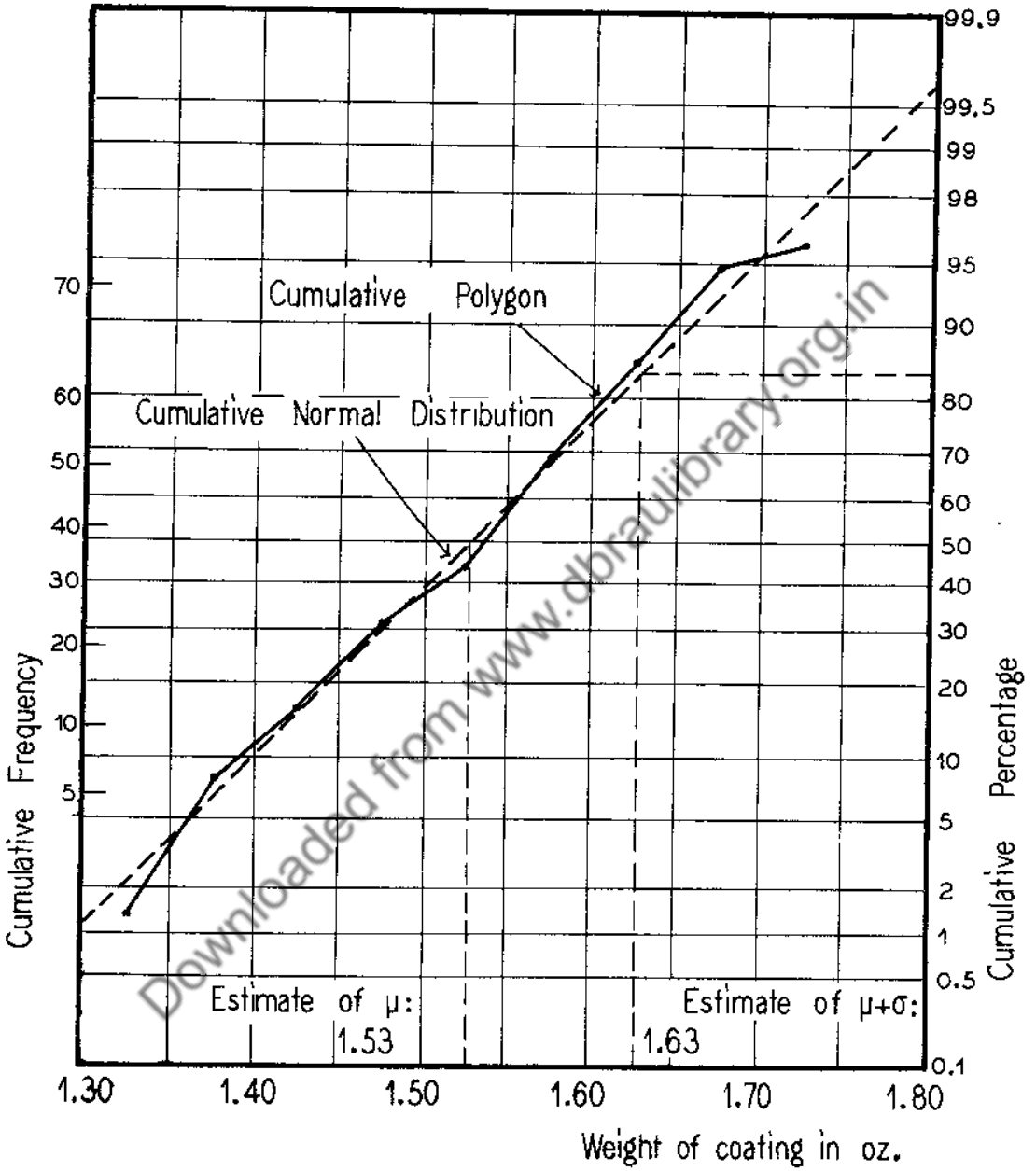
You will recall from Section 2.3 that reference was made to probability graph paper which had the property of making cumulative polygons appear as approximately straight lines. This is accomplished by the way in which the percentage scales for low percentages and high percentages are stretched. Now if the cumulative normal distribution $F_N(x)$ is graphed on probability paper, the graph obtained is exactly a straight line. Figure 8.6 shows how the graph of the cumulative normal distribution $F_N(x)$ (as graphed in Figure 8.1) becomes a straight line on probability graph paper.

Notice that the 50th percentile of the cumulative normal distribution is μ, the mean. The 84th (more precisely the 84.13th) percentile is $\mu + \sigma$. Thus σ is the difference between the 84th percentile and the 50th percentile. Thus if one were to draw any straight line through the center of a sheet of probability graph paper in the direction from lower left to upper right so that it would intersect the upper and lower edges of the graph paper, one would have a graph of a cumulative normal distribution. If the scale for X has been laid off, one could determine the value of μ graphically by getting the 50th percentile, and the value of σ graphically by taking the difference between the 84th



Graph of Cumulative Normal Distribution $F_N(x)$ on Probability Graph Paper

Figure 8.6



Cumulative Polygon of Figure 2.5 and "Fitted" Cumulative Normal Distribution, Plotted on Probability Graph Paper

Figure 8.7

and 50th percentiles. This procedure is very useful in quickly making rough "estimates" of the mean and standard deviation of a sample distribution of measurements, when the distribution is "approximately" normal.

As an example, suppose we return to Figure 2.5 and draw "by eye" a straight line that seems to "fit" the polygon satisfactorily. A good way to "fit" such a line is to adjust a stretched black thread by trial and error until the fit looks "reasonable", and then mark off a point through which the string passes near each end of the cumulative polygon. Then draw a straight line through the two points. Performing this operation, we get Figure 8.7. Taking the 50th and 84th percentiles with respect to the straight line, we find 1.53 and 1.63. These are the rough graphical estimates of $\overline{X}$ and $\overline{X} + s_x$, respectively. Thus the probability graph paper estimates of $\overline{X}$ and s_x are 1.53 and .1, respectively, which are to be compared with the arithmetically determined values 1.527 and .101, respectively. There are ways of "fitting" straight lines which are mathematically more "precise" than the one just described. One of these, the method of "least squares", will be discussed in Chapter 13.

Exercise 8.

1. If X is a chance quantity having a normal distribution with $\mu = 13$ and $\sigma = 4$, find the value of the probability

   (a) that $X \leq 20$.
   (b) that $10 < X$.
   (c) that $10 < X \leq 20$.
   (d) that X differs from 13 by more than 6.

2. The scores made by candidates on the Scholastic Aptitude Test of the College Entrance Examination Board are normally distributed with mean 500 and standard deviation 100. What percent of the candidates receive scores

   (a) exceeding 700?
   (b) less than 400?
   (c) between 400 and 600?
   (d) which differ from 500 by more than 150 points?
   (e) If a candidate gets a score of 680, what percent of the candidates have higher scores than he?

3. Fit a cumulative normal distribution to the data in problem No. 6 of Exercise 2.2. Graph the cumulative normal distribution you obtain, and also the cumulative polygon (a) on ordinary graph paper and (b) on probability graph paper.

4. Fit a normal distribution to the data in problem No. 7 of Exercise 2.2. Graph the cumulative normal distribution you obtain, and also the cumulative polygon (a) on ordinary graph paper and (b) on probability graph paper.

5. Fit a cumulative normal distribution to the data in problem No. 4 of Exercise 3.3. Graph the cumulative normal distribution you obtain as well as the cumulative polygon (a) on ordinary graph paper and (b) on probability graph paper.

6. You are supposed to have done at least one of the following problems in Exercise 2.2: Nos. 8, 9, 10, 11, 12, 13, 14. Fit a normal distribution to the distribution (or distributions) of sample data on measurements you obtained. Graph the fitted cumulative normal distribution and the cumulative polygon (a) on ordinary graph paper, and (b) on probability graph paper.

7. A sack of 400 nickels is emptied onto a table. Using a cumulative normal distribution approximate the probability that:

(a) more than 250 heads will turn up.
(b) the number of heads will be less than 190.
(c) the number of heads will lie between 170 and 230 inclusive.

8. A die is rolled 720 times. Using an approximating cumulative normal distribution, estimate the probability that:

(a) more than 130 "sixes" will turn up.
(b) the number of "sixes" obtained will lie between 100 and 140 inclusive.

9. It is known that the probability of dealing a bridge hand with at least one ace is approximately .7. If a person plays 100 hands of bridge, what is the approximate probability that he will receive at most 20 hands which will contain no aces?

10. Fit a cumulative normal distribution to the cumulative binomial distribution for the case in which $n = 5$, $p = .4$. Graph both cumulative distributions on ordinary graph paper.

11. Let X be a (discrete) chance quantity denoting the total number of dots obtained in throwing three dice. Work out the probability distribution of X.

Fit a cumulative normal distribution to the cumulative distribution of X. Graph the two cumulative distributions on ordinary graph paper. From each cumulative distribution, calculate the probability that the total number of dots will be (a) at least 12, (b) 8, 9, 10, 11 or 12.

12. Fit the cumulative probability distribution of the discrete chance quantity X in problem No. 6 of Exercise 5.1 by a cumulative normal distribution, and graph the two cumulative distributions.

13. Fit a normal distribution to the cumulative probability distribution of the continuous chance quantity X in problem No. 3 of Exercise 5.3.

CHAPTER 9. ELEMENTS OF SAMPLING

9.1 Introductory Remarks.

In Chapter 4 we mentioned that there are two approaches to the study of sample-to-sample fluctuations of sample statistics: experimental and mathematical. The mathematical approach is founded on the theory of probability and we shall find that the normal distribution plays a very important role in it.

There are two types of sampling which we shall consider: (1) sampling from a finite population, and (2) sampling from an indefinitely large population.

Each of these two types of sampling has an experimental and a mathematical or theoretical aspect. We shall discuss the two types of sampling in turn.

9.2 Sampling from a Finite Population.

First let us consider a very simple example and illustrate what we mean by experimental sampling and what we mean by mathematical or theoretical sampling from a finite population.

9.21 Experimental sampling from a finite population.

Suppose we have 6 chips, marked with the numbers 1, 2, 3, 4, 5, 6, respectively, and placed in a bowl. The composition of this bowl of chips may be described by Table 9.1. If we stir these chips thoroughly and draw out 3 chips simultaneously (or one after another without replacement until 3 chips are drawn) we are experimentally drawing a sample of 3 chips from the finite population of 6 chips having the following distribution of values of X:

TABLE 9.1

Composition of Bowl of Six Chips

x	frequency	relative frequency (probability)
1	1	1/6
2	1	1/6
3	1	1/6
4	1	1/6
5	1	1/6
6	1	1/6
Total	6	1

If we put the 3 chips back in the bowl and stir them with the others, we can repeat the process again and again as many times as we please. Every time we repeat the process we get an experimental sample of 3 chips. Now suppose we are interested in the mean $\bar{X}$ of the numbers on the three chips in a sample. In drawing 100 experimental samples, let us say, from the bowl we would get 100 means, ranging between 2 and 5. In an actual experiment of drawing 100 small samples, the distribution of values of sample means $\bar{X}$ obtained is given in Table 9.2. A column showing the sample sum $S(X)$ is also given. The relation between $\bar{X}$ and $S(X)$ in this example is $3\bar{X} = S(X)$.

TABLE 9.2

Frequency Distribution of $\bar{X}$ (and $S(X)$) in 100 Experimentally Drawn Samples of 3 Chips from the Bowl with Composition Given in Table 9.1

$S(X)$	$\bar{X}$	Frequency	Relative Frequency
6	2.00	6	.06
7	2.33	6	.06
8	2.67	8	.08
9	3.00	17	.17
10	3.33	18	.18
11	3.67	12	.12
12	4.00	14	.14
13	4.33	8	.08
14	4.67	5	.05
15	5.00	6	.06
Total		100	1.00

The second and fourth columns of Table 9.2 constitute the experimental sampling distribution of means $\bar{X}$ in the set of 100 experimentally drawn samples of three chips from the given population of six chips. (Similarly the first and fourth columns constitute the experimental sampling distribution of sample sums $S(X)$.) We could consider other statistics than the mean $\bar{X}$ or sum $S(X)$ calculated from the successive samples, e.g., the standard deviation, the range, largest value, median, smallest value, etc. These statistics all have experimental sampling distributions among the 100 samples too. But the mean is an especially simple statistic to deal with and we will stick to it in most of our discussion of sampling.

If another 100 samples were drawn one would get a slightly different

frequency distribution or relative frequency distribution of values of $\bar{X}$ (or $S(X)$). If a larger number of samples were drawn, say 1000, one would find that the relative frequencies would be more "stable" from one set of 1000 samples to another, than from one set of 100 samples to another. But the distribution of $\bar{X}$ for 100 samples as shown in Table 9.2 gives a good idea of how means of samples of 3 from the given population of 6 chips vary. One can describe the distribution in Table 9.2 graphically by the methods of Chapter 2, or arithmetically by finding the mean $\bar{\bar{X}}$ and standard deviation $s_{\bar{X}}$ of the sample means. In fact, we have

$$\bar{\bar{X}} = 3.46 \tag{9.1}$$

$$s_{\bar{X}} = .79 .$$

We give these values because we shall want to compare them with theoretical values to be worked out in the following paragraphs.

The idea of experimental sampling from a finite population, as illustrated by the foregoing simple example, can clearly be extended to drawing samples of n "elements" from a finite population of N "elements".

Let us now turn to theoretical considerations and show how we can make predictions as to what will actually happen in experimental sampling from a finite population.

9.22 Theoretical sampling from a finite population.

Let us return to the example of a finite population of six chips marked 1, 2, 3, 4, 5, 6, respectively. The number of samples of three chips it is possible to draw out of this population of six chips is simply the number of combinations of 6 objects taken 3 at a time, i.e., $C_3^6 = 20$. These 20 samples are as follows:

1, 2, 3	1, 3, 5	1, 4, 6	3, 4, 5
1, 2, 4	2, 3, 4	2, 3, 6	2, 5, 6
1, 2, 5	1, 3, 6	2, 4, 5	3, 4, 6
1, 3, 4	1, 4, 5	1, 5, 6	3, 5, 6
1, 2, 6	2, 3, 5	2, 4, 6	4, 5, 6 .

The frequency distribution of the sum $S(X)$ and mean $\bar{X}$ in this set of possible samples is given in Table 9.3. The second and fourth columns in Table 9.3 constitute the theoretical sampling distribution of the mean $\bar{X}$ of samples of three chips from the finite population of six chips.

TABLE 9.3

Frequency Distribution of $\overline{X}$ and S(X) Among the 20 Possible Samples of 3 Chips from the Bowl with Composition Given in Table 9.1

S(X)	$\overline{X}$	Frequency	Probability
6	2.00	1	.05
7	2.33	1	.05
8	2.67	2	.10
9	3.00	3	.15
10	3.33	3	.15
11	3.67	3	.15
12	4.00	3	.15
13	4.33	2	.10
14	4.67	1	.05
15	5.00	1	.05
Total		20	1

(Similarly the first and fourth columns constitute the theoretical sampling distribution of the sum S(X).)

The theoretical sampling distribution in Table 9.3 is really nothing but a probability distribution. It is a prediction of the distribution one would get in experimental sampling by drawing a larger and larger number of samples of 3. The accuracy of the prediction for the 100 experimental samples we actually drew may be seen by comparing column 4 of Table 9.2 with column 4 of Table 9.3. As we pointed out before, if we had drawn another 100 samples we would, in general, get a slightly different relative frequency distribution, but not "drastically different" from the theoretical distribution in column 4 of Table 9.3. If we had drawn 1000 samples we would "almost certainly" get a relative frequency distribution closer to the theoretical distribution in column 4 of Table 9.3, assuming "thorough" stirring after each sample.

Now we can get the mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ of the theoretical sampling distribution of $\overline{X}$ in Table 9.3 by the usual formulas (5.1) and (5.3) for the mean and standard deviation of a probability distribution. We have

(9.2)
$$\mu_{\overline{X}} = 3.5$$
$$\sigma_{\overline{X}} = \sqrt{\frac{7}{12}} = .7638 .$$

These theoretical values computed from Table 9.3 are seen to predict closely the experimental values as in (9.1), $\overline{\overline{X}} = 3.46$ and $s_{\overline{X}} = .79$ respectively, computed from Table 9.2.

We can also find from Table 9.3 the values of the mean $\mu_{S(X)}$ and standard deviation $\sigma_{S(X)}$ of the sample sums $S(X)$ in the 20 possible samples from the population. They are

$$\mu_{S(X)} = 10.5 \tag{9.3}$$

$$\sigma_{S(X)} = \sqrt{\frac{63}{12}},$$

which, of course, can also be found directly from $\mu_{\overline{X}}$ and $\sigma_{\overline{X}}$. For since $3\overline{X} = S(X)$, we have $3\mu_{\overline{X}} = \mu_{S(X)}$ and $3\sigma_{\overline{X}} = \sigma_{S(X)}$.

If we should consider samples of one chip from our population of six chips, there would be six possible samples. The theoretical sampling distribution in this case is given by the first and third columns of Table 9.1; it is the same as the distribution of values of X in the population itself. This distribution has its own mean μ and standard deviation σ having values

$$\mu = 3.5 \tag{9.4}$$

$$\sigma = \sqrt{\frac{35}{12}}.$$

In other words, (9.4) gives the values of the mean and standard deviation of the population of chips. Expression (9.2) gives the values of the mean and standard deviation of the distribution of the means of all possible samples of three chips from this population. It will be seen that $\mu_{\overline{X}} = \mu$, and $\sigma_{\overline{X}} = \sigma/\sqrt{5}$. (Also $\mu_{S(X)} = 3\mu$ and $\sigma_{S(X)} = 3\sigma/\sqrt{5}$.) But, of course, we have been discussing a very simple case of theoretical sampling from a finite population. Let us consider the general case.

9.23 The mean and standard deviation of means of all possible samples from a finite population.

The real question here is this: What are the relationships between $\mu_{\overline{X}}$ and μ and between $\sigma_{\overline{X}}$ and σ for more general theoretical sampling from finite populations?

To answer this question, suppose we have a finite population of N "elements", each "element" (e.g., a chip) having a number belonging to it, so that there will be some distribution of these numbers. (The numbers do not have to be integers 1, 2, 3, ... N; any "element" can have any number belonging to it, provided all "elements" do not have the same number on them. In such a case we would not really have a distribution of numbers.) Suppose this distribution has mean μ and standard deviation σ. Now consider all possible samples of n "elements" from this population. There are C_n^N such samples. The theoretical sampling distribution of the means $\overline{X}$ of these samples has the following mean and standard deviation respectively:

(9.5)
$$\mu_{\overline{X}} = \mu$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} .$$

The first formula states that the mean of the means of all possible samples of n elements is equal to the population mean. The second formula states that the standard deviation of means of all possible samples of n elements is equal to the population standard deviation times the factor $\frac{1}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

The theoretical sampling distribution of the sample sums S(X) has the following mean and standard deviation respectively:

(9.6)
$$\mu_{S(X)} = n\mu$$

$$\sigma_{S(X)} = \sigma\sqrt{n}\sqrt{\frac{N-n}{N-1}} .$$

This may be seen at once from the fact that $\mu_{S(X)} = n\mu_{\overline{X}}$, and $\sigma^2_{S(X)} = n^2\sigma^2_{\overline{X}}$.

Note that if the number of elements N in the population is very large compared with the number of elements n in the sample then the factor $\sqrt{\frac{N-n}{N-1}}$ is nearly 1, and the value of $\sigma_{\overline{X}}$ reduces to $\frac{\sigma}{\sqrt{n}}$ approximately, and $\sigma_{S(X)}$ reduces to $\sigma\sqrt{n}$ approximately.

Before we discuss the derivation of formulas (9.5), let us consider the following

Example: Suppose we remove all face cards from a deck of playing cards. Of the 40 remaining cards, 4 will be marked 1, 4 will be marked 2, ..., 4 will be marked 10. What is the mean and standard deviation of the theoretical sampling distribution of means of all possible samples of 10 cards from this "population" of 40 cards?

If X denotes the number on a card, the distribution of X in the population of 40 cards is shown in Table 9.4.

TABLE 9.4

Distribution of the Numbers on the 40 Non-Face Cards in a Deck of Playing Cards

x	Frequency f(x)	Relative Frequency (Probability)
1	4	.1
2	4	.1
3	4	.1
4	4	.1
5	4	.1
6	4	.1
7	4	.1
8	4	.1
9	4	.1
10	4	.1
Total	40	1.0

The mean and standard deviation of this distribution are

$$\mu = 5.5$$

$$\sigma = \sqrt{8.5}.$$

The values of N and n are 40 and 10 respectively. Therefore, we have for the mean and standard deviation of the distribution of sample means

$$\mu_{\bar{X}} = 5.5$$

$$\sigma_{\bar{X}} = \frac{\sqrt{8.5}}{\sqrt{10}}\sqrt{\frac{40-10}{40-1}} = \sqrt{\frac{8.5}{13}},$$

respectively. For the mean and standard deviation of sample sums we would have

$$\mu_{S(X)} = 55$$

$$\sigma_{S(X)} = 10\sqrt{\frac{8.5}{13}}.$$

The derivation of the formula $\mu_{\bar{X}} = \mu$ in (9.5) proceeds as follows: Let $x_1, x_2, \ldots, x_N$ be the numbers belonging to the N elements in the population. We may think of $(x_1, x_2, \ldots, x_{n-1}, x_n)$, $(x_1, x_2, \ldots, x_{n-1}, x_{n+1})$, ..., and finally $(x_{N-n+1}, x_{N-n+2}, \ldots, x_N)$ as the C_n^N samples. Now $\mu_{\bar{X}}$ is the mathematical expectation of the mean of the means of all these samples. Since the probability for each sample is $\frac{1}{C_n^N}$, this means that $\mu_{\bar{X}}$ is the mean of all sample means. Hence, $C_n^N \mu_{\bar{X}}$ is the sum of all sample means, i.e.,

$$C_n^N \mu_{\bar{X}} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) + \frac{1}{n}(x_1 + x_2 + \ldots + x_{n-1} + x_{n+1}) \tag{9.7}$$

$$+ \ldots + \frac{1}{n}(x_{N-n+1} + x_{N-n+2} + \ldots + x_N).$$

Now each one of the x's in the population will occur in C_{n-1}^{N-1} of the samples. For if a particular x is selected, a sample of n elements containing that x can be formed by selecting n-1 other x's. There are C_{n-1}^{N-1} ways of selecting the other n-1 x's. Hence we can rewrite (9.7) as follows by collecting the x_1's, the x_2's, etc.

$$C_n^N \mu_{\bar{X}} = \frac{1}{n} C_{n-1}^{N-1} x_1 + \frac{1}{n} C_{n-1}^{N-1} x_2 + \ldots + \frac{1}{n} C_{n-1}^{N-1} x_N. \tag{9.8}$$

Now if we divide both sides of (9.8) by C_n^N and note that $C_{n-1}^{N-1} / C_n^N = \frac{n}{N}$, we have

$$\mu_{\bar{X}} = \frac{1}{N}(x_1 + x_2 + \ldots + x_N). \tag{9.9}$$

But $\frac{1}{N}(x_1 + x_2 + \ldots + x_N)$ is the mean of the x's in the population and we have denoted this by μ. Therefore

$$\mu_{\bar{X}} = \mu \tag{9.10}$$

which establishes the first formula of (9.5).

The derivation of the variance $\sigma_{\bar{X}}^2$ is a little more involved, but will be given for the benefit of those who wish to follow it through. To obtain the value of $\sigma_{\bar{X}}^2$ we use the definition of the variance of a probability distribution given in formula (5.3b). This means that $\sigma_{\bar{X}}^2$ is equal to the mean of the squares of means of all possible samples minus the square of $\mu_{\bar{X}}$ (the mean of means of all samples), or written more briefly

$$\sigma_{\bar{X}}^2 = \left[\frac{1}{n}(x_1+x_2+\ldots+x_n)\right]^2 \cdot \frac{1}{C_n^N} + \left[\frac{1}{n}(x_1+x_2+\ldots+x_{n-1}+x_{n+1}\right]^2 \cdot \frac{1}{C_n^N} \tag{9.11}$$

$$+\ldots+\left[\frac{1}{n}(x_{N-n+1}+x_{N-n+2}+\ldots+x_N)\right]^2 \cdot \frac{1}{C_n^N} - \mu_{\bar{X}}^2 .$$

When the quantity in each square bracket is squared, we find the sum of squares of all x's in that bracket plus the sum of twice the product of all possible pairs of x's in that square bracket. But we know, by the same argument given in arriving at formula (9.10), that the square of any given x, say x_1, occurs in a total of C_{n-1}^{N-1} samples. By the same method of reasoning any particular product of x's, e.g., x_1x_2, occurs in C_{n-2}^{N-2} samples. Hence, by squaring out the terms in (9.11) and collecting, we have

$$\sigma_{\bar{X}}^2 = \frac{1}{n^2}\frac{1}{C_n^N} \cdot C_{n-1}^{N-1}\left[x_1^2 + x_2^2 + \ldots + x_N^2\right] \tag{9.12}$$

$$+ \frac{2}{n^2}\frac{1}{C_n^N} \cdot C_{n-2}^{N-2}\left[x_1x_2+x_1x_3+\ldots+x_{N-1}x_N\right] - \mu_{\bar{X}}^2 .$$

But from (9.9) we have

$$\mu_{\bar{X}}^2 = \frac{1}{N^2}(x_1 + x_2 + \ldots + x_N)^2 \tag{9.13}$$

or, squaring the quantity in the parentheses,

$$\mu_{\bar{X}}^2 = \frac{1}{N^2}\left[x_1^2 + x_2^2 + \ldots + x_N^2\right] + \frac{2}{N^2}\left[x_1x_2 + x_1x_3 + \ldots + x_{N-1}x_N\right].$$

Inserting in (9.12) the value of $\mu_{\bar{X}}^2$ as given by (9.13), and noticing that $\frac{1}{n^2}\cdot\frac{1}{C_n^N}\cdot C_{n-1}^{N-1} = \frac{1}{nN}$ and

$$\frac{2}{n^2}\cdot\frac{1}{C_n^N}\cdot C_{n-2}^{N-2} = \frac{2(n-1)}{nN(N-1)},$$

we may write

$$\sigma_{\bar{X}}^2 = \left(\frac{1}{nN} - \frac{1}{N^2}\right)\left[x_1^2 + x_2^2 + \ldots + x_N^2\right] + 2\left(\frac{n-1}{nN(N-1)} - \frac{1}{N^2}\right)\left[x_1x_2 + x_1x_3 + \ldots + x_{N-1}x_N\right]$$

$$= \left\{\left[\frac{1}{N} - \frac{1}{N^2}\right]\left[x_1^2 + x_2^2 + \ldots + x_N^2\right] - \frac{2}{N^2}\left[x_1x_2 + x_1x_3 + \ldots + x_{N-1}x_N\right]\right\}\frac{(N-n)}{n(N-1)}. \tag{9.14}$$

The quantity in $\{\}$ may be written as

$$\frac{1}{N}(x_1^2 + x_2^2 + \ldots + x_N^2) - \left[\frac{x_1 + x_2 + \ldots + x_N}{N}\right]^2 \tag{9.15}$$

which is simply σ^2, the variance of a finite population. Notice that <u>in defining the variance of a finite population, we divide by N and not N-1</u>. (In the case of the variance of a sample of n cases, from a population, remember that we divide by n-1 and not n.) We finally have, for the variance of the sample means,

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \tag{9.16}$$

from which the value of $\sigma_{\bar{X}}$ in (9.5) may be obtained by taking square roots.

9.24 Approximation of distribution of sample means by normal distribution.

You will recall that there are C_n^N possible samples of n elements in a population of N elements. Starting with a fairly large population and considering samples even of moderate size, it would be a hopeless task to compute a table similar to Table 9.3 showing the probabilities in the distribution of sample means. But we can calculate these probabilities approximately. Under a very wide set of conditions this distribution can be approximately fitted by a normal distribution. Roughly speaking these conditions are that n be fairly large (say 30 or greater) and N much larger (say 100 or greater). But if one were to examine the conditions more precisely one would find, in certain special cases, that the fit is very good for n as small as 10 and N as small as 25. For example, this would be true if the values of x in the population were equally likely and equally spaced. On the other hand, the fit can be very bad for such small values of n and N, particularly if the distribution of values of x's in the population is very lopsided or skewed. A full mathematical statement of the degree of accuracy with which a cumulative theoretical sampling distribution of sample means can be approximated by a cumulative normal distribution is far beyond the scope of this course.

Now how is a normal distribution fitted to a distribution of sample means? Just as in Chapter 8, we would fit the cumulative normal distribution to the cumulative distribution of sample means. To carry out this fitting process, we would need only the values of $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ and Table 8.1, and proceed as we did in fitting the cumulative normal distribution to a cumulative polygon or to a cumulative binomial distribution.

The fitted distribution of $\bar{X}$ has the equation

$$F_N(\bar{x}) = \frac{1}{\sqrt{2\pi}\sigma_{\bar{X}}} \int_{-\infty}^{\bar{x}} e^{\frac{1}{-2\sigma_{\bar{X}}^2}(x - \mu_{\bar{X}})^2} dx , \tag{9.17}$$

where the values of $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ are given by formulas (9.5). The table of fitted probabilities would be given by Table 8.1 by replacing μ by $\mu_{\bar{X}}$ and σ by $\sigma_{\bar{X}}$ and X by $\bar{X}$. The graph of the fitted cumulative normal distribution would look like Figure 8.1, where μ is replaced by $\mu_{\bar{X}}$ and σ by $\sigma_{\bar{X}}$. In any particular example we would have numerical values of the population size

N, the sample size n, the population mean μ, and the standard deviation σ. Actually we would rarely carry out the actual process of fitting. We only use the fitted cumulative normal distribution to approximate any particular probability in which we may be interested. In determining such a probability, it is not necessary to actually fit the complete distribution. An example will make this clear.

Example: The distribution of weights in a population of 1000 students has mean 148.2 lb., and standard deviation 5.4 lb. If a sample of 100 students is picked "at random" from this population, what is the probability that the mean weight of these 100 students will exceed 149 lbs.?

We have N = 1000, n = 100, μ = 148.2, and σ = 5.4.

Hence

$$\mu_{\overline{X}} = 148.2$$

$$\sigma_{\overline{X}} = \frac{5.4}{\sqrt{100}}\sqrt{\frac{900}{999}} = .513 .$$

The relationship (see Section 8.1) between Z and $\overline{X}$ for this problem is

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} .$$

Using the values of $\mu_{\overline{X}}$ and $\sigma_{\overline{X}}$ found, we have

$$Z = \frac{\overline{X} - 148.2}{.513} .$$

Now our original question was this: What is the probability that $\overline{X}$ will exceed 149? $\overline{X}$ will exceed 149 if and only if Z exceeds $\frac{149 - 148.2}{.513}$ (= 1.56). We may write this briefly as follows:

$$\Pr(\overline{X} > 149) = \Pr\left(\frac{X-148.2}{.513} > \frac{149-148.2}{.513}\right) = \Pr(Z > 1.56) .$$

But $\Pr(Z > 1.56) = 1 - \Pr(Z < 1.56)$. Looking in Table 8.1 and interpolating for z = 1.56, we find $\Pr(Z < 1.56) = .9406$. Hence the answer to our question is

$$\Pr(\overline{X} > 149) = 1 - .9406 = .059 .$$

This means that approximately 5.9% of the possible samples of 100 from the given

population of 1000 students have means exceeding 149 lbs.

In a similar manner one can find approximately the probabilities in the distribution of the sums S(X) in all possible samples. In this case the relation between Z and S(X) is

$$Z = \frac{S(X) - \mu_{S(X)}}{\sigma_{S(X)}},$$

where $\mu_{S(X)}$ and $\sigma_{S(X)}$ are given in terms of N, n, μ and σ by formulas (9.6).

Asking probability questions in terms of $\overline{X}$ is equivalent to asking probability questions in terms of S(X). One can express a probability dependent on $\overline{X}$ as a probability dependent on S(X). For instance in the example just given, $\mu_{S(X)}$ would have the value 14820 lbs., and $\sigma_{S(X)}$ the value $(5.4)\sqrt{100}\sqrt{\frac{900}{999}} = 51.3$ pounds, and the relation between Z and S(X) would be

$$Z = \frac{S(X) - 14820}{51.3}.$$

From this relation one could ask: What is the approximate probability that the total weight of 100 students would exceed 14,900 lbs.? The value of Z for S(X) = 14,900 is $\frac{14900 - 14820}{51.3} = \frac{80}{51.3} = 1.56$

$= \Pr(\overline{X} > 149)$.

Thus

$$\Pr(S(X) > 14{,}900) = \Pr(Z > 1.56) = .059 .$$

In other words, the probability of the mean weight of 100 students exceeding 149 lbs., is exactly the same as the probability of the total weight of the 100 students exceeding 14900 lbs.

Exercise 9.2.

1. In Problem No. 1 of Exercise 2.1, suppose a sample of 2 students is picked at random. What is the probability that their average score would exceed 17? That the sum of their scores would lie between 13 and 75 (excluding 13 and 75)?

2. Suppose a population of 50,000 candidates takes the Scholastic Aptitude Test

and that the scores are scaled in such a way that the mean of the scores in the population is 500 and the standard deviation is 100. If 25 candidates are picked at random from this population, what is the probability that their average score will be less than 475? Greater than 550? Between 475 and 525?

3. If 45 percent of the 3600 Princeton undergraduates were the number who would answer "yes" to a certain question, what is the approximate probability that in a random sample of 100 students a majority would answer "yes" to the question? (Hint: Each individual in the population can be considered as being "marked" 1 or 0: 1 if he says yes, and 0 otherwise. Thus in drawing a person, X is a chance quantity which is equal to 1 if the person says "yes" and 0 otherwise. The distribution of X is therefore as follows:

x	frequency	p(x)
0	1980	.55
1	1620	.45
Total	3600	1.00

From this distribution you can easily find the value of μ and σ.)

4. Suppose a lot containing 10,000 articles contains 20 percent defectives. What is the approximate probability that a random sample of 400 articles from this lot will contain more than 25% defectives?

5. Suppose 1000 chips, marked 1, 2, 3, ..., 1000 respectively, are put in a bowl and mixed thoroughly. If 10 chips are drawn at random, approximately what is the probability that the sum of the numbers will exceed 6000? (For your information:

$$1 + 2 + \ldots + k = \frac{k(k+1)}{2}$$

$$1^2 + 2^2 + \ldots + k^2 = \frac{k(k+1)(2k+1)}{6}$$.)

6. Consider a fictitious life insurance company which has an insurance policy on each of 100,000 persons, aged 65, and suppose that the distribution of size (x) of policies (in \$1000's) is as follows:

x	1	2	3	4	5	6	7	8	9	10	15	20	25
relative frequency	.25	.03	.02	.02	.40	.02	.02	.01	.01	.15	.05	.01	.01

Suppose 1000 of these persons will die within a year. Approximately, what is the probability that the company will have to pay more than $6,000,000 in death claims to beneficiaries of these persons?

9.3 Sampling from an Indefinitely Large Population.

Just as in the case of sampling from a finite population, there are two approaches to the study of sample-to-sample fluctuations of sample statistics, in samples from an indefinitely large population: experimental and mathematical. Experimental sampling from an indefinitely large population is not essentially different from sampling from a finite population. It consists of actually carrying out experimental operations. The experiments may be less obviously sampling operations. For example, if a single die is thrown 10 times successively we may regard this as drawing a sample of 10 elements from an indefinitely large population of potential throws of the die. This may be repeated again and again as often as we like. If we stop with 100 samples, say, we could make a frequency distribution of $\bar{X}$ or $S(X)$ analogous to Table 9.2. We shall not actually do this, however, but shall base most of our discussion on theoretical sampling from an indefinitely large population.

Conceptually, we may think of "drawing" such a sample as equivalent to drawing 10 chips out of a bowl containing indefinitely many chips, where each chip is marked 1, 2, 3, 4, 5 or 6. If the bowl of chips is to correspond to a "true" die, the proportion of chips with each number marked on would be 1/6, and the mixing would be "thorough". Or, alternatively, we can think of putting six chips marked 1, 2, 3, 4, 5, 6 in the bowl, mixing thoroughly, and "drawing" one chip, returning it to the bowl, and repeating this process 10 times. Thus 10 "drawings" of a single chip with replacement after each drawing is equivalent to "drawing" a sample of 10 chips from an indefinitely large population.

9.31 Mean and standard deviation of theoretical distributions of means and sums of samples from an indefinitely large population.

Actually, it is theoretical sampling rather than experimental sampling from an indefinitely large population which is of greatest interest here. For we can establish some fairly simple sampling theory of sample means and sample sums in this case which can usefully predict what will happen in actual experimental sampling from an indefinitely large population. We have discussed the case of a finite population first because it is simpler conceptually and because we can make use of some results obtained in that case for the case of the indefinitely large population. The main results for the case of the finite population are the formulas given by (9.5) and (9.6) and the fact that the distribution of sample means for large n and much larger N is approximately a normal distribution. It is clear from formulas (9.5) that the mean and standard deviation of the distribution of means of samples of n elements from an indefinitely large population (i.e., N indefinitely large) are

(9.18)
$$\mu_{\overline{X}} = \mu$$
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

where μ and σ are the mean and standard deviation of the indefinitely large population from which the samples are drawn. Similarly, when N is indefinitely large, formulas (9.6) become

(9.19)
$$\mu_{S(X)} = n\mu$$
$$\sigma_{S(X)} = \sigma\sqrt{n}.$$

You will recall from the discussion in Section 5.35 that the distribution for an indefinitely large population is simply a probability distribution. For example, the probability distribution

x	1	2	3	4	5	6
f(x)	1/6	1/6	1/6	1/6	1/6	1/6

may be regarded as the distribution of the number of dots obtained at a throw in an indefinitely large population of throws of a single "true" die. The cumulative probability distribution F(x) = x which is graphed in Figure 5.5 may be regarded as the cumulative distribution of values of pointer readings in an

indefinitely large population of spins of the "perfectly balanced" pointer described in Section 5.21.

In drawing the successive elements of a sample from a finite population, the probabilities change as more elements are withdrawn from the population. This is why we consider sampling from a finite population as a problem in combinations. But in drawing the successive elements of a sample from an indefinitely large population, the probability that a chance quantity X has a particular value in one drawing is not affected by the value of X in any other drawing. In other words, the results of successive drawings from an indefinitely large population are <u>independent</u> of each other.

Although the formulas (9.18) and (9.19) were obtained from (9.5) and (9.6) merely by letting N become indefinitely large, one can actually derive them from scratch by another process. This process is important in the mathematical theory of sampling and may be illustrated by considering samples of 2 elements from an indefinitely large population having the following general distribution of a discrete chance quantity X for the population.

x	x_1	x_2	x_3	...	x_k
f(x)	$f(x_1)$	$f(x_2)$	$f(x_3)$	...	$f(x_k)$

The probability that the chance quantity X has the value x_α in the first drawing is $f(x_\alpha)$, and that it has the value x_β in the second drawing is $f(x_\beta)$. The drawings are independent, since we are sampling from an indefinitely large population, and hence the probability that X has the values x_α and x_β in the two successive drawings is, by the law of multiplication of probabilities, $f(x_\alpha)\cdot f(x_\beta)$. The mean of this sample of 2 elements is $(x_\alpha + x_\beta)/2$. The mean value $\mu_{\overline{X}}$ of the sample mean is obtained by multiplying $(x_\alpha + x_\beta)/2$ by the probability $f(x_\alpha) \cdot f(x_\beta)$ and summing over all possible values of α and β, i.e.,

$$\mu_{\overline{X}} = \sum_{\alpha=1}^{k} \sum_{\beta=1}^{k} \left[(x_\alpha + x_\beta)/2\right] \cdot f(x_\alpha) \cdot f(x_\beta) , \tag{9.20}$$

which may be broken into two sums as follows:

$$\mu_{\overline{X}} = \frac{1}{2} \sum_{\alpha=1}^{k} \sum_{\beta=1}^{k} x_\alpha \cdot f(x_\alpha) \cdot f(x_\beta) + \frac{1}{2} \sum_{\alpha=1}^{k} \sum_{\beta=1}^{k} x_\beta \cdot f(x_\beta) \cdot f(x_\alpha) \tag{9.21}$$

$$= \frac{1}{2}\left[\sum_{\alpha=1}^{k} x_\alpha \cdot f(x_\alpha)\right]\left[\sum_{\beta=1}^{k} f(x_\beta)\right] + \frac{1}{2}\left[\sum_{\beta=1}^{k} x_\beta \cdot f(x_\beta)\right]\left[\sum_{\alpha=1}^{k} f(x_\alpha)\right] .$$

But

$$\sum_{\alpha=1}^{k} f(x_\alpha) = \sum_{\beta=1}^{k} f(x_\beta) = 1 ,$$

and

$$\sum_{\alpha=1}^{k} x_\alpha \cdot f(x_\alpha) = \sum_{\beta=1}^{k} x_\beta \cdot f(x_\beta) = \sum_{i=1}^{k} x_i \cdot f(x_i) = \mu ,$$

(it does not matter what letter we use for the subscripts), where μ is the population mean. Hence

$$\mu_{\overline{X}} = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu ,$$

or

$$\mu_{\overline{X}} = \mu .$$

By following the same line of reasoning, we find that for samples of size n

$$\mu_{\overline{X}} = \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu \qquad \text{(n terms)} \tag{9.22}$$

or

$$\mu_{\overline{X}} = \mu .$$

In other words, the mean of the distribution of means of an indefinitely large number of samples from an indefinitely large population is equal to the mean of the population.

Similarly, by applying the principle expressed by (5.3b), the variance of $\overline{X}$ for samples of 2 elements is given by

$$\sigma_{\overline{X}}^2 = \sum_{\beta=1}^{k}\sum_{\alpha=1}^{k} \left[(x_\alpha + x_\beta)/2\right]^2 \cdot f(x_\alpha) \cdot f(x_\beta) - \mu_{\overline{X}}^2 . \tag{9.23}$$

Squaring the quantity in $[\ ]$ and summing the resulting three terms just as we did in passing from (9.20) to (9.21), we find that (9.23) reduces to

$$(9.24)\quad \sigma_{\bar{X}}^2 = \frac{1}{4}\sum_{\alpha=1}^{k} x_\alpha^2 \cdot f(x_\alpha) + \frac{2}{4}\left[\sum_{\alpha=1}^{k} x_\alpha \cdot f(x_\alpha)\right]\left[\sum_{\beta=1}^{k} x_\beta \cdot f(x_\beta)\right] + \frac{1}{4}\sum_{\beta=1}^{k} x_\beta^2 f(x_\beta) - \mu_{\bar{X}}^2.$$

But we know from (5.3b) that

$$\sum_{\alpha=1}^{k} x_\alpha^2 \cdot f(x_\alpha) = \sum_{\beta=1}^{k} x_\beta^2 \cdot f(x_\beta) = \sum_{i=1}^{k} x_i^2 \cdot f(x_i) = \sigma^2 + \mu^2.$$

Hence

$$(9.25)\quad \sigma_{\bar{X}}^2 = \frac{\sigma^2 + \mu^2}{4} + \frac{2\mu^2}{4} + \frac{\sigma^2 + \mu^2}{4} - \mu_{\bar{X}}^2.$$

But we know that $\mu_{\bar{X}}^2 = \mu^2$. Hence (9.25) reduces to

$$(9.26)\quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{2},$$

for samples of 2 elements.

By similar reasoning, we would find for samples of n elements that

$$(9.27)\quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n},$$

or

$$(9.28)\quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

This states that the standard deviation of the distribution of means of samples of n elements from an indefinitely large population is equal to the standard deviation of the population divided by $\sqrt{n}$.

It should be emphasized that formulas (9.22) and (9.28) are essentially the mean and standard deviation respectively of a probability distribution.

Formulas (9.22) and (9.28) were actually derived for the case of sampling from a population with a discrete chance quantity X. But the same formulas hold for sampling from a population having a continuous probability distribution of a chance quantity X. In such a case, the population would have a probability density function $f(x)$ and μ and σ would be obtained by integration using formulas (5.14) and (5.15a).

It should be pointed out that the formulas for the mean and standard deviation of the distribution of the sample sum S(X) in indefinitely many samples from an indefinitely large population are

(9.29)
$$\mu_{S(X)} = n\mu$$
$$\sigma^2_{S(X)} = n\sigma^2.$$

The validity of these formulas can be seen by letting N increase indefinitely in formulas (9.6).

9.32 Approximate normality of distribution of sample mean in large samples from an indefinitely large population.

As in the case of theoretical sampling from a finite population, the distribution of means of samples from an indefinitely large population can be approximated by a normal distribution under certain conditions. In fact, we may make the following statement:

For large values of n the distribution of means of samples of n elements from an indefinitely large population with mean μ and standard deviation σ is approximately normal with mean $\mu_{\overline{X}}$ $(=\mu)$ and standard deviation $\sigma_{\overline{X}}\left(=\frac{\sigma}{\sqrt{n}}\right)$, the accuracy of the approximation becoming perfect in the limit as n increases indefinitely.

The accuracy of approximation depends on the size of n, the degree of lopsidedness or skewness of the population distribution, and other things. But in some situations, the approximation is good for practical purposes for values of n as small as 10. If the population has a normal distribution with mean μ and standard deviation σ, then $\overline{X}$ is exactly normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

The procedure by which the normal distribution is used to approximate probabilities concerning $\overline{X}$ or S(X) is very similar to that discussed in Section 9.24. An example will make this clear.

Example: Suppose a "true" die is rolled 25 times. Approximately what is the probability that the average of the 25 numbers of dots obtained will be less than 4? (This is equivalent to asking: What is the probability that the total number of dots obtained will be less than 100?)

In this problem the population distribution is as follows:

x	1	2	3	4	5	6
f(x)	1/6	1/6	1/6	1/6	1/6	1/6

The population mean $\mu = 3.5$ and $\sigma = \sqrt{\frac{35}{12}}$; $n = 25$, hence

$$\mu_{\overline{X}} = 3.5$$

$$\sigma_{\overline{X}} = \frac{1}{\sqrt{25}} \cdot \sqrt{\frac{35}{12}} = \sqrt{\frac{7}{60}} = .342 \ .$$

The relation between Z and $\overline{X}$ for this problem is

$$Z = \frac{\overline{X} - 3.5}{.342} \ .$$

For $\overline{X} = 4$, we have $Z = \frac{.5}{.342} = 1.46$ and hence

$$Pr(\overline{X} < 4) = Pr(Z < 1.46) \ .$$

Looking in Table 8.1 and interpolating, we find $Pr(Z < 1.46) = .928$ as the answer to the question.

9.33 Remarks on the binomial distribution as a theoretical sampling distribution.

The binomial distribution given by formula (6.3) is essentially a theoretical sampling distribution; it is the sampling distribution of the sample sum S(X) in samples of n elements from an indefinitely large population in which the chance quantity X has the following distribution:

x	0	1
f(x)	q	p

The occurrence of event E corresponds to x = 1, and the occurrence of "not E" corresponds to x = 0. In other words, the chance quantity X takes a value equal to the number of E's in a single trial (either 1 or 0).

The mean μ of this population is seen to be $0 \cdot q + 1 \cdot p = p$, i.e.,

$$\mu = p ,$$

and the variance of the distribution is

$$(0-p)^2 \cdot q + (1-p)^2 \cdot p = p^2 q + q^2 p = pq(p+q) = pq, \text{ i.e.,}$$

$$\sigma^2 = pq .$$

If samples of n elements are drawn from this population, then S(X), the sum of the X's in the sample is simply the total number of 1's in the sample (i.e., the number of times event E occurred in the sample of n elements or trials). Actually, formula (6.3) gives the probability that S(X) has the value x; it follows from formula (9.19) that the mean and variance of S(X) for the case of samples from the binomial population having distribution

x	0	1
f(x)	q	p

are given by

$$\mu_{S(X)} = np$$

and

$$\sigma^2_{S(X)} = npq$$

respectively. These values of the mean and variance of a binomial distribution were established by direct, but more cumbersome argument, in Section 6.2.

If we are interested in the mean $\overline{X}$ (the proportion of trials in which event E occurs), of the binomial distribution, we find that the distribution has the following mean and standard deviation:

(9.30)

$$\mu_{\overline{X}} = p$$

$$\sigma_{\overline{X}} = \sqrt{\frac{pq}{n}} .$$

You should notice that not only do we know the mean and variance of $\overline{X}$ and S(X) in samples of n elements from the binomial population having the distribution mentioned above, but we also know the exact sampling distribution

of $\overline{X}$ and S(X). This distribution is, in fact, the binomial distribution

$$f(x) = C_x^n p^x (1-p)^{n-x} .$$

More precisely this formula gives the probability that S(X) = x, or that $\overline{X} = \frac{x}{n}$.

Exercise 9.3.

1. Suppose the net weight of individual packages in a population of "half-pound" packages has mean .51 lb., and standard deviation .02 lb., and that the packages are put up in lots of 2500 packages. What proportion of the lots can be expected to weigh more than 1276 pounds net? Between 1273 and 1277 pounds net weight?

2. When a one-foot ruler and a pencil are repeatedly used to mark off "one-foot lengths" suppose, in fact, that a population (considered indefinitely large) of lengths is generated with mean 1.001 feet, and standard deviation .005 ft. If a distance of 100 ruler lengths is marked off, what is the probability that the distance marked off will be less than 100 ft.? Will lie between 100 and 100.2 feet?

3. Suppose the population of men travelling in airplanes from a large city has a distribution of gross weights with mean 162 lb., and standard deviation 7 lb. What is the approximate probability that a DC-3 load of 21 men passengers would have a combined gross weight of more than 3500 lbs.?

4. Eight persons sitting around a table are provided with 10 matches each. Each person takes some number of matches from 1 to 10 and clenches them in his hand. At a given instant everyone shows how many matches he is holding. Approximately what is the probability that the number of matches that will turn up will lie between 35 and 45 inclusive?

5. If a hand of 13 cards is dealt from a pack of 52 playing cards, the probability (to 3 decimal places) of getting 0, 1, 2, 3, 4 aces is given by the following probability distribution:

x	0	1	2	3	4
f(x)	.304	.439	.213	.041	.003

In playing 100 hands of bridge, approximately what is the probability of getting a total of less than 90 aces? More than 120 aces?

6. Suppose pamphlets are counted out in packets of 25 by weighing them. Suppose the distribution of weights of individual pamphlets has mean 1 oz., and standard deviation .05 oz. Any weighed pile of pamphlets is "counted" as 25 pamphlets only if it makes the scales read between 24.5 and 25.5 ounces. What is the probability that a pile of 24 pamphlets will be "counted" as 25 pamphlets by this method? That a pile of 25 pamphlets will not be "counted" as 25 pamphlets?

7. If 60% of the voters of a certain city (the population of voters to be considered indefinitely large) are in favor of a given change, what is the probability that a random sample of 100 of the voters would not show a majority in favor of the change?

8. Suppose certain numbers are calculated to four decimal places. If we drop the fourth decimal (which means we drop .000x, where x = 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9) and add 25 resulting 3-decimal numbers, what is the approximate probability that the total of the dropped numbers will exceed .01?

9. Suppose breaking strengths of individual pieces of a certain type of plastic fiber have a distribution with mean 2.5 lb., and standard deviation .2 lb. What is the probability that a bundle of 100 of these fibers would support a weight of 255 lbs.?

9.4 The Theoretical Sampling Distribution of Sums and Differences of Sample Means.

In statistical problems, we frequently have to deal with the difference between the means of two samples, or the sum or average (weighted or unweighted) of two or more sample means. In this section we shall discuss briefly the sampling theory of such sums and differences.

9.41 Differences of sample means.

The main results for differences between sample means which we want to consider may be stated as follows:

Suppose $\bar{X}$ is the mean of a sample of n elements drawn from a population having mean μ and variance σ^2, and $\bar{X}'$ is the mean of a sample of n' elements from

a population with mean μ' and variance σ'^2. Then the mean of the distribution of the difference of means $\bar{X} - \bar{X}'$ is given by

$$\mu_{\bar{X}-\bar{X}'} = \mu_{\bar{X}} - \mu_{\bar{X}'} \tag{9.31}$$

$$= \mu - \mu' .$$

This formula holds for finite or indefinitely large populations. The variance of the distribution of the difference of means is given by

$$\sigma^2_{\bar{X}-\bar{X}'} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{X}'} , \tag{9.32}$$

or

$$\sigma^2_{\bar{X}-\bar{X}'} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} + \frac{\sigma'^2}{n'} \cdot \frac{N'-n'}{N'-1} \tag{9.33}$$

in case of finite populations, where N and N' are the numbers of elements in the two populations respectively, and

$$\sigma^2_{\bar{X}-\bar{X}'} = \frac{\sigma^2}{n} + \frac{\sigma'^2}{n'} \tag{9.34}$$

in case of indefinitely large populations.

As in the case of single sample means, the difference between two sample means is, under certain conditions, approximately normally distributed with mean $(\mu-\mu')$ and variance given by (9.33) in case of finite populations or (9.34) in case of indefinitely large populations. The main conditions under which the approximation is valid are similar to those for approximate normality of the distribution of means of single samples, namely large n and much larger N, and also large n' and much larger N'.

Let us consider an example.

Example: Suppose 80% of the population (considered indefinitely large) of voters in a certain city is in favor of a certain proposal. Two samples of 100 voters each are polled. Approximately what is the probability that the difference between the percentages favoring the proposal will exceed 10%?

Here we are drawing both samples from the same population. The chance quantity X has the value 1 for a voter if he favors the proposal and 0 otherwise. Hence we have from (9.30)

$$\mu_{\overline{X}} = \mu_{\overline{X}'} = .8$$

$$\sigma_{\overline{X}} = \sigma_{\overline{X}'} = \sqrt{\frac{(.8)(.2)}{100}} = .04 .$$

Hence from (9.31)

$$\mu_{\overline{X}-\overline{X}'} = 0$$

and from (9.32)

$$\sigma^2_{\overline{X}-\overline{X}'} = \sigma^2_{\overline{X}} + \sigma^2_{\overline{X}'} = 2(.04)^2$$

or

$$\sigma_{\overline{X}-\overline{X}'} = .04\sqrt{2} = .057 .$$

The question to be answered in the example may be expressed as follows: What is the value of $Pr(|\overline{X}-\overline{X}'| > .1)$? Now $Pr(|\overline{X}-\overline{X}'| > .1) = 1-Pr(-.1 < \overline{X}-\overline{X}' < .1)$. For this problem we have

$$Z = \frac{(\overline{X} - \overline{X}') - \mu_{\overline{X}-\overline{X}'}}{\sigma_{\overline{X}-\overline{X}'}} = \frac{\overline{X} - \overline{X}'}{.057} .$$

The values of Z for $\overline{X}-\overline{X}' = .1$ and $-.1$ are $\frac{.1}{.057} = 1.76$ and $\frac{-.1}{.057} = -1.76$. Hence, we find from Table 8.1

$$Pr(-.1 < \overline{X} - \overline{X}' < .1) = Pr(-1.76 < Z < 1.76) = .9212.$$

The probability is approximately $1 - .9212 = .0788$ that the difference between the percentages of voters favorable to the proposal in the two samples will differ by more than 10%.

9.42 Sums of sample means.

If we take the sum of the sample means, this sum will have a distribution with the following mean and variance:

(9.35)
$$\mu_{\overline{X}+\overline{X}'} = \mu_{\overline{X}} + \mu_{\overline{X}'}$$

(9.36)
$$\sigma^2_{\overline{X}+\overline{X}'} = \sigma^2_{\overline{X}} + \sigma^2_{\overline{X}'} .$$

Note that formula for $\sigma^2_{\overline{X}+\overline{X}'}$ is the same as that for $\sigma^2_{\overline{X}-\overline{X}'}$.

As a matter of fact, if we take any linear combination $a\overline{X} + a'\overline{X}'$

where a and a' are any given constants, this linear combination will have the following mean and variance:

$$\mu_{a\bar{X}+a'\bar{X}'} = a\mu_{\bar{X}} + a'\mu_{\bar{X}'} , \tag{9.37}$$

$$\sigma^2_{a\bar{X}+a'\bar{X}'} = a^2\sigma^2_{\bar{X}} + a'^2\sigma^2_{\bar{X}'} . \tag{9.38}$$

9.43 Derivations.

We can derive the basic formulas (9.31), (9.32), (9.35) and (9.36) by considering two general discrete chance quantities X and X'. These derivations are given below for those who want to see how the argument proceeds. Similarly, derivations can be made by using continuous chance quantities X and X'. In this case the summations involved in the derivation would be replaced by integrations.

Suppose X is a discrete chance quantity having the probability distribution

x	x_1	x_2	...	x_k
$f(x)$	$f(x_1)$	$f(x_2)$	...	$f(x_k)$

and X' is a discrete chance quantity having the probability distribution

x'	x_1'	x_2'	...	$x'_{k'}$
$f'(x')$	$f'(x_1')$	$f'(x_2')$	...	$f'(x'_{k'})$

Let the mean and standard deviation of X be μ and σ, and those for X' be μ' and σ', respectively. Suppose that X and X' are independent chance quantities, i.e., assume that the probability of X taking on any value x_α and X' taking on any value x'_β is equal to the product of the two probabilities, i.e.

$$\Pr(X = x_\alpha \text{ and } X' = x'_\beta) = f(x_\alpha) \cdot f'(x'_\beta) . \tag{9.39}$$

Now consider the chance quantity $L = aX + a'X'$. It will have a distribution of its own, for the probability that L has a particular value l is obtained by adding the values of all products $f(x_\alpha)\cdot f'(x'_\beta)$ for which $ax_\alpha + a'x'_\beta = l$. The mean of the distribution of L is obtained by multiplying $(ax_\alpha + a'x'_\beta)$ by $f(x_\alpha)\cdot f'(x'_\beta)$ and summing over all values of α and of β, i.e.

$$\mu_L = \sum_{\beta=1}^{k'} \sum_{\alpha=1}^{k} (ax_\alpha + a'x'_\beta)\, f(x_\alpha) \cdot f'(x'_\beta) \,. \tag{9.40}$$

But this may be expressed as two sums:

$$\mu_L = \sum_{\beta=1}^{k'} \sum_{\alpha=1}^{k} ax_\alpha f(x_\alpha) \cdot f'(x'_\beta) + \sum_{\beta=1}^{k'} \sum_{\alpha=1}^{k} a'x'_\beta\, f(x_\alpha) \cdot f'(x'_\beta) \,. \tag{9.41}$$

Rearranging summation signs,

$$\mu_L = a\left[\sum_{\alpha=1}^{k} x_\alpha f(x_\alpha)\right]\left[\sum_{\beta=1}^{k'} f'(x'_\beta)\right] + a'\left[\sum_{\beta=1}^{k'} x'_\beta f'(x'_\beta)\right]\left[\sum_{\alpha=1}^{k} f(x_\alpha)\right] \,. \tag{9.42}$$

But the quantity in the first $[\,]$ is the mean μ of X, that in the second $[\,]$ is 1, that in the third $[\,]$ is the mean μ' of X', that in the fourth $[\,]$ is 1. Therefore (9.42) reduces to

$$\mu_L = a\mu + a'\mu' \,, \tag{9.43}$$

which states that *if X and X' are any two independent chance quantities having means μ and μ' respectively, then the mean μ_L of the distribution of the linear combination $L = aX + a'X'$ has the value $a\mu + a'\mu'$* .

Now let us consider the variance of L. By the definition of a variance (see (5.3b)) it is given by

$$\sigma_L^2 = \sum_{\beta=1}^{k'} \sum_{\alpha=1}^{k} (ax_\alpha + a'x'_\beta)^2 \cdot f(x_\alpha) \cdot f'(x'_\beta) - \mu_L^2 \,. \tag{9.44}$$

Squaring the quantity in the parentheses, we have

$$\sigma_L^2 = \sum_{\beta=1}^{k'} \sum_{\alpha=1}^{k} (a^2x_\alpha^2 + 2aa'x_\alpha x'_\beta + a'^2x'^2_\beta)\, f(x_\alpha) \cdot f'(x'_\beta) - \mu_L^2 \,. \tag{9.45}$$

Rearranging the summation signs, and remembering that

$$\sum_{\alpha=1}^{k} x_\alpha f(x_\alpha) = \mu \,, \quad \sum_{\alpha=1}^{k} x_\alpha^2 f(x_\alpha) = \sigma^2 + \mu^2 \,, \quad \sum_{\beta=1}^{k'} x'_\beta f'(x'_\beta) = \mu' \,,$$

$\sum_{\beta=1}^{k'} x'^2 f'(x'_\beta) = \sigma'^2 + \mu'^2$, and $\mu_L = a\mu + a'\mu'$, we find

$$\sigma_L^2 = a^2(\sigma^2 + \mu^2) + 2aa'\mu\mu' + a'^2(\sigma'^2 + \mu'^2) - (a\mu + a'\mu')^2 . \tag{9.46}$$

Simplifying, we find

$$\sigma_L^2 = a^2\sigma^2 + a'^2\sigma'^2 , \tag{9.47}$$

which states that if X and X' are any two independent chance quantities, having variances σ^2 and σ'^2 respectively, then the variance σ_L^2 of the linear combination $L = aX + a'X'$ has the value $a^2\sigma^2 + a'^2\sigma'^2$.

Now let us return to the sampling theory of the differences and sums of means. Formula (9.31) follows immediately from (9.43) which holds for any two chance quantities X and X'. If we replace X by $\overline{X}$, X' by $\overline{X}'$, a by 1, a' by -1, then μ is replaced by $\mu_{\overline{X}}$, μ' is replaced by $\mu_{\overline{X}'}$, $aX + a'X'$ becomes $\overline{X} - \overline{X}'$ and (9.43) reduces to (9.31). In a similar manner it can be seen that (9.35) and (9.37) are special cases of (9.43).

Similarly, formula (9.32) is a special case of (9.47), as one will see by replacing X by $\overline{X}$ and X' by $\overline{X}'$ and letting a = 1, and a' = -1. Also (9.36) is a special case of (9.47).

The extension of formulas (9.43) and (9.47) to the case of a linear combination of any (finite) number of independent chance quantities is straightforward. For instance, in the case of three independent chance quantities X, X', X'', with means and standard deviations μ, σ; μ', σ'; μ'', σ''; respectively, the mean and variance of the linear combination $L = aX + a'X' + a''X''$ are

$$\mu_L = a\mu + a'\mu' + a''\mu''$$

and

$$\sigma_L^2 = a^2\sigma^2 + a'^2\sigma'^2 + a''^2\sigma''^2$$

respectively.

Exercise 9.4.

1. A rolls a die 100 times and B rolls a die 100 times. What is the approximate

probability that A will get a total of at least 25 points more than B?

2. The population of Scholastic Aptitude Test scores has mean 500 and standard deviation 100. Approximately what is the probability that the mean score of a randomly selected group of 50 students will exceed the mean score of a randomly selected group of 25 students by at least 10 points?

3. In problem No. 6 of Exercise 9.3, suppose two packets of 25 booklets are weighed. What is the approximate probability that the difference between the weights will be less than 1/2 oz.?

CHAPTER 10. CONFIDENCE LIMITS OF POPULATION PARAMETERS

10.1 Introductory Remarks.

In Chapter 9 we discussed the rudiments of sampling theory with special reference to means of samples. In all cases we started with a specified population and considered the problem of finding out something about the distribution of means from that population. In particular, we found formulas for obtaining the mean and standard deviation of such a distribution of sample means; we stated (without mathematical proof) that, for large samples, the distribution of means is almost a normal distribution; and we showed how to calculate approximate probabilities from a normal distribution so that the mean of a random sample would fall in any specified interval. Similar results were presented for differences between sample means.

Thus, in sampling theory, we start from a population having a known distribution and deal with the problem of calculating probabilities about samples (e.g., about their means) from it. In practical statistical situations we start from known samples and have to deal with the problem of making inferences about the population (e.g., estimating its mean) from which the sample was drawn. In this problem of statistical inference we make use of sampling theory as a tool for drawing conclusions from a sample about a population.

In this chapter we shall consider the problem of finding confidence limits of population parameters on the basis of sample sums and means for various kinds of problems.

10.2 Confidence Limits of p in a Binomial Distribution.

In Section 8.2 we stated that under certain conditions (large n and value of p not too "near" 0 or 1, particularly) the binomial distribution can be approximated by a normal distribution. More specifically, if X is a chance quantity (number of times an event E occurred in n trials) distributed according to the binomial distribution

$$f_B(x) = C_x^n \, p^x \, q^{n-x}$$

for x = 0, 1, 2, ..., n, then X is approximately normally distributed with mean

np and standard deviation $\sqrt{npq}$. This means that for any value of z, the probability of X being less than $np + z \cdot \sqrt{npq}$ is given approximately by the value of $F_N(x)$ for that value of z in Table 8.1. For example, if z = 2.0,

$$\Pr(X < np + 2\sqrt{npq}) \cong .9773 .$$

Similarly

$$\Pr(np - 2\sqrt{npq} < X < np + 2\sqrt{npq}) \cong .9773 - .0227$$
$$= .9546 .$$

If we knew the values of p and n, we would have specific numerical values for $np \pm 2\sqrt{npq}$ and hence a specific interval such that the probability is approximately .9546 that X would lie in that interval. For example if n = 400, p = .2, q = .8, then $np + 2\sqrt{npq} = 80 \pm 16$ and the interval would be (64, 96). This means that, if X is a chance quantity having the binomial distribution

$$f(x) = C_x^{400}(.2)^x(.8)^{400-x} ,$$

the probability is about .9546 that X would fall between 64 and 96 inclusive.

Now suppose we know that X is a chance quantity which has the binomial distribution

$$f(x) = C_x^{400}p^xq^{400-x}$$

in which the value of p is unknown and that in an experiment (of 400 trials) we found X to be 280. What can we say about the value of p on the basis of this sample evidence? Can we make some kind of estimate of the value of p? The answer is that for a designated "degree of confidence" we can establish a confidence interval on the basis of the sample evidence, such that we cay say with the given degree of confidence that this interval contains the value of p. Suppose we choose a probability of .9546 as the designated "degree of confidence". More precisely we refer to .9546 as the confidence coefficient. Now we know from the preceding discussion that whatever value p may have, it is true that

(10.1) $$\Pr(400p - 2\sqrt{400pq} < X < 400p + 2\sqrt{400pq}) \cong .9546 .$$

The double inequality in the parenthesis is equivalent to the following

double inequality (as will be seen by subtracting 400p from each of the three members of the inequality)

$$-2\sqrt{400pq} < X - 400p < 2\sqrt{400pq} . \tag{10.1}$$

Dividing each member of this inequality by $\sqrt{400pq}$, we obtain the following double inequality

$$-2 < \frac{X-400p}{\sqrt{400pq}} < 2. \tag{10.2}$$

Hence, probability statement (10.1) is equivalent to the following probability statement

$$Pr\left(-2 < \frac{X-400p}{\sqrt{400pq}} < 2\right) \cong .9546 . \tag{10.3}$$

Now for a given value of X in a sample of 400 cases, there is a range of values which p can have such that the inequality (10.2) is satisfied. How can we find this range of values of p? Simply by setting

$$\frac{X-400p}{\sqrt{400p(1-p)}} = +2 \tag{10.4}$$

and

$$\frac{X-400p}{\sqrt{400p(1-p)}} = -2 \tag{10.5}$$

and solving for p. The two values of p that will be found are called the 95.46% confidence limits of p for this problem. To find these values of p we square the two equations and note that in either case we get

$$\frac{(X-400p)^2}{400(p-p^2)} = 4 , \tag{10.6}$$

which reduces to

$$(X-400p)^2 = 1600(p-p^2) . \tag{10.7}$$

(10.7) is a quadratic equation in p, and may be rewritten as

$$161600p^2 - (800X + 1600)p + X^2 = 0 . \tag{10.8}$$

Now the solutions of a general quadratic of the form

$$a p^2 + bp + c = 0 \tag{10.9}$$

are

$$p = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} . \tag{10.10}$$

Applying this formula to the quadratic equation (10.8) and simplifying a bit we obtain

$$p = \frac{X + 2 \pm \sqrt{4X + 4 - .01X^2}}{404} . \tag{10.11}$$

The two values of p given by (10.11) are the two 95.46% confidence limits of p.

It should be realized that what we have done amounts to finding another way to express the double inequality in (10.2). That inequality is equivalent to the following one:

$$\frac{X + 2 - \sqrt{4X + 4 - .01X^2}}{404} < p < \frac{X + 2 + \sqrt{4X + 4 - .01X^2}}{404} . \tag{10.12}$$

Hence we may rewrite (10.1) as follows:

$$\Pr\left\{\frac{X + 2 - \sqrt{4X + 4 - .01X^2}}{404} < p < \frac{X + 2 + \sqrt{4X + 4 - .01X^2}}{404}\right\} \cong .9546 . \tag{10.13}$$

Written in this form the confidence limits of p show up explicitly.

To recapitulate, we may say this: If repeated samples of 400 cases are drawn from the binomial population having the distribution

$$f(x) = C_x^{400} p^x q^{400-x} ,$$

the 95.46% confidence limits in (10.13) will vary from sample to sample since they depend on X, the quantity which varies from sample to sample. For some samples the confidence limits will include the value of the unknown parameter p between them, and for others the confidence limits will not include the value of p between them. But the important point is this: in about 95.46% of the samples, in the long run, the confidence limits will include the unknown value of p between them.

Remember that in the original question we found X = 280 in the sample of 400. Hence, for this particular sample the two confidence limits are obtained by putting X = 280 in (10.11). We find

$$p = .698 \pm .046 .$$

Hence, we say that the true value of the unknown value of p is included between the two values .698 ± .046. We repeat: we can attach about 95.46% confidence to this statement, in the sense that, if we were to repeat the procedure for many other samples, then for about 95.46% of the samples the value of p would be included between the values of the confidence limits for those samples.

We may make the foregoing discussion more concrete by summarizing in the form of an example.

Example: Suppose 400 voters are selected at random from a large city (considered to have an indefinitely large population of voters) and are asked whether they are in favor of Candidate A as a presidential candidate. Suppose 280 voters say "yes". What are the 95.46% confidence limits of the proportion of voters in the city who would say "yes" if they were asked?

In this case the proportion of voters in the city who would say "yes" if asked is p and is obviously unknown. If X is the number who would say "yes" in a sample of 400, then X would theoretically have a binomial distribution if repeated samples of 400 were taken. In a single sample X = 280. The 95.46% confidence limits of p are found by the foregoing procedure to be

$$.698 \pm .046 ,$$

i.e., .652 and .744. This means that we can state on the basis of our sample with "about 95.46% confidence" that the percent of voters who would say "yes" would lie between 65.2 and 74.4.

The procedure we have just discussed can be extended to the case of a general sample of size n (large) from a binomial population and a general confidence coefficient α. In this case we would have a chance quantity X distributed according to the binomial distribution

$$C_x^n p^x (1-p)^{n-x} . \tag{10.14}$$

Then in place of (10.1) we would have

$$\Pr(np - z_\alpha \sqrt{npq} < X < np + z_\alpha \sqrt{npq}) \cong \alpha , \tag{10.15}$$

where values of z_α can be obtained from Table 8.1 for any value of α to be considered. In practice, values of α from .9 to .99 are used most widely. Table 10.1 shows the values of α and z_α most commonly used.

TABLE 10.1

Values of z_α

Confidence coefficient α	z_α
.9000	1.645
.9500	1.960
.9546	2.000
.9900	2.576
.9973	3.000

In place of (10.3) we would have

$$\Pr\left(-z_\alpha < \frac{X-np}{\sqrt{npq}} < +z_\alpha\right) \cong \alpha \,. \tag{10.16}$$

The two confidence limits of p are given by solving the two equations

$$\frac{X - np}{\sqrt{np(1-p)}} = \pm z_\alpha \,. \tag{10.17}$$

Squaring both sides and collecting terms, we obtain the quadratic equation

$$p^2(n^2 + nz_\alpha^2) - p(2nX + nz_\alpha^2) + X^2 = 0 \,. \tag{10.18}$$

The two values of p obtained by solving this equation are the confidence limits of p for confidence coefficient α.

10.21 Confidence interval chart for p.

Figure 10.1 shows a confidence interval chart for determining graphically the two solutions of (10.18) for p; the chart can be used for any given value of $\overline{p} = \frac{X}{n}$ (the relative frequency of "successes" in the sample), for various values of n, and for $\alpha = .95$. In this case $z_\alpha = 1.96$.

To illustrate the use of the chart, suppose a sample of 1000 interviews

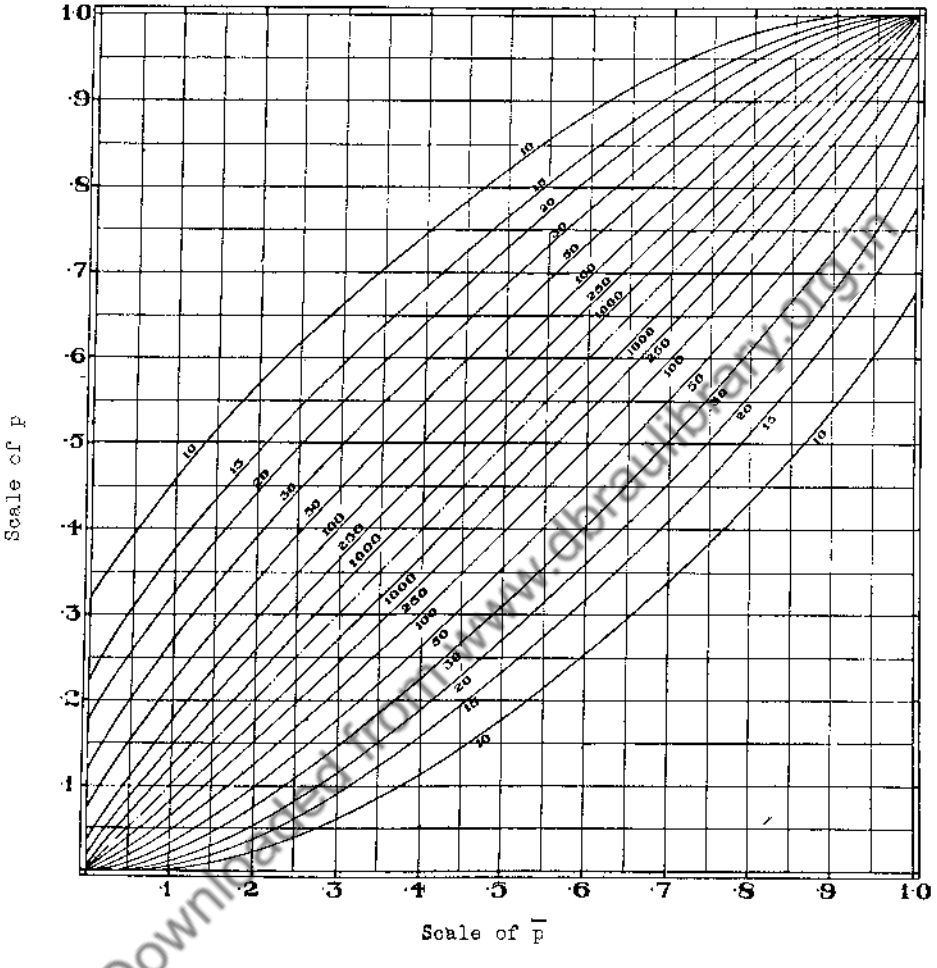


Chart for Determining 95% Confidence Limits of p

in a Binomial Distribution

for n = 10, 15, 20, 30, 50, 100, 250, 1000

(Reprinted by permission of the authors,
C. J. Clopper and E. S. Pearson, and the publishers, the Biometrika Office)

Figure 10.1

is taken at random from the voting population (considered indefinitely large) of a large city. Suppose 240 of the voters answer "yes" to a certain question. What are 95% confidence limits of the percentage of voters in this population who would answer "yes" to the question if asked? Here $\bar{p} = \frac{240}{1000} = 24\%$ and $n = 1000$. Looking at the ordinates in Figure 10.1, determined by the two curves marked $n = 1000$, and for $\bar{p} = .24$, we find .22 and .27 as the required limits.

10.22 Remarks on sampling from a finite binomial population.

Suppose we are sampling from a finite binomial population, i.e., one in which there are only two kinds of elements: A and B. Suppose p is the fraction of A's and q is the fraction of B's. Then there are Np A's and Nq B's in the population. Suppose a sample of n elements is drawn from this population, and let X = number of A's in the sample. Then it follows from (9.6) that X has a distribution which has mean np and standard deviation $\sqrt{npq}\sqrt{\frac{N-n}{N-1}}$; if n is large and N much larger, X is approximately normally distributed with this mean and standard deviation. It follows that we have a way of constructing confidence limits of p in the case of a large finite population. We would proceed by replacing npq in (10.16) by $(npq)\left(\frac{N-n}{N-1}\right)$ and thus obtain

$$\frac{X-np}{\sqrt{np(1-p)}\sqrt{\frac{N-n}{N-1}}} = \pm z_\alpha$$

as the equations to solve for p. The resulting solutions for p would be the $100\alpha\%$ confidence limits.

Exercise 10.2.

1. Suppose a new Roosevelt dime is tossed 10,000 times and turns up heads 5,270 times. Construct 95% confidence limits for p, the probability of getting a head with this dime.

2. If 75 out of a random sample of 225 telephone-subscribing residences in Princeton do not respond to a telephone call between 7:00 and 8:00 p.m., on a particular evening, what would you construct as the 90% confidence limits of the percentage of telephone-subscribing homes having someone at home during those hours? (Assume no answer means no one is at home, and consider the population as indefinitely large.)

3. In a random sample of n voters of State K suppose 51% of the ballots cast are for Candidate A. How large should n be in order for the smaller of the two 99% confidence limits of the percent for A in the population of voters of State K to be 50%?

4. Suppose that at a certain time there are 3600 Princeton undergraduates. A random sample of 400 shows 240 in favor of a certain student proposal. Establish 95% confidence limits of the percent of students in the entire undergraduate body in favor of the proposal.

5. Construct 90% confidence limits of p, the probability that the type of thumb tack discussed in Section 6.3 will fall point up, from the data presented in Table 6.3.

10.3 Confidence Limits of Population Means Determined from Large Samples.

In Section 9.32 it was stated that in large samples of n elements from an indefinitely large population having mean μ and standard deviation σ, the sample mean $\overline{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. This means that for a given α we have

$$\Pr\left(-z_\alpha < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha\right) \cong \alpha \,. \tag{10.19}$$

If the population distribution were such that μ and σ were each expressible in terms of a single parameter θ, ther one could find confidence limits of θ by taking the end points of the range of all possible values of θ for which the inequality in (10.19) holds. These limits would involve $\overline{X}$ and n in general. A simple example will make this clear.

Example: Suppose X is a continuous chance quantity such that all values of X on the interval (0, θ) are equally likely, where θ is unknown. (This is equivalent to the simple scale problem of Section 5.21 where the length of the scale is θ rather than 1.) Suppose a sample of 20 values of X is "drawn", and the sample mean $\overline{X}$ has the value 3.2. What are the 90% confidence limits of θ?

In this problem, the probability density function is $f(x) = \frac{1}{\theta}$, over the interval (0, θ). Hence

$$\mu = \int_0^\theta x \cdot \frac{1}{\theta} \cdot dx = \frac{x^2}{2\theta}\Big]_0^\theta = \frac{\theta}{2}$$

$$\sigma^2 = \int_0^\theta x^2 \cdot \frac{1}{\theta}\, dx - \mu^2 = \frac{\theta^3}{3\theta}\Big]_0^\theta - \mu^2$$

$$= \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12} .$$

That is

$$\mu = \frac{\theta}{2} , \text{ and } \sigma = \frac{\theta}{\sqrt{12}} .$$

For $\alpha = .90$ we see from Table 10.1 that $z_\alpha = 1.645$. Hence (10.19) for this example becomes

$$\Pr\left(-1.645 < \frac{\bar{X} - \frac{\theta}{2}}{\theta/\sqrt{12n}} < +1.645\right) \cong .90 . \tag{10.20}$$

Solving the two equations

$$\frac{\bar{X} - \frac{\theta}{2}}{\theta/\sqrt{12n}} = \pm 1.645$$

we obtain the following two values of θ,

$$\theta = \frac{\bar{X}}{\frac{1}{2} \pm \frac{1.645}{\sqrt{12n}}} . \tag{10.21}$$

which are the approximate 90% confidence limits of θ in terms of $\bar{X}$ and n. But in our specific sample we have $\bar{X} = 3.2$ and $n = 20$. Substituting in (10.21) we find the approximate 90% confidence limits of θ for our example to be

$$\theta = 5.28 \text{ and } 8.12 .$$

If σ and μ are not expressible in terms of the same parameter, and if σ is known, then for large samples we may write (10.19) as follows:

$$\Pr\left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \cong \alpha , \tag{10.22}$$

from which it is clear that $100\alpha\%$ confidence limits of μ are

$$\bar{X} \pm z_\alpha \frac{\sigma}{\sqrt{n}} . \tag{10.23}$$

As a matter of fact, if σ is unknown and n is large, we may replace σ in (10.23) by the sample standard deviation s, and obtain

$$\bar{X} \pm z_{\alpha} \frac{s}{\sqrt{n}}, \tag{10.24}$$

as approximate 100α% confidence limits of the population mean μ.

Example: A sample of 100 washers is taken at random from a stamping machine during a certain afternoon and the thicknesses of the washers are measured. The mean and standard deviation of the thicknesses of these 100 washers are .1022" and .0021" respectively. What are (approximate) 90% confidence limits of the mean of the population of washers turned out that afternoon?

Here $\bar{X}$ = .1022, and s = .0021, and n = 100. For 90% confidence limits we see from Table 10.1 that z = 1.645. Hence the (approximate) 90% confidence limits of the population mean μ are

$$.1022 \pm (1.645) \frac{.0021}{\sqrt{100}},$$

i.e.,

$$.1022 \pm .00035.$$

Thus, the probability is about .90 that the population mean μ is included between .10185" and .10255".

10.31 Remarks about confidence limits of means of finite populations.

Sometimes we have to consider the problem of determining confidence limits of the mean μ of a finite population of N elements by using a sample of n elements from this population (assuming n and N large enough for the mean $\bar{X}$ to be approximately normally distributed). In this case we replace σ by $\sigma\sqrt{\frac{N-n}{N-1}}$ in (10.23) and s by $s\sqrt{\frac{N-n}{N-1}}$ in (10.24).

Exercise 10.3.

1. Establish 90% confidence limits of μ, the mean of the population of weights of zinc coatings from which the sample in Table 2.1 may be considered as having been drawn. (The sample mean and standard deviation are given in Section 3.2.)

2. In a sample of 270 bricks from a certain population, the mean of the transverse

strengths is 999.8, and the standard deviation is 202.1. Construct 95% confidence limits of the mean transverse strength for the population.

3. It is known that a chance quantity X is distributed in a population in accordance with a Poisson distribution, but the constant m in the distribution is unknown. A sample of 200 elements from this population has a mean equal to 3.4. Construct 95% confidence limits of m. (Refer to Chapter 7 for mean and variance of a Poisson distribution.)

4. Suppose that each person at a large national convention is supplied with a badge with a serial number on it. At this convention you look around through the corridors and record a sample of 100 numbers. You find that these 100 numbers add up to 24,520. Set up 90% confidence limits of k, the number of people registered at the convention. (See Problem No. 5, Exercise 9.2 for the sum of integers and sum of squares of integers from 1 to k.)

10.4 Confidence Limits of Means Determined from Small Samples.

If n, the sample size, is small, we cannot in general use the confidence limits (10.23) or (10.24) for μ. Under certain conditions, however, we can establish confidence limits for μ which are accurate enough for many practical situations. In many problems of practical statistical importance, the indefinitely large population from which we actually draw our sample is approximately normal itself. This is true of many populations of measurements. For example, we can see from Chapter 2 and from Figures 2.4 and 8.7 that it would probably not be rash to make the assumption that the population from which the zinc coating weights could be assumed to have come is approximately normal. It will be remembered from Section 9.3 that it was stated that the theoretical sampling distribution of the mean $\bar{X}$ in samples of size n from an indefinitely large normal population is exactly (not approximately) normal. This means that if $\bar{X}$ is the mean of a sample from an indefinitely large normal population the probability expressed on the left of (10.19) and (10.22) is exactly (not approximately) equal to α, no matter what value n has. If σ were known, then exact $100\alpha\%$ confidence limits of μ would be given by (10.23). But if the population is unknown, (although assumed to be normal), we could not use these confidence limits because σ would be unknown. We can get around the difficulty by substituting the sample standard deviation s for the unknown population standard

deviation σ. This is not unreasonable, since we may regard s as an estimate of σ. We then write

$$\Pr\left(-t_\alpha < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_\alpha\right) \cong \alpha \,, \tag{10.25}$$

where t_α is a quantity depending on n and α. More precisely t_α depends on α and the number of degrees of freedom of s which replaced σ in (10.19). It was stated in Section 3.1 that the number of degrees of freedom of s is n-1. Values of t for the practically useful values of α and various numbers of degrees of freedom are given in Table 10.2. It will be seen that where the number of degrees of freedom = ∞, we have $t_\alpha = z_\alpha$. In other words, for indefinitely large samples, it makes no difference if we replace σ by s; one might expect this to be the case, since the standard deviation of an indefinitely large sample is the same as the standard deviation of the population from which the sample came.

The mathematical argument for the validity of statement (10.25) is beyond the scope of this course. The problem amounts to finding the theoretical sampling distribution of the quantity

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

in samples from a normal population with mean μ. The sampling distribution of this quantity does not involve the standard deviation of the normal population; it is called the "Student" t distribution and resembles the normal distribution quite a lot. In fact, as n increases indefinitely, the t distribution approaches a normal distribution with mean 0 and variance 1 as its limiting distribution.

Returning to (10.25) we see that for a given value of α and a given sample, the only unknown quantity is μ. Thus if we are sampling from an indefinitely large normal population with unknown mean μ, the confidence limits of μ for confidence coefficient α are given by solving the equations

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \pm t_\alpha \,, \tag{10.26}$$

where t_α is determined from Table 10.2 for n-1 degrees of freedom. The confidence limits of μ are

$$\mu = \bar{X} \pm t_\alpha \frac{s}{\sqrt{n}} \,. \tag{10.27}$$

TABLE 10.2

Values of t_a for a = .99, .95, .90 and Degrees of Freedom from 1 to 30

Degrees of Freedom	a		
	.99	.95	.90
1	63.657	12.706	6.314
2	9.925	4.303	2.920
3	5.841	3.182	2.353
4	4.604	2.776	2.132
5	4.032	2.571	2.015
6	3.707	2.447	1.943
7	3.499	2.365	1.895
8	3.355	2.306	1.860
9	3.250	2.262	1.833
10	3.169	2.228	1.812
11	3.106	2.201	1.796
12	3.055	2.179	1.782
13	3.012	2.160	1.771
14	2.977	2.145	1.761
15	2.947	2.131	1.753
16	2.921	2.120	1.746
17	2.898	2.110	1.740
18	2.878	2.101	1.734
19	2.861	2.093	1.729
20	2.845	2.086	1.725
21	2.831	2.080	1.721
22	2.819	2.074	1.717
23	2.807	2.069	1.714
24	2.797	2.064	1.711
25	2.787	2.060	1.708
26	2.779	2.056	1.706
27	2.771	2.052	1.703
28	2.763	2.048	1.701
29	2.756	2.045	1.699
30	2.750	2.042	1.697
∞	2.576	1.960	1.645

(The entries of this table taken from Statistical Methods for Research Workers by permission of the author, R. A. Fisher, and the publishers, Oliver and Boyd, Edinburgh)

These confidence limits of μ have been found to be satisfactory even for small samples (less than 30) from populations which depart "moderately" from normal populations - such as those frequently met in statistical practice (population of weights, lengths, scores, and other such measurements).

Example: A sample of 11 rats from a "control" population had an average blood viscosity of 3.92 (units need not be specified) with a standard deviation of .61. On the basis of this sample, establish 95% limits of μ, the mean blood viscosity of the "control" population.

Here we have $\bar{X} = 3.92$, $n = 11$, $s = .61$, and $\alpha = .95$. Looking at Table 10.2 for $n-1 = 10$ degrees of freedom, and $\alpha = .95$, we find $t_\alpha = 2.228$. Hence the 95% confidence limits for μ are given by substituting these quantities in (10.27), i.e.,

$$3.92 \pm 2.228 \frac{(.61)}{\sqrt{11}}$$

or

$$3.92 \pm .41 .$$

We say that the probability is about .95 that the two confidence limits 3.51 and 4.33 include the mean of the blood viscosity measurements of the population of "control" rats being sampled. The main assumptions here are that we are drawing a sample of 11 "at random" from a population in which viscosity measurements are "almost" normally distributed.

Exercise 10.4.

1. Ten short pieces of copper wire from 10 rolls of wire have the following breaking strengths (in lbs.): 578, 572, 570, 568, 572, 570, 570, 572, 596, 584. Construct 90% confidence limits of μ, the mean of the breaking strengths for the population from which the sample is considered to have been drawn.

2. Chemical determinations of percent of iron in five random batches of iron ore from a certain deposit had an average of 22.3% and a standard deviation of 1.8%. Establish 95% confidence limits of the mean percent of iron in the deposit.

3. Suppose a plot of land is surveyed by 5 student surveyors and they find the following areas for the plot (in acres): 7.27, 7.24, 7.21, 7.28, 7.23. On the

basis of this information, construct 90% confidence limits of the area of the plot.

10.5 Confidence Limits of Difference between Population Means Determined from Large Samples.

In Section 9.4 we stated that if two large samples are drawn respectively from two much larger finite populations or two indefinitely large populations, then the difference between the sample means has a theoretical distribution which is approximately normal. We may use this fact in getting approximate confidence limits of the difference $(\mu - \mu')$ of the two population means.

More specifically, suppose $\overline{X}$ is the mean of a large sample of n elements from an indefinitely large population with mean μ and standard deviation σ, and $\overline{X}'$ is the mean of a large sample of n' elements from an indefinitely large population with mean μ' and standard deviation σ'. Then it follows from Section 9.41 that we may write the following approximation:

$$(10.28)\quad \Pr\left[(\mu-\mu') - z_\alpha\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}} < (\overline{X}-\overline{X}') < (\mu-\mu') + z_\alpha\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}}\right] \cong \alpha ,$$

where z_α is the same z_α we have mentioned before, for which the useful values are given in Table 10.1. Expression (10.28) may be rewritten as

$$(10.29)\quad \Pr\left[(\overline{X}-\overline{X}') - z_\alpha\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}} < (\mu-\mu') < (\overline{X}-\overline{X}') + z_\alpha\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}}\right] \cong \alpha ,$$

showing explicitly that the $100\alpha\%$ confidence limits of $\mu - \mu'$ are

$$(10.30)\quad \overline{X} - \overline{X}' \pm z_\alpha\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}} .$$

Again we state that if σ^2 and σ'^2 are unknown (remember that we are considering large values of n and n') we may substitute the sample variances s^2 and s'^2 for the unknown σ^2 and σ'^2, thus giving us the confidence limits

$$(10.31)\quad \overline{X} - \overline{X}' \pm z_\alpha\sqrt{\frac{s^2}{n} + \frac{s'^2}{n'}} .$$

An example will perhaps clarify the situation.

Example: A certain psychological test was given to two groups (samples) of Army prisoners: (a) first offenders and (b) recidivists. The sample statistics were as follows (Betts data):

Population	Sample size	Sample mean	Sample standard deviation
(a) first offenders	580	34.45	8.83
(b) recidivists	786	28.02	8.81

What are 95% confidence limits of the difference of the means for the two populations?

Let μ and μ' be the means of the scores for the populations of first offenders and of recidivists, respectively. We have $n = 580$, $n' = 786$, $\overline{X} = 34.45$, $\overline{X}' = 28.02$, $s = 8.83$, and $s' = 8.81$. For $\alpha = .95$ we see from Table 10.1 that $z_\alpha = 1.96$. Hence the 95% confidence limits of $(\mu-\mu')$ are found by substituting in (10.31):

$$\left[(34.45) - (28.02)\right] \pm (1.96)\sqrt{\frac{(8.83)^2}{580} + \frac{(8.81)^2}{786}}$$

or

$$6.43 \pm .95 .$$

Thus, the probability is about .95 that the difference between the two population means, i.e., $(\mu - \mu')$ is included between 5.48 and 7.38.

If we are working with large finite populations instead of indefinitely large ones, then the only alteration we make in the confidence limits for $(\mu - \mu')$ is to replace σ^2 by $\sigma^2\left(\frac{N-n}{N-1}\right)$ and σ'^2 by $\sigma'^2\left(\frac{N'-n'}{N'-1}\right)$ in (10.30), where N and N' are the numbers of elements in the two populations respectively.

10.51 Confidence limits of the difference p-p' in two binomial populations.

In case the two populations involved are binomial populations, where p is the probability of "success" in the first population, and p' is that in the second population, then $\mu = p$ and $\mu' = p'$, $\sigma^2 = pq$ and $\sigma'^2 = p'q'$. In this case the sample means $\overline{X}$ and $\overline{X}'$ simply become the proportions of "successes" in the samples, and $(\mu-\mu')$ is simply $(p-p')$, the difference between the proportions of "successes" in the population. The 100α% confidence limits of $(p-p')$ are therefore (for large values of n and n')

$$(\overline{X} - \overline{X}') \pm z_\alpha\sqrt{\frac{pq}{n} + \frac{p'q'}{n'}} \tag{10.32}$$

which is a special case of (10.30). Since p and p' are unknown, we replace p

by $\bar{X}$ and p' by $\bar{X}'$ (which can be done only in the case of samples from binomial populations) and obtain 100α% confidence limits of p-p' as follows:

$$(10.33) \qquad (\bar{X} - \bar{X}') \pm z_\alpha \sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{X}'(1-\bar{X}')}{n'}} .$$

Let us illustrate this by an example.

Example: Suppose a sample of 400 voters of town A showed 230 in favor of candidate K and a sample of 500 voters of town B showed 200 voters in favor of candidate K. Construct 95% confidence limits of (p-p') (the fraction of voters in town A who support K minus the fraction in town B who support K).

Let us regard the population of voters in A and in B as being indefinitely large. We have the following sample values

$$n = 400 \qquad \bar{X} = \frac{230}{400} = .575 ,$$

$$n' = 500 \qquad \bar{X}' = \frac{200}{500} = .400 .$$

For $\alpha = .95$, $z_\alpha = 1.96$. Hence the 95% confidence limits of p-p' are

$$(.575-.400) \pm 1.96\sqrt{\frac{(.575)(.425)}{400} + \frac{(.400)(.600)}{500}}$$

or

$$.175 \pm .065 .$$

Hence the probability is about .95 that the fraction of voters in town A supporting candidate K minus the fraction of voters in town B supporting candidate K is included between .110 and .240 .

10.52 Confidence limits of the difference of two population means in case of small samples.

In experimental work, it frequently happens that we have to deal with two small samples (less than 30, say) which can be assumed to come from two indefinitely large populations which are "nearly" normal, and which have equal variances. Thus in an experiment, we may make measurements on a sample of individuals in a "control" group, and make similar measurements on a sample of individuals in an "experimental" group. In such a situation it often happens that the measurements in the "experimental" group "appear" to come from a population which has nearly the same variance as the population from which the

"control" group comes but which has a different mean.

If the two population variances are equal, i.e., $\sigma'^2 = \sigma^2$, then $\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}} = \sigma\sqrt{\frac{1}{n} + \frac{1}{n'}}$, and (10.29) may be written as

$$(10.34) \qquad \Pr\left[(\bar{X}-\bar{X}') - z_\alpha \sigma\sqrt{\frac{1}{n} + \frac{1}{n'}} < (\mu - \mu') < (\bar{X} - \bar{X}') + z_\alpha \sigma\sqrt{\frac{1}{n} + \frac{1}{n'}}\right] \equiv \alpha .$$

If the two populations are normal the expression on the left in (10.34) is exactly equal to α. Now if σ is unknown, as is in fact usually the case, we shall proceed to replace it by an estimate of σ which we can make from the two sample standard deviations. The estimate we use for σ is

$$(10.35) \qquad \sqrt{\frac{(n-1)s^2 + (n'-1)s'^2}{n + n' - 2}} .$$

This does not look unreasonable. For $(n-1)s^2$ is the sum of squares of deviations of X's in the first sample from their mean, and $(n'-1)s'^2$ is a similar quantity for the second sample. Adding them we get sums of squares of deviations for both samples which we are assuming to come from populations with the same variance. The number of degrees of freedom "contributed" by $(n-1)s^2$ is $n-1$, and the number "contributed" by $(n'-1)s'^2$ is $n'-1$; thus the total number of degrees of freedom is $n + n' - 2$, which is seen to be the divisor in (10.35).

Now we can construct confidence limits for $(\mu-\mu')$ from two small samples on the basis of the following statement:

If two samples of n and n' elements respectively are drawn from two normal populations having the same variance, and if $\bar{X}$ and $\bar{X}'$ are the sample means, and s^2 and s'^2 are the sample variances, then the probability is α that $(\mu-\mu')$ will be included between the two values

$$(10.36) \qquad (\bar{X}-\bar{X}') \pm t_\alpha \sqrt{\frac{(n-1)s^2 + (n'-1)s'^2}{n + n' - 2}} \cdot \sqrt{\frac{1}{n} + \frac{1}{n'}} ,$$

where t_α is determined from Table 10.2 for $n + n' - 2$ degrees of freedom. This is simply to say that (10.36) gives the $100\alpha\%$ confidence limits of $(\mu-\mu')$. In (10.36) the number of degrees of freedom of the estimate we have used for σ^2 is $n + n' - 2$. This means that for a given α we look up t_α in Table 10.2 under that value of α and for the number of degrees of freedom equal to $n + n' - 2$.

Example: Two methods of determining nickel content of steel, say A and B, are

tried out on a certain kind of steel. Samples of four determinations are made by each method, with the following results (Raithel data):

Method	Sample size	Sample mean	Sample variance
A	4	3.285%	.000033
B	4	3.258%	.000092

Construct 95% confidence limits of the difference of the means of the populations associated with methods A and B respectively.

We have $n = n' = 4$, $\overline{X} = 3.285$, $\overline{X}' = 3.258$, $3s^2 = .000099$, $3s'^2 = .000276$, and the number of degrees of freedom $(n + n' - 2) = 6$. For $\alpha = .95$ and 6 degrees of freedom, we have from Table 10.2, $t_\alpha = 2.447$. Hence the 95% confidence limits of $\mu - \mu'$ are obtained by substituting in (10.36):

$$.027 \pm 2.447\sqrt{\frac{.000375}{6}} \cdot \sqrt{\frac{1}{4} + \frac{1}{4}}$$

or

$$.027 \pm .014 .$$

Thus the probability is about .95 that the difference $(\mu-\mu')$ between the means of the two populations is included between .013% and .041%.

Exercise 10.5.

1. An attitude test was given to two groups of soldiers: (a) a group of operative soldiers (who had been in the Army for a while) and (b) new selectees. The information on the two samples of soldiers was as follows (Betts data):

Sample	Sample number	Sample mean	Sample standard deviation
(a)	1050	47.65	6.77
(b)	531	46.10	6.79

If μ is the mean of the population from which (a) came and μ' that of the population from which (b) came, establish 95% confidence limits of $\mu-\mu'$.

2. The following question was asked in a poll of 148 men and 152 women in Trenton: "Do you approve or disapprove of the practice of tipping by and large?"

The results were as follows (Crespi data):

Group	Sample size	No. who answered "yes"
Men	148	89
Women	152	116

Construct 95% confidence limits of $(\mu-\mu')$, the difference between proportion of "yeses" among population of men in Trenton and proportion of "yeses" among women.

3. In an experiment on two groups of rats: (A) "Normal", and (B) "Adrenalectomized", the following blood viscosity readings were found (Nice and Fishman data):

(A)	(B)
3.29	3.45
3.91	4.26
4.64	4.71
3.55	3.14
3.67	3.45
4.18	5.01
3.74	4.43
4.67	4.91
3.03	4.22
4.61	4.83
3.84	3.55

Construct 95% confidence limits of $(\mu-\mu')$, the difference between the mean of the blood viscosity of the population of "normal" and that of the population of "adrenalectomized" rats.

CHAPTER 11. STATISTICAL SIGNIFICANCE TESTS.

11.1 A Simple Significance Test.

The problem of making statistical significance tests is very closely related to that of determining confidence limits. Roughly speaking, a statistical significance test is a probability test, based on sampling theory, to determine whether a sample could have "reasonably" come from a specified population, or from a population with specified values of its parameters. Let us illustrate by a simple example.

Example: A claims he can roll a die in such a way that he can make aces come up more often than the "average person" can. He demonstrates by rolling a die 600 times and turning up 120 aces. He claims that this is an average of one ace in five rolls instead of one in six which proves his case. How can we test whether 120 aces in 600 rolls is "unreasonably" large on the basis of the behavior of a normal die?

Our approach to this question is this: Suppose the die is true and that it is falling according to the "laws" of a true die. We want to find the probability of getting 120 aces or more with a true die. If this probability is not "too small" we can discredit A's claim. If X is the number of aces in 600 rolls of a true die, it has the binomial distribution

$$f(x) = C_x^{600}(\tfrac{1}{6})^x(\tfrac{5}{6})^{600-x} .$$

This is a binomial distribution with $p = 1/6$, $n = 600$. X is approximately normally distributed with mean $np = 100$, and standard deviation $\sqrt{npq} = \sqrt{\frac{500}{6}} = 9.13$. If we set

$$Z = \frac{X-100}{9.13}$$

then Z is approximately normally distributed with mean 0 and variance 1. Let us choose some small probability level, say .01, and find a number so that the probability of X exceeding that number is .01. We find from Table 8.1 that the value of z for this probability level is 2.33. The value of x corresponding to this value of z is $100 + (2.33)(9.13) = 121.3$ which is larger than the

sample 120. We say that the 120 as obtained by A does not significantly exceed the 100 aces expected from a true die, at the 1% probability level. Another way to say it is this: We tested the hypothesis that $p = 1/6$ on the basis of the sample, and we did not reject the hypothesis at the 1% probability level.

The hypothesis that $p = 1/6$ would be referred to as the null hypothesis in this problem. This means that if we agree to work at the 1% probability level A cannot be regarded as having proved his claim on the basis of the evidence provided by the 600 throws. 120 aces or more in 600 rolls will occur with a probability greater than .01 with an ordinary "true" die behaving like a "true" die. If A had obtained 150 aces, say, (or more than 121 aces), then the difference between this number and the 100 aces expected from a true die would have been statistically significant at the 1% probability level. We would then have to admit that the data available support A's claim at the 1% probability level.

11.2 Significance Tests by using Confidence Limits.

We can also look at this problem from the point of view of confidence limits. For let p be the probability of getting an ace, and let us ask whether the 99% confidence limits of p will include the value 1/6 (the value of p in the case of a true die) between them? We do not have to find the 99% confidence limits of p to answer this question. All we have to do is to note from Table 10.1 that for $\alpha = .99$, $z_\alpha = 2.576$ and see whether $p = 1/6$ will satisfy the double inequality

$$-2.576 < \frac{120-600p}{\sqrt{600p(1-p)}} < +2.576 .$$

For $p = 1/6$, the expression in the middle part of the inequality is

$$\frac{120 - 100}{\sqrt{600(\frac{1}{6})(\frac{5}{6})}} = \frac{20}{\sqrt{\frac{500}{6}}} = \frac{20}{9.13} = 2.19$$

which clearly lies between -2.576 and + 2.576. In other words, since 2.19 lies between -2.576 and 2.576, 120 aces does not differ significantly from the expected number 100, and A's claim is not supported at the 99% confidence level.

In using the confidence interval method, we ask whether the absolute value of the difference between the number of aces obtained and that expected is statistically significant, i.e., whether the number of aces obtained were high or

low. This is the reason that $z_\alpha = 2.576$ in the case of two confidence limits, and $z_\alpha = 2.33$ in the case originally considered, in which we were asking whether 120 is significantly larger than 100.

In many situations a statistical significance test made on the basis of a sample can be easily made by simply checking whether a value of a population parameter in which we are particularly interested (which is specified in the hypothesis being tested) lies between confidence limits for that parameter as determined from the sample. This can often be done without actually finding the confidence limits.

Let us consider an example involving the difference between two means.

Example: Suppose two machines, say A and B, are packaging "6-ounce" cans of talcum powder, and that 100 cans filled by each machine are emptied and the contents carefully weighed. Suppose the following sample values are found:

Machine	Sample size	Mean weight of contents (oz.)	Standard deviations
A	100	6.11	.04
B	100	6.14	.05

Are these means significantly different at the 99% confidence level?

What we are asking here is this: Could these two samples reasonably have come from populations having equal means? To answer this we find out whether the value $(\mu-\mu') = 0$ is included between the 99% confidence limits of $(\mu-\mu')$. In this case, no additional effort is required to actually find the 99% confidence limits of $(\mu-\mu')$. They are

$$(6.14-6.11) \pm 2.576\sqrt{\frac{.0016}{100} + \frac{.0025}{100}}$$

or

$$.03 \pm .016 .$$

The confidence limits are .014 and .046, which do not include 0 between them. Hence we conclude that at the 99% confidence level the difference between the two sample means is significantly different from 0. In other words, we have tested the hypothesis (null hypothesis) that $(\mu-\mu') = 0$ and we reject it at the 99% confidence level. This means that we can be practically certain on the basis of the two samples that machine B is putting more powder into the boxes

on the average than machine A.

11.3 Significance Tests without the use of Population Parameters.

In making some significance tests one often has first to set up a null hypothesis and then to determine the probability distribution or theoretical sampling distribution under the null hypothesis. In such a case one does not have an opportunity to use confidence limits. An example will make this clear.

Example: Refer to problem No. 5 in Exercise 5.1. Suppose the person who is smoking the cigarettes claims he can distinguish brands A, B, C, D blindfolded. How would we test his claim?

An experiment is set up as described in that problem. The null hypothesis here is that he cannot distinguish the cigarettes and that any assignment of the letters A, B, C, D he makes is done "at random", which means that all assignments are considered to be equally likely. If X is a chance quantity denoting the number of correct assignments, then under the null hypothesis the 4! = 24 different assignments provide the following probability distribution of X:

x	0	1	2	3	4
f(x)	$\frac{9}{24}$	$\frac{8}{24}$	$\frac{6}{24}$	0	$\frac{1}{24}$

Now we consider ability to identify to be indicated or measured by large values of X. The largest value of X possible is 4 and its probability is $\frac{1}{24}$, which is about .0417. Since .0417 is greater than .01 it is clear that if we adopt a 1% probability level we do not have an opportunity to reject the null hypothesis at this probability level. Not enough experimenting has been done. But suppose this entire experiment were repeated twice and all four brands were identified in both experiments. The probability of this happening under the null hypothesis is $(\frac{1}{24})^2 = \frac{1}{576}$. Under this condition we would reject the null hypothesis at the 1% probability level and say that this person has made a significantly large number of correct identifications; we would conclude that he has some ability to discriminate among these brands of cigarettes.

Exercise 11.

1. Suppose a sack of 400 nickels of a new design is emptied on a table, and 235 heads appear. Does this number differ significantly at the 1% probability level from that expected of a "true" coin (for which heads and tails are equally likely)?

2. A poll of 400 men and of 400 women in a certain town shows 270 of the men and 240 of the women in favor of a certain proposal. Is there a significant difference between the opinion of the men and the opinion of the women on this proposal at the 5% probability level?

3. Suppose a random sample of 50 entering freshmen at College A has a mean S.A.T. score of 560 and standard deviation of 90, while a sample of 50 entering freshmen at College B has a mean S.A.T. score of 565 and a standard deviation of 95. Test, at the 5% probability level, the null hypothesis that the populations of entering freshmen at the two colleges have the same mean S.A.T. score. Assume each college is admitting 500 freshmen.

4. Methods A (Dichromate method) and B (Thioglycolic acid method) of determining iron content of ore were tried out on each of 21 batches of iron ore. The difference between the percent of iron found by method A and that found by method B was obtained for each batch. These differences in percent were as follows (Mehlig and Shepherd data):

+0.05	-0.05	+0.04
+0.04	+0.03	+0.03
-0.05	-0.02	+0.09
-0.10	0.00	+0.05
+0.02	-0.07	+0.05
+0.01	-0.01	0.00
+0.09	-0.05	-0.01

Is there any significant difference between the mean iron content yield as determined by the two methods at the 5% probability level?

5. It is known that the probability of getting no aces in a hand of bridge under perfect shuffling is about .3. A complains about the cards and/or shuffling on the basis of the fact that in 4 hands he got only one ace. Does he have any grounds for complaint at the 1% probability level? (Use binomial distribution in making significance test.)

6. A bag has 100 chips in it; some are white and the rest are red. A draws a chip and notices it is white. He returns the chip to the bag, mixes up the chips and draws again. He repeats this 5 times, obtaining a white chip each time. What is the smallest number of white chips which can be in the bag without making A's 5 white chips a significantly large number of white chips to be drawn in 5 draws at the 1% probability level?

7. A 5-cent plastic die was rolled 100 times and a total of 370 dots was obtained. Test the null hypothesis that this die is true at the 1% probability level. (Hint: make use of approximate normality of S(X).)

CHAPTER 12. TESTING RANDOMNESS IN SAMPLES.

12.1 The Idea of Random Sampling.

The most fundamental concept underlying probability theory and statistical inference is that of randomness. We have essentially taken this as an undefined concept throughout all of the previous chapters. We have merely implied, in a general way, that randomness means something like haphazardness with which values of a chance quantity X vary from trial to trial, or from drawing to drawing in an actual experiment or sample. In order to gain some idea as to whether the successive measurements in a sample are exhibiting this property of randomness, we must make use of the information contained in the order in which the measurements occur in building up the sample. In all of the discussion in the previous chapters we have ignored this order information and have considered only the frequency distribution of the magnitudes of the individual measurements in the sample. But we are now in a position to make use of what we know about sampling and significance tests, to make a more definite statement about whether the successive observations on a chance quantity X (i.e., whether the successive measurements in a sample) are behaving in a "random" manner.

12.2 Runs.

Many forms of non-randomness could conceivably exhibit themselves in a sample of n elements. Suppose, for example, that a coin is tossed 20 times and let us consider three "kinds" of sequences as follows:

Sample I: T T T T H H T T T T T T H H H H H H H H

Sample II: H T H T H T H T H T H T H T H T H T H T

Sample III: T T H H T H T T H T H H T T T H T H H H .

What features of these sequences (samples), if any, are most likely to arouse suspicion of non-randomness? In the case of I, it is the fact that there are long runs (and few of them) of H's and T's. In the case of II, it is the fact that there are short runs (and many of them) as well as regularity of H's and T's. In the case of III, we probably would not be suspicious.

Again, let us consider outside diameters of 15 shafts being turned out by an automatic lathe (a shaft being picked at random every half hour). Consider three samples of shaft diameters and suppose they look like this (measurements in inches):

TABLE 12.1

Samples of Shaft Diameters (in inches)

Sample A	Sample B	Sample C
.501	.502	.505
.502	.501	.501
.502	.500	.502
.503	.503	.502
.502	.501	.506
.503	.502	.502
.503	.503	.507
.504	.509	.505
.505	.508	.503
.505	.509	.504
.506	.503	.502
.505	.510	.505
.507	.509	.506
.507	.510	.502
.508	.510	.501

Presented graphically, these three samples are pictured in Figure 12.1 (the a's and b's are explained later). What features of these sequences are most likely to arouse suspicion as to whether they are "sufficiently" random? In the case of Sample A, the feature which makes one wonder is the way in which the 15 measurements in the sample seem to rise in a general way as the sampling progresses. In the case of Sample B, the suspicious feature is the jump after the seventh drawing to a generally higher level. Sample C seems to have a reasonable degree of randomness.

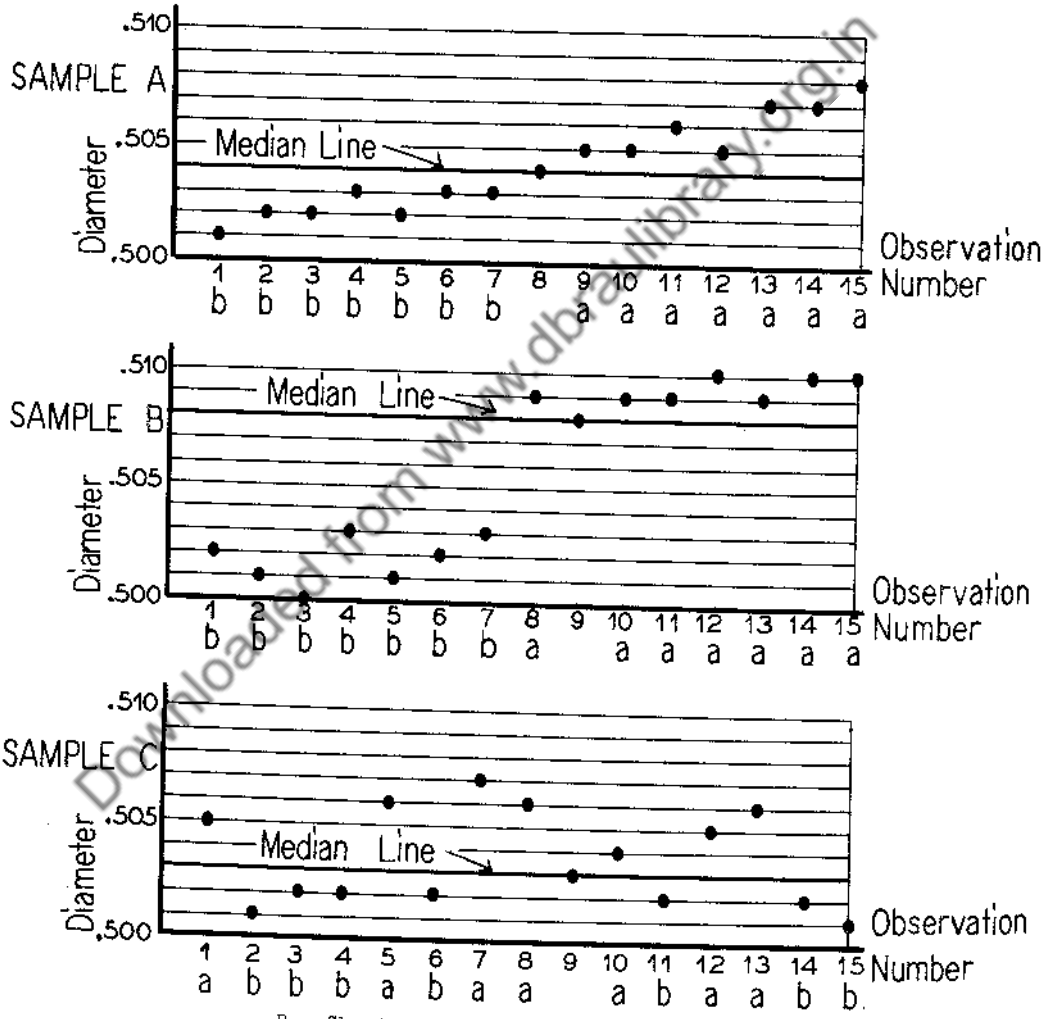
These features can be made more objective in the following way: Referring to Figure 12.1, suppose we draw a horizontal line (median line) so that there are as many points above the line as below it. Then for every point below the median line write b under that point and below the horizontal axis; for every point above the median line write a in a similar way. (See Figure 12.1.) For samples A, B and C we have the following rows of a's and b's (14 in each sample):

Sample A: b b b b b b b a a a a a a a

Sample B: b b b b b b b a a a a a a a

Sample C: a b b b a b a a a b a a b b .

In other words, we have reduced the three samples of diameter readings to three samples of a's and b's. Samples A and B, which looked suspiciously non-random when viewed graphically, also look suspiciously non-random when viewed in terms of a's and b's.



Run Chart for Data in Table 12.1

Figure 12.1

In trying to draw a median line for samples of 10 to 50 one may find that there is no such line, and that the line coming nearest to dividing the points into two sets having equal numbers of points will pass through several points (usually two or three in practice). In this case one may assign one or more of these points on the line to the side having the smaller number of points in order to make the number of points above the median line equal to the number below. The assignment of each point should be made so as to increase the number of runs.

So far, we have just used our common sense or intuition as to whether the observations in these various samples appear to be random. In this intuitive analysis we have associated non-randomness with a few very long runs (as in Sample I of the coin-tossing example and Samples A and B of the shaft-turning example); or many very short runs (as in Sample II of the coin-tossing example). Now, it is a matter of experience that the case of many very short runs is not very important in the usual statistical applications. "Naturally occurring" causes or factors which produce many very short runs do not occur very often. But there are many factors in statistical situations which may cause few very long runs. For example, in Sample I, poor tossing could cause the long runs. In Sample A, wear of the cutting tool on the lathe could cause a gradual increase in diameters of shafts and easily account for the apparent non-randomness. In example B, the cause of non-randomness could be a slip in tool setting at about the time that the eighth element in the sample was drawn. For practical purposes, it has been found that a satisfactory indication of non-randomness is the total number of runs in the sample, i.e., the total number of isolated "bunches" of H's and T's, (or of a's and b's), each "bunch" containing one or more similar letters. If we call the total number of runs U, then U has the following values in the 6 illustrative samples we have discussed:

TABLE 12.2

Sample	I	II	III	A	B	C
U	4	20	12	2	2	8

If we consider the case of many very short runs as not occurring often enough in statistical situations to be of much practical interest, we may then regard small values of U as a more important indication of non-randomness than large values of U. U is a criterion for testing randomness with respect to bunching, which is an order characteristic. It should be emphasized that other criteria

could be considered for testing randomness with respect to other characteristics.

But how do we actually decide how small (or how large) U has to be in any given case to show non-randomness beyond a "reasonable doubt"? We consider all possible permutations of the elements in a sample and find out the value of U (the total number of runs) for each permutation. Considering all of these permutations equally likely, we can then obtain a probability distribution of U. In practical situations in which graphs containing from 10 to 60 sample points similar to those shown in Figure 12.1 are constructed and runs of a's and b's are determined, one finds that it is sufficient to consider the case in which the number of a's is equal to the number of b's. Suppose there are m a's and m b's. Then the probability distribution of U (i.e., the value of $\Pr(U = u)$) is given by

$$(12.1) \qquad f(u) \begin{cases} = \dfrac{2\left(C_{\frac{u}{2}-1}^{m-1}\right)^2}{C_m^{2m}} & \text{if } u \text{ is even.} \\ = \dfrac{2\left(C_{\frac{u+1}{2}-1}^{m-1}\right)\left(C_{\frac{u+1}{2}-2}^{m-1}\right)}{C_m^{2m}} & \text{if } u \text{ is odd.} \end{cases}$$

This probability distribution is derived under the null hypothesis that the sequence is random, which means that all permutations of a's and b's are considered equally likely.

The argument for formula (12.1) involves permutation analysis and we shall not present it here. A similar formula, and not much more complicated, could be written down for m a's and m' b's where m and m' are not equal, but the case m = m' is satisfactory for present purposes.

In making a statistical significance test of the value of U obtained in a sample, we follow the usual practice and choose a probability level, say .01. Then for a given value of m (number of a's or b's) we find a critical value of u, say $u_{.01}$, such that the probability of U being less than or equal to $u_{.01}$ is at most .01. Then in a given example if U turns out to be less than the value of $u_{.01}$ applicable to that example, we say that there is a significant amount of non-randomness in the sample drawings at the 1% probability level.

Similarly, if we should be interested in the use of large values of

U as an indication of non-randomness, we would choose a critical value $u'_{.01}$ so that the probability of U being greater than or equal to $u'_{.01}$ is at most .01. Table 12.3 shows values of $u_{.01}$ and $u_{.05}$ (significantly small values of U at the 1% and 5% probability levels, respectively), and values of $u'_{.01}$ and $u'_{.05}$ (significantly large values of U at the 1% and 5% probability levels, respectively) for values of m running from 5 to 30.

TABLE 12.3

Tables of Critical Values of U

m (= no. a's = no. b's)	Significantly small critical values of U		Significantly large critical values of U	
	$u_{.05}$	$u_{.01}$	$u'_{.05}$	$u'_{.01}$
5	3	2	8	9
6	3	2	10	11
7	4	3	11	12
8	5	4	12	13
9	6	4	13	15
10	6	5	15	16
11	7	6	16	17
12	8	6	17	19
13	9	7	18	20
14	10	8	19	21
15	11	9	20	22
16	11	10	22	23
17	12	10	23	25
18	13	11	24	26
19	14	12	25	27
20	15	13	26	28
21	16	14	27	29
22	17	14	28	31
23	17	15	30	32
24	18	16	31	33
25	19	17	32	34
26	20	18	33	35
27	21	19	34	36
28	22	19	35	38
29	23	20	36	39
30	24	21	37	40

(Reproduced by courtesy of C. Eisenhart and Freda S. Swed)

As an illustrative example, we may ask whether the value of U in Sample A (see Table 12.2) is significantly low. The value of U here is 2, and m = 7. But the critical value of U at the 1% probability level is 3.

Hence we state that the value of U in Sample A is significantly low at that probability level, indicating a significant degree of non-randomness in the sample. Similarly in the case of Sample B. In the case of Sample I (considering H's as a's and T's as b's) we have U = 4, and m = 10. The critical value at the 1% probability level is $u_{.01} = 5$, thus indicating that there is a significant amount of non-randomness of H's and T's in Sample I. In the case of Sample II, the number of runs is 20 and the critical number at the 1% probability level is (m = 10) $u'_{.01} = 16$, thus showing that non-randomness in Sample II was indicated by too many runs.

12.3 Quality Control Charts.

In the discussion of runs we talked about testing non-randomness with respect to bunching of a's and b's (i.e., bunching of values above the median line and below the median line). But in many cases we find it useful to ask whether the sample is behaving in a random fashion with respect to "excessively high" values or "excessively low" values of the measurements. For example, in sampling articles coming off a production line, one may be interested in breaking strength, weight, or a critical length measurement of each article. One of the most immediate and foolproof indications that something is going wrong with the manufacturing process with respect to one or more of these important characteristics is the appearance of "excessively high" or "excessively low" values of the measurements in the sample.

So the question that arises is: What does one mean by an "excessively high" or an "excessively low" value of a measurement? How can we make it definite and unambiguous when such a value has occurred so that one can be reasonably sure that something has crept into the manufacturing process to cause such a value?

A practical procedure which is very widely used in industry in connection with mass production operations has been developed for answering these questions. The procedure is carried out graphically on what are called quality control charts.

The simplest way to show what a quality control chart is, how it is established for a given mass production operation, and how it is used, is to consider an example.

Example: In a certain plant rheostat knobs are mass produced by a plastic molding process. Each knob contains a metal insert. A certain dimension is

critical in the fitting of this knob into the assembly for which the knob is intended. This critical dimension is affected by slight variations of the size of the molded-in metal part and in the molding operation.

In establishing a quality control chart for this critical dimension, a sample of 5 knobs was taken every hour from the molding machine and the critical dimension measured on each of the 5 knobs in the set. The mean of this sample of 5 measurements was taken. This procedure was repeated for 25 successive working hours and the data in Table 12.4 were obtained (Grant data).

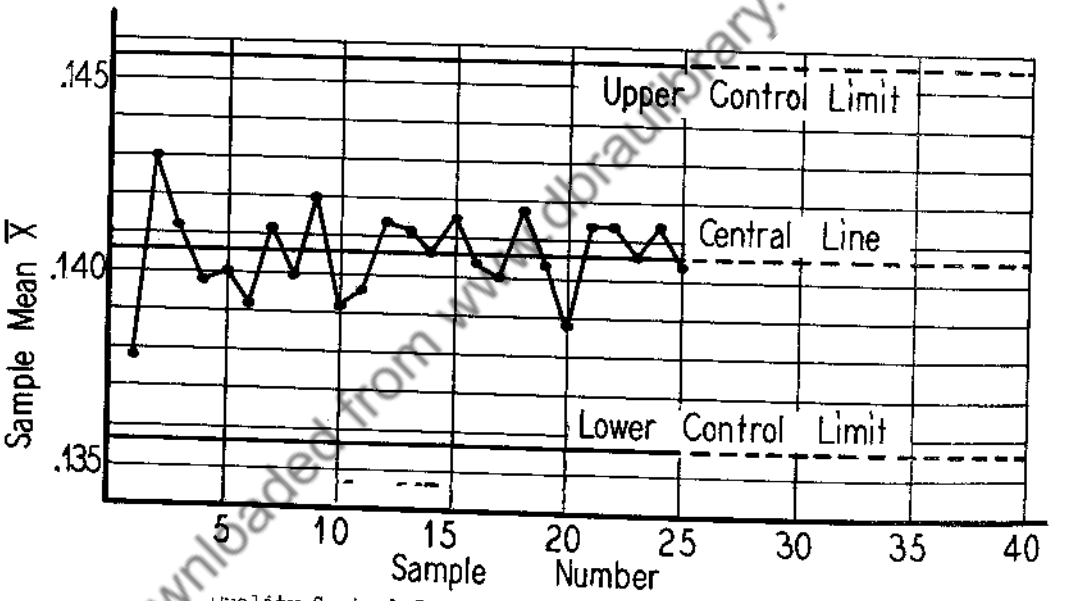
TABLE 12.4

Sample Measurements of Rheostat Knobs

Sample number	Measurements on each knob in a set of 5 picked for a given hour (measurements in thousandths of an inch)					Mean $\bar{X}$	Range R
1	140	143	137	134	135	137.8	9
2	138	143	143	145	146	143.0	8
3	139	133	147	148	139	141.2	15
4	143	141	137	138	140	139.8	6
5	142	142	145	135	136	140.0	10
6	136	144	143	136	137	139.2	8
7	142	147	137	142	138	141.2	10
8	143	137	145	137	138	140.0	8
9	141	142	147	140	140	142.0	7
10	142	137	145	140	132	139.2	13
11	137	147	142	137	135	139.6	12
12	137	146	142	142	140	141.4	9
13	142	142	139	141	142	141.2	3
14	137	145	144	137	140	140.6	8
15	144	142	143	135	144	141.6	9
16	140	132	144	145	141	140.4	13
17	137	137	142	143	141	140.0	6
18	137	142	142	145	143	141.8	8
19	142	142	143	140	135	140.4	8
20	136	142	140	139	137	138.8	6
21	142	144	140	138	143	141.4	6
22	139	146	143	140	139	141.4	7
23	140	145	142	139	137	140.6	8
24	134	147	143	141	142	141.4	13
25	138	145	141	137	141	140.4	8
						$\bar{\bar{X}}$=140.6	$\bar{R}$=8.7

In Table 12.4 it will be seen that the mean and range of the measurements for the critical dimension for each sample of 5 are given in the last two columns.

The value of $\overline{\overline{X}}$, the mean of the sample means, and the value of $\overline{R}$, the mean of the sample ranges, are shown at the bottom of Table 12.4. Now suppose we plot each mean against the corresponding sample number in the order in which it was drawn. We get the 25 points as plotted in Figure 12.2. In actual practice, each sample mean is plotted on the graph as soon as it is obtained, and straight lines connecting successive points are drawn; this is done for visual convenience in following the means from sample to sample (i.e., from hour to hour in this case).



Quality Control Chart for Data in Table 12.4

Figure 12.2

We now draw the following three horizontal lines on the graph:

(1) The central line through the mean $\overline{\overline{X}}$ of all 25 sample means, i.e., through 140.6.

(2) The upper control limit through the value $\overline{\overline{X}} + 3(.193\overline{R})$, i.e., through the value $140.6 + 3(.193)(8.7) = 140.6 + 5.0 = 145.6$.

(3) The lower control limit through the value $\overline{\overline{X}} - 3(.193\overline{R})$, i.e., through $140.6 - 3(.193)(8.7) = 140.6 - 5.0 = 135.6$.

These three lines are drawn as solid lines to Sample No. 25 and dotted from there on. The graphical result of these operations is called a quality control chart for means. Note that if we were to push all these 25 points horizontally to the left so as to pile them up against the vertical axis, we would simply have a dot diagram of 25 points with mean 140.6 and the two control limits 140.6 ± 5.0. What we have done in constructing the quality control chart is essentially this: We have taken the 25 successive sample means as a sort of temporary working standard of the amount of variability in the critical dimension. If the variation of the critical dimension were "purely random" from one rheostat knob to another, then one could consider that he had an indefinitely large population in which this dimension would follow some unknown probability distribution (probably not violently different from a normal distribution). Then means of samples of 5 from this population would tend to be much more nearly normally distributed than individual measurements (i.e., samples of 1). The larger the sample drawn each hour the more nearly normal will be the distribution of sample means if the critical dimension varies in a purely random way. In practice, however, it has been found that samples of 5 are satisfactory.

If we knew the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the theoretical sampling distribution of means of samples of 5 from the population of knobs, then we could conclude, from the normal probability table (Table 8.1), that about 99.74% of means of 5 would lie between $\mu_{\bar{X}} \pm 3\sigma_{\bar{X}}$. But we do not know $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$. We therefore use the mean $\bar{\bar{X}}$ of the 25 sample means of 5 as an estimate of $\mu_{\bar{X}}$ and the quantity $.193\bar{R}$ as an estimate of $\sigma_{\bar{X}}$, where $\bar{R}$ is the average of the 25 sample ranges. We can then state that, if the critical dimension continues to vary in a "purely random" manner from knob to knob as it did during the period covered by the 25 samples, then about 99% (99.74% to be more precise) of means of samples of 5 will fall between $\bar{\bar{X}} \pm 3(.193\bar{R})$, i.e., between the dotted control limits, in future sampling.

The justification of the value $.193\bar{R}$ as an estimate of $\sigma_{\bar{X}}$, on the assumption that the critical dimension varies in a "purely random" way, is beyond the scope of this course. We could have estimated $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{5}}$ by pooling the variances of the 25 samples just as we pooled two sample variances to obtain (10.35) as the estimate of σ in the two-sample case. In practice, however, this involves a great deal of computation. Estimating $\sigma_{\bar{X}}$ by using $.193\bar{R}$ is quite

satisfactory when as many as 25 samples of 5 measurements each are involved. The factor .193 is used only when samples of 5 measurements are used. The factors for samples of sizes 3, 4, 6, 7, 8, 9, 10 are .341, .243, .161, .140, .124, .112, .103, respectively.

In practice we continue to take a sample of 5 knobs every hour beyond the 25th hour to see whether the sample means continue to fall within the dotted control lines $140.6 \pm 3(.193\bar{R})$. If a point falls outside the region bounded by the dotted lines, it is considered that something has gone wrong with the process and an examination of the process is made to see what the trouble might be. Of course, even when nothing has gone wrong, there is a probability of less than 1% of a point falling outside the control limits. But the engineer takes this amount of risk of looking for trouble when it is not present, just to make sure that a cause of trouble which might exist does not go undetected. In other words, when a point falls outside he will bet more on process trouble than on pure chance as the cause of the point falling outside.

If, as the sampling proceeds, hour by hour one finds the means of samples of 5 jumping about in a haphazard way between the control limits, we say that the manufacturing process is under statistical control with respect to the critical dimension under consideration. What usually happens in practice when a quality control chart procedure is introduced is this: After establishing control limits on the basis of 25 or more samples, and after searching for and eliminating causes for process trouble every time a point falls outside the control limits, one soon finds that the variability from mean to mean becomes smaller than it was for the initial set of 25 samples. One can then take a new set of samples and establish a new central line and new upper and lower control limits. The control limits, in practice, will usually be closer together than the original ones. After a few stages of this kind one arrives at a state of statistical control involving control limits which are about as close together as one can hope to make them without revolutionary changes in the manufacturing process.

If you look at Figure 12.2 you will see that the mean of each of the 25 samples drawn is well within the control limits; this fact indicates that no causes of exceptional variation have crept into the manufacturing process. The tolerance limits specified by the engineering designer for this problem were 140.6 ± 5.0 (in units of .001 in.); this means that any rheostat knob having its critical dimension outside these limits will be rejected. Looking at Table 12.4 you will see that there are 19 knobs (about 14%) among the 125 in that table that

would be rejected upon inspection. This is a high figure for rejections. In good manufacturing practice, the percentage of rejections should not be more than 1% to 5%. Since the present manufacturing process seems to be nicely under statistical control, the only way to get the variation in critical dimension down to the point where the rejects would not be more than 1% to 5% would probably be to make a radical change in manufacturing the metal parts and in the molding process -- i.e., to "tighten" up on them so there would not be so much variation in either of them.

Quality control charts are very widely used in industry. They provide a very simple and effective way to see graphically what goes on in successive sampling so as to be in a position to know when a drawing yields an excessively large or excessively small value, and to take appropriate action when it occurs. If one should want to examine the data in a control chart a little more closely he could make a run test on it.

Control charts and runs provide simple and practical methods of checking randomness in samples from any indefinitely large population. This randomness is necessary in order for sampling theory to be applicable. The usual practice in statistics is to assume that the randomness is good enough without actually checking it. Some of the serious pitfalls in statistics occur in sampling situations in which the requirement of randomness is not satisfactorily fulfilled.

Quality control charts can be constructed for sample statistics other than means, e.g., sums, ranges, etc. In fact, control charts for means and ranges are usually run parallel to each other on the samples and on the same sheet of paper, the range chart being placed directly below the mean chart.

If one has enough information, theoretical or experimental, about a population from which one is sampling, it is possible to set up the control limits on a quality control chart completely in advance of any sampling. For example, suppose a person proposes to make a critical study of the performance of a die thrown under a given set of conditions (the die may be shaken in a leather cup, thrown on a card table, etc.). Suppose the study is made by considering means of the number of dots obtained in successive samples of 10 throws. If the hypothesis of perfect performance is set up, then the central line and the two control limits can be established without any performance data from the die. In fact, the central line would pass through the mean 3.5

and the control lines would pass through $3.5 \pm 3\sqrt{\frac{35}{120}} = 3.5 \pm 1.62$.

Exercise 12.

1. Consider the first 3 columns of data in Table 2.1 as a sample of 45 individual measurements drawn in the order 1.47, 1.62, ..., 1.57. Make a run test of these measurements for randomness at the 1% probability level. (Test for too many runs as well as too few.)

2. You are supposed to have the original data on at least one of the following problems in Exercise 2.2: Nos. 8, 9, 10, 11, 12, 13, 14. Make a run test on your data at the 1% probability level.

3. Do the best you can to write down what you would consider to be a random sequence of 31 numbers between 500 and 1000. Make a run test on this sequence of numbers at the 1% probability level.

4. The number of divorces per 1000 persons in the U. S. for each of the years 1920 to 1940 was: 1.6, 1.5, 1.4, 1.5, 1.5, 1.5, 1.5, 1.6, 1.6, 1.7, 1.6, 1.5, 1.3, 1.3, 1.6, 1.7, 1.8, 1.9, 1.9, 1.9, 2.0. Test this sequence for randomness by the method of runs at the 1% probability level. Interpret your results.

5. Make a quality control chart for the data in Example No. 7 of Exercise 2.2, using means of samples of 5. The first sample consists of the first row of numbers across the page, the second sample consists of the second row, etc.

6. Using the probability distribution in Example No. 5 of Exercise 9.3, set up the central line and control limits for a quality control chart for the total number of aces in successive sets of 10 hands of bridge, which you would use in studying the thoroughness of shuffling as measured by number of aces dealt.

7. The following means and ranges of muzzle velocities (in ft./sec.) were obtained for samples of five from 25 consecutive lots of ammunition (Simon data):

Lot	Mean	Range
1	1710	42
2	1711	40
3	1713	39
4	1718	26
5	1735	10
6	1739	25
7	1723	14
8	1741	15
9	1738	11
10	1725	31
11	1731	25
12	1721	19
13	1719	43
14	1735	39
15	1741	17
16	1783	51
17	1777	9
18	1794	15
19	1773	37
20	1789	54
21	1798	15
22	1789	29
23	1788	39
24	1799	30
25	1807	44
	$\bar{\bar{X}}$=1751.88	$\bar{R}$=28.76

Construct a quality control chart for the means in this problem.

8. Suppose you have 16 dice of a certain kind (all look exactly alike). Construct the central line and control limits for a quality control chart you would use in studying whether the total number of dots appearing on these dice is behaving as one would expect for unbiased dice, assuming that all 16 dice are thrown successively. The following set of data shows the total number of dots which appeared in 100 throws of 16 5-cent dice. Try your control chart (with its central and control limits) on these data:

61	57	56	48	63	49	56	57	66	61
54	62	57	63	59	56	54	60	56	62
54	62	60	50	52	62	60	47	63	32
52	53	65	57	56	61	57	63	67	47
63	51	55	53	57	58	55	54	55	53
53	67	59	59	67	47	65	55	59	53
59	56	54	65	57	50	68	53	59	63
49	51	67	69	56	61	52	56	51	63
51	54	58	52	55	65	54	71	52	56
52	54	55	71	57	58	57	69	53	45

CHAPTER 13. ANALYSIS OF PAIRS OF MEASUREMENTS

13.1 Introductory Comments.

In all the preceding chapters we have dealt with elementary statistical analysis of samples involving only a single measurement, and with elementary probability analysis of one chance quantity X. Many statistical problems arise in which the sample consists of pairs (triplets, or a higher number) of measurements.

In some cases the relationship between the two measurements may be very strong and the statistical analysis required may be very simple. For example, a chemist interested in finding a quick method of determining alpha-resin content of hops first studies the relationship between colorimeter readings for certain standard flasks containing various concentrations of alpha-resin. He performs an experiment and obtains the following pairs of measurements (Bullis and Alderton data):

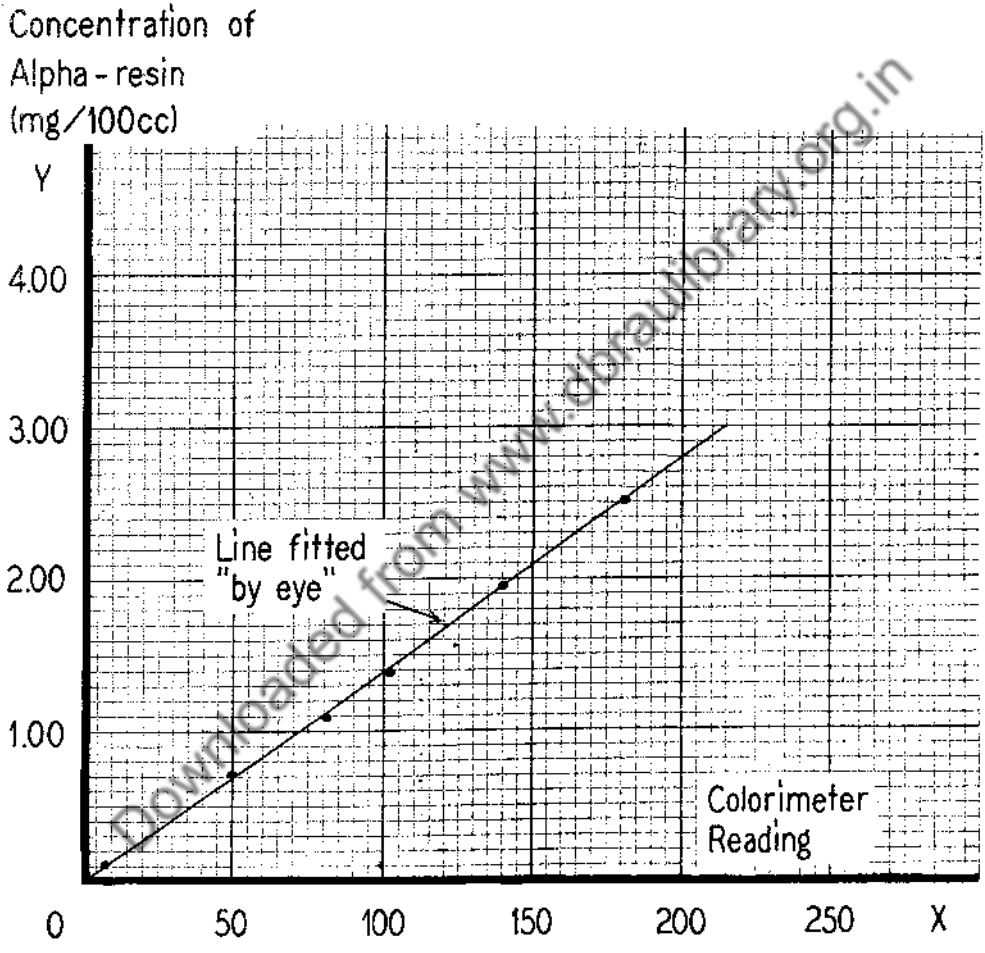
TABLE 13.1

Colorimeter reading X	Concentration of alpha-resin (milligrams per 100 cc.) Y
8	.12
50	.71
81	1.09
102	1.38
140	1.95
181	2.50

The simplest way to analyze these data is to plot them on graph paper as six points, the X and Y coordinates of the six observed points being (8, .12), (50, .71), (81, 1.09), (102, 1.38), (140, 1.95) and (181, 2.50), and to note that the points lie almost on a straight line. In fact, if a straight line were fitted "by eye", it would be quite accurate. The result of plotting the points and the straight line is shown in Figure 13.1.

From Figure 13.1 we can make a reasonably accurate and quick estimate of the concentration of alpha-resin from a given colorimeter reading for a flask

containing an unknown concentration of alpha-resin. For instance, if a standard flask containing an unknown concentration of alpha-resin has a colorimeter reading of 125, we would estimate the alpha-resin concentration to be 1.75 mg/100cc. as you will see from Figure 13.1. A colorimeter reading is easy and quick to make, but a direct determination of the alpha-resin content in the unknown concentration would be very time-consuming.



Graph of the Data in Table 13.1 and a Straight Line Fitted "by eye"

Figure 13.1

The line determined from the observed points in Figure 13.1 for estimating concentration of alpha-resin (Y) from colorimeter reading (X) may be referred to as a regression line of Y on X. We have used the line in Figure 13.1 as a graphical regression line of alpha-resin concentration (Y) on colorimeter reading (X). Because the fit is so close, this particular line can also be used as a graphical regression line of colorimeter reading (X) on alpha-resin concentration (Y). We would use it in the latter way if we wished to estimate the colorimeter reading for a known alpha-resin concentration. In case the observed points do not lie closely along a straight line, there will be two regression lines as will be discussed in section 13.22 (see Figure 13.3).

Many other simple examples of this type could be given, in which the plotted points would fall almost along a straight line or some kind of a smooth curve. Satisfactory elementary statistical analysis for such examples consists of plotting the points and drawing the line or curve by eye, and using the resulting line or curve for estimating one of the values of a pair of measurements from the other measurement. Fitting "by eye" can be done more effectively by using certain tools to aid the eye. In fitting straight lines, a piece of transparent celluloid or plastic with a fine black line is best. A piece of fine black thread streched from hand to hand is good. The edge of a transparent triangle is satisfactory. In fitting curved lines, drafting curves or splines are effective.

In problems where the plotted points do not fall nearly along a line or a curve, the problem of statistical analysis of the relationship between the two variables is usually more complicated. We can, of course, still try to draw some kind of a line or curve through the points "by eye" so as to fit them as "best" we can. The difficulty is that if we should try to repeat this operation on the same set of points, or if several people were to try to do it, they might get quite different results because of the rather wide scatter of the points. Hence some procedure for fitting the line or curve is needed which will result in greater consistency from repetition to repetition than fitting "by eye". Not only do we need a procedure to fit a line or a curve objectively, but we need some way of measuring the amount of scatter of the points about the curve. The main purpose of this chapter is to give methods for handling these problems. The most widely used method for objectively fitting lines and curves is the method of least squares. We shall consider this method for fitting straight lines and

curves.

Exercise 13.1.

1. The following data were obtained in a certain series of experiments on the relationship between concentration of penicillin solution in units/ml.(X) and mean circle diameter of the zone of inhibition in mm.(Y):

X	1	2	4	8	16	32
Y	15.87	17.78	19.52	21.35	23.13	24.77

Graph these six points and construct a regression line "by eye". From this regression line, estimate the mean circle diameter of zone of inhibition for 25 units/ml. of penicillin solution. How many units of the penicillin solution would be required to produce a mean circle diameter of 20 mm.?

2. Specimens of several brands of mayonnaise were analyzed for fat content by a rapid method and by a method of the Association of Official Agricultural Chemists (A.O.A.C.). Denoting the result of the Rapid Method by X and that by the A.O.A.C. Method by Y, the following data were obtained (by Kaufman):

X	Y
80.5	79.3
30.3	30.4
25.2	26.0
77.4	77.9
48.1	47.5
35.7	35.3
18.6	18.7

Graph these seven points and fit a regression line "by eye". If the Rapid Method showed a fat content of 50% for a certain specimen, what percent content would you estimate the A.O.A.C. Method to show for the same specimen?

3. The following data were obtained in an experiment to study the relationship between the amount of beta-erythroidine in mg/l. (X) in an aqueous solution and colorimeter reading of turbidity (Y) of the solution:

X	40	50	60	70	80	90
Y	69	175	272	335	390	415

Graph these six points and construct a regression curve "by eye". Estimate the concentration of beta-erythroidine in a solution if the colorimeter reading is 370. What concentration of beta-erythroidine is required to give a colorimeter reading of 200?

13.2 The Method of Least Squares for Fitting Straight Lines.

13.21 An example.

We shall first consider the following

Example: In a class of 17 students in a sophomore mathematics course at Princeton, the following Term Scores (X) and Final Examination Scores (Y) were obtained:

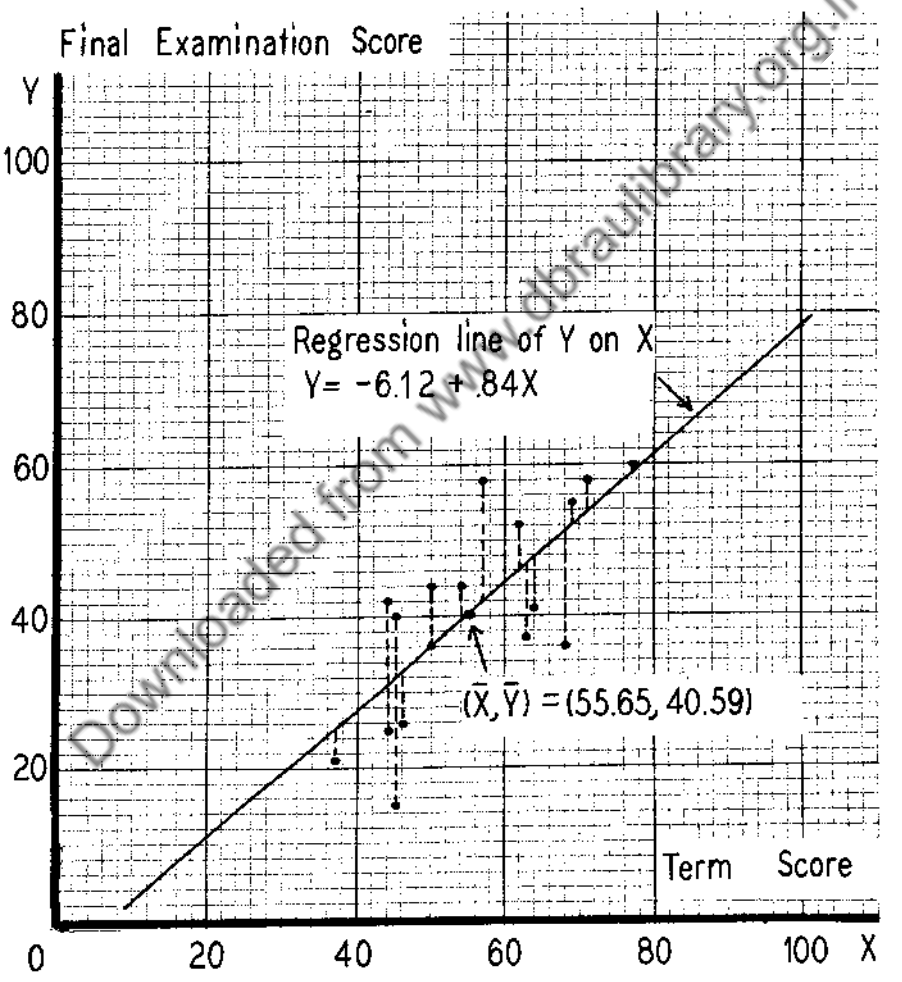
TABLE 13.2

Term Score X	Final Examination Score Y
50	44
44	25
54	44
46	26
64	41
37	21
77	60
62	52
45	15
69	55
44	42
57	58
63	37
68	36
45	40
71	58
50	36

We are to fit a straight line to these data which can be used for making an estimate of the Final Examination Score (Y) for a student for whom only the Term

Score (X) is available.

The first thing to do is to plot the data as 17 dots in the XY plane, as shown in Figure 13.2. The resulting plot is called a scatter diagram. Note that if the 17 dots were moved vertically downward to the X axis we would have a dot frequency diagram of X (the Term Score). Similarly, if the dots were moved horizontally left to the Y axis, we would have a dot frequency diagram of Y (the Final Examination Score). The distribution of the X scores has its own mean and variance. Similarly for the distribution of Y scores.



Scatter Diagram for Data in Table 13.2 and
Regression Line of Final Examination Score (Y) on Term Score (X)

Figure 13.2

Usually, some useful and interesting information can be obtained by inspection from a scatter diagram. For example, suppose 60 is the passing score on the Term Score and 50 is the passing score on the Final Examination. By drawing a vertical line through X = 60 and a horizontal line through Y = 50, we cut the plotted points into four sets, falling into four quadrants. The four points in the upper right quadrant correspond to 4 students who passed on both Term and Examination. The 9 points falling into the lower left quadrant failed on both. One student passed on the Examination and failed on the Term, while 3 passed on the Term and failed on the Examination.

Now if we want to study the relationship between the Term Score and Examination Score, the question is this: How do we construct a line through the scattered points which could be used for estimating Y from a given X? If we were to try to do it several times "by eye" the result might not be very consistent from trial to trial. Or if several persons were to try to fit by such a method the resulting line might not be very consistent from person to person. This variability would certainly be greater for the scatter diagram of Figure 13.2 than for that of Figure 13.1. We need an objective way to fit a line.

Now, the equation of any straight line may be written in the form

$$Y = a + bX \tag{13.1}$$

where a and b are constants. Every straight line that can be drawn on Figure 13.2 can be obtained from (13.1) by substituting some numerical values for a and b. An objectively fitted regression line is one of these lines. The problem of objectively fitting the line (13.1) to the data therefore amounts to objectively determining values of a and b. The principle we shall use to determine a and b is this: Use formula (13.1) for estimating the value of Y from the X value of each of the 17 points. Then square the difference between the actual value of Y and the estimated value of Y for each point and sum these squares. Choose values of a and b so as to minimize this sum of squares of differences or residuals. If we think of these differences as errors in estimating values of Y from values of X, then what we are really doing is to choose a and b so as to minimize the sum of squares of the errors.

The values of a and b determined in this manner will be called the least squares values of a and b.

Now for the details. The actual value of X for the first point (first entry in Table 13.2) is 50. The actual value of Y is 44. The estimated value of Y is

$$a + 50\,b\ ,$$

obtained by substituting $X = 50$ in formula (13.1). The squared difference between the actual and estimated values of Y is

$$(44 - a - 50\,b)^2\ .$$

Doing a similar operation on all of the points in the scatter diagram (i.e., pairs of scores in Table 13.2), we have for the sum of the 17 squared differences:

$$(13.2)\quad S = (44-a-50b)^2 + (25-a-44b)^2 + (44-a-54b)^2 + \ldots + (36-a-50b)^2\ .$$

Now we must choose the values of a and b so as to make the value of S as small as possible. Such values are obtained by setting the following two derivatives equal to zero and solving simultaneously for a and b:

$$(13.3)\quad \frac{\partial S}{\partial a} = 0$$

$$\frac{\partial S}{\partial b} = 0\ .$$

Carrying out the two differentiations with respect to a and b, we find

$$(13.4)\quad \begin{aligned} &-2(44-a-50b) - 2(25-a-44b) - \ldots - 2(36-a-50b) = 0 \\ &2(50)(44-a-50b) - 2(44)(25-a-44b) - \ldots - 2(50)(36-a-50b) = 0\ . \end{aligned}$$

Dividing each fraction by two and performing the arithmetic we find that these two equations simplify to

$$(13.5)\quad \begin{aligned} &-690 + 17a + 946b = 0 \\ &40{,}238 + 946a + 54{,}836b = 0\ . \end{aligned}$$

Let $\hat{a}$ and $\hat{b}$ be the values of a and b which satisfy these equations. Multiplying the first equation by $\frac{946}{17} = 55.6471$ and subtracting from the second we find

$$\hat{b} = .8394 .$$

Substituting this value for b in the first of the equations in (13.5) we find

$$\hat{a} = -6.1215 .$$

The two equations (13.5) may also be solved by the use of determinants as follows:

$$\hat{b} = \frac{\begin{vmatrix} 17 & 690 \\ 946 & 40,238 \end{vmatrix}}{\begin{vmatrix} 17 & 946 \\ 946 & 54,836 \end{vmatrix}} = \frac{(40,238)(17)-(690)(946)}{(54,836)(17)-(946)^2}$$

$$= \frac{31,306}{37,296} = .8394$$

and

$$\hat{a} = \frac{\begin{vmatrix} 690 & 946 \\ 40,238 & 54,836 \end{vmatrix}}{\begin{vmatrix} 17 & 946 \\ 946 & 54,836 \end{vmatrix}} = \frac{(54,836)(690)-(946)(40,238)}{37,296}$$

$$= \frac{-228,308}{37,296} = -6.1215 .$$

Hence the equation of the least squares line for estimating Y from X is

(13.6) $$Y = -6.1215 + .8394 X .$$

This line is called the regression line of Y on X. Rounding the coefficients off to two decimal places (which provides sufficient accuracy for the present problem) the graph of the line is shown in Figure 13.2. It is an objectively obtained line which we would use for estimating the Y-score (Final Examination Score) for a student of the class whose X score (Term Score) is known. For instance, suppose a student in the class having a Term Score of 70 is unable to take the Final Examination. What estimate would we make for his Final Examination Score if he had not taken the Examination? Substituting in (13.6) we would have

$$Y = -6.12 + .84(70)$$

$$= 52.68$$

or, rounding off, we would estimate his Final Examination Score to be 53.

13.22 The general case.

It will be seen by examining equations (13.4) and (13.5) that the constants and coefficients occurring in (13.5) are obtained by performing summation operations on X, Y, X^2, and XY. More specifically, if we let (X_1, Y_1), (X_2, Y_2), ..., (X_{17}, Y_{17}) denote the 17 pairs of measurements in our sample, and if we are interested in estimating values of Y from values of X, then equation (13.5) may be written as

$$-\sum_{j=1}^{17} Y_j + 17a + b\sum_{j=1}^{17} X_j = 0$$

(13.5a)

$$-\sum_{j=1}^{17} X_j Y_j + a\sum_{j=1}^{17} X_j + b\sum_{j=1}^{17} X_j^2 = 0 .$$

Or writing these with our abbreviated notation introduced in Chapter 3 (and noting that the coefficient 17 in the first equation may be written as $\sum_{j=1}^{17}(1)$ or S(1) in this example) we have

$$\begin{aligned} -S(Y) + a\,S(1) + b\,S(X) &= 0 \\ -S(XY) + a\,S(X) + b\,S(X^2) &= 0 . \end{aligned} \qquad (13.7)$$

The values of the coefficients may be conveniently found by setting up a table such as Table 13.3. Unless one has a computing machine it pays to use a simpler computing scheme than that involved in Table 13.3. Simplified computing schemes require special discussion and illustration, and will be considered in Section 13.3. The column of values of Y^2 is given for later use.

If we consider a sample of n pairs of measurements (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), and fit the line $Y = a + bX$ by the method of least squares, we will end up with equations of form (13.7) for determining a and b. Consequently we can consider equations (13.7) as general enough to hold for any sample of n pairs of measurements. In the general case $S(1) = n$.

In fitting the line $Y = a + bX$ so as to be able to estimate values of Y from values of X, we refer to X as the *independent variable* or *predictor* and Y as the *dependent variable* or *predictand*.

Considering (13.7) as general equations, let us solve them for a and

b. Let us call the values of a and b which are found $\hat{a}$ and $\hat{b}$. Multiplying the first equation in (13.7) by $\frac{S(X)}{S(1)}$, and subtracting from the second equation, and remembering that $S(1) = n$, $S(X) = n\bar{X}$, and $S(Y) = n\bar{Y}$, we have

$$\hat{b} = \frac{S(XY) - \frac{1}{n}S(X)S(Y)}{S(X^2) - \frac{1}{n}[S(X)]^2} = \frac{S(XY) - n\bar{X}\bar{Y}}{S(X^2) - n\bar{X}^2}. \tag{13.8}$$

TABLE 13.3

Table for Computing Values of Coefficients for Equations (13.5)

Pair No.	X	Y	X^2	Y^2	XY
1	50	44	2500	1936	2200
2	44	25	1936	625	1100
3	54	44	2916	1936	2376
4	46	26	2116	676	1196
5	64	41	4096	1681	2624
6	37	21	1369	441	777
7	77	60	5929	3600	4620
8	62	52	3844	2704	3224
9	45	15	2025	225	675
10	69	55	4761	3025	3795
11	44	42	1936	1764	1848
12	57	58	3249	3364	3306
13	63	37	3969	1369	2331
14	68	36	4624	1296	2448
15	45	40	2025	1600	1800
16	71	58	5041	3364	4118
17	50	36	2500	1296	1800
Total	946	690	54,836	30,902	40,238

$S(1) = 17$ $S(X^2) = 54,836$
$S(X) = 946$ $S(Y^2) = 30,902$
$S(Y) = 690$ $S(XY) = 40,238$

The quantity $\frac{1}{n-1}[S(XY) - n\bar{X}\bar{Y}]$ is called the covariance between X and Y, and will be abbreviated as cov(X,Y). It will be remembered from Section 3.12 that the quantity $\frac{1}{n-1}[S(X^2) - n\bar{X}^2]$ is s_X^2, the variance of X. Hence, we may write

(13.8) as follows:

$$\hat{b} = \frac{\mathrm{cov}(X,Y)}{s_X^2} . \tag{13.8a}$$

From (13.8a) it is evident that equations (13.7) can always be solved if s_X^2 is not zero; s_X^2 will be zero only if the X measurements all have the same value, in which case we would not be able to study the relationship between X and Y by least squares or any other method.

It will be seen that

$$S[(X-\overline{X})(Y-\overline{Y})] = S[XY-\overline{X}Y-X\overline{Y}+\overline{X}\overline{Y}]$$

$$= S(XY) - \overline{X}S(Y) - \overline{Y}S(X) + n\overline{X}\overline{Y} = S(XY) - n\overline{X}\overline{Y} .$$

In other words, the covariance between X and Y may also be written as

$$\frac{1}{n-1} S[(X-\overline{X})(Y-\overline{Y})] .$$

Just as $\frac{1}{(n-1)}\left[S(X^2) - \frac{1}{n}[S(X)]^2\right]$ is the most convenient formula for computing the value of s_X^2 (and similarly for s_Y^2), the formula $\frac{1}{n-1}\left[S(XY) - \frac{1}{n}[S(X)\cdot S(Y)]\right]$ is the most convenient formula for computing the value of cov(X,Y).

We may write $\hat{a}$ in terms of $\hat{b}$ as follows:

$$\hat{a} = \overline{Y} - \hat{b}\overline{X} ,$$

and substituting this value for a and $\hat{b}$ for b in the equation $Y = a + bX$, the equation of the straight line which fits the n observed points "best" in the sense of least squares when Y is used to estimate X, is

$$Y - \overline{Y} = \hat{b}(X-\overline{X}) . \tag{13.9}$$

This line is called the regression line of Y on X and is used for estimating the Y value of a pair of measurements if only the X value of the pair is known.

If we let s_Y^2 be the variance of the Y measurements in the sample, and let

$$r = \frac{S[(X-\overline{X})(Y-\overline{Y})]}{(n-1)s_X s_Y} = \frac{\mathrm{cov}(X,Y)}{s_X s_Y} , \tag{13.10}$$

then we may write

$$\hat{b} = r\frac{s_Y}{s_X} \tag{13.11}$$

$$\hat{a} = \bar{Y} - r\frac{s_Y}{s_X}\bar{X}.$$

Using these values for a and b, the equation of the regression line of Y on X may therefore be written in terms of $\bar{X}$, $\bar{Y}$, s_X, s_Y, and r as follows:

$$(Y-\bar{Y}) = r\frac{s_Y}{s_X}(X-\bar{X}) . \tag{13.12}$$

This is merely another way of writing equation (13.9).

The quantity r as defined by (13.10) is called the correlation coefficient between X and Y; its properties and uses will be discussed in Section 13.25.

The quantity $\hat{b}$ or $r\frac{s_Y}{s_X}$, which is the coefficient of X in the equation of the regression line of Y on X, is called the regression coefficient of Y on X. The constant term $\hat{a}(= \bar{Y} - \hat{b}\bar{X} = \bar{Y} - r\frac{s_Y}{s_X}\bar{X})$ is called the Y intercept. It is the value of Y at which the regression line cuts the Y axis. It will be seen from equation (13.12) that the regression line always passes through the point $(\bar{X},\bar{Y})$, i.e., the point whose X-coordinate is the mean of the X measurements and whose Y-coordinate is the mean of the Y measurements.

It is clear from considerations of symmetry that if we should need to estimate X for a pair of observations when the value of Y is known, we would determine a line so that the sum of squares of differences between the actual values of X and estimated values of X is a minimum. This leads to the regression line of X on Y,

$$(X-\bar{X}) = r\frac{s_X}{s_Y}(Y-\bar{Y}) . \tag{13.13}$$

Note that the regression coefficient of X on Y is $r\frac{s_X}{s_Y}$ and not $r\frac{s_Y}{s_X}$, i.e., it is $\frac{cov(X,Y)}{s_Y^2}$ and not $\frac{cov(X,Y)}{s_X^2}$.

For any given scatter diagram of points such as that shown in Figure 13.3,

the two regression lines will be something like those shown in Figure 13.3.

In practical statistical work only rarely, if ever, are both regression lines needed in a single problem. For when regression analysis is used, the usual situation is that we know only one of the values in a pair of measurements and want to estimate the value of the other measurement. In any given problem, we can always label the measurements in the n pairs of measurements so that the independent variable (predictor) is X and the dependent variable is Y. In doing this we would have no reason for wanting to determine the regression line of X on Y.

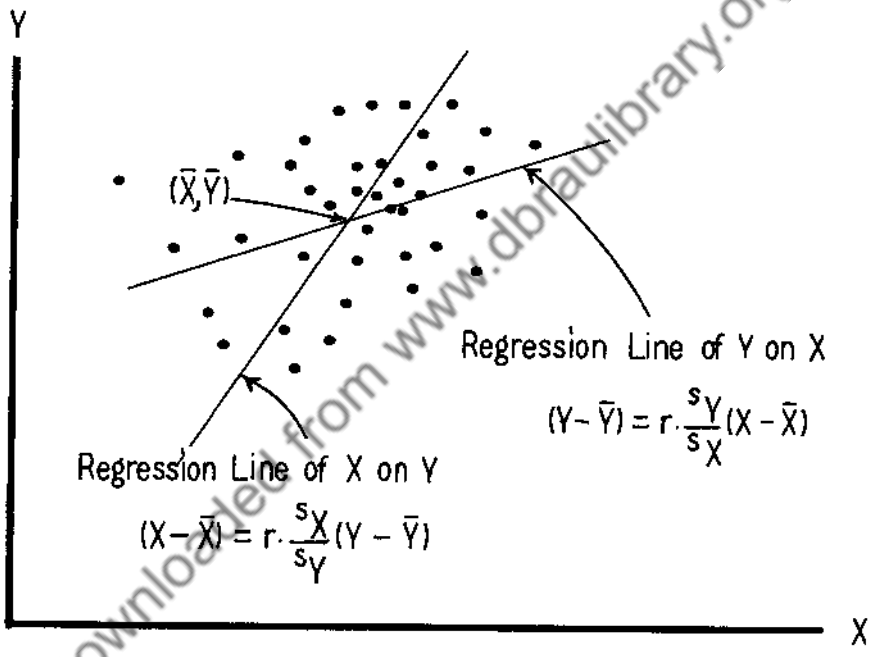


Figure Showing the Two Regression Lines for a Scatter Diagram

Figure 13.3

Returning to the example of the 17 pairs of scores, we may calculate the values of $\bar{X}$, $\bar{Y}$, $\hat{b}$, s_X, s_Y and r from Table 13.3 as follows:

$$\bar{X} = \frac{946}{17} = 55.65$$

$$\bar{Y} = \frac{690}{17} = 40.59$$

$$s_X^2 = \frac{1}{16}\left[(54,836) - \frac{1}{17}(946)^2\right] = 137.1176;\ s_X = 11.71$$

$$s_Y^2 = \frac{1}{16}\left[(30,902) - \frac{1}{17}(690)^2\right] = 181.0074;\ s_Y = 13.45$$

$$\mathrm{cov}(X,Y) = \frac{1}{16}\left[(40,238) - \frac{1}{17}(946)(690)\right] = 115.0956$$

$$\hat{b} = \frac{115.0956}{137.1176} = .8394$$

$$r = \frac{115.0956}{\sqrt{(137.1176)(181.0074)}} = .7306\ .$$

If these values of $\bar{X}$, $\bar{Y}$, and $\hat{b}$ were substituted in (13.12), we would find, as the equation of the regression line of Y on X,

(13.14) $$(Y - 40.59) = .8394(X - 55.65)\ ,$$

which is merely an alternative way of writing (13.6).

13.23 The variance of estimates of Y from X.

If we were to use the regression equation (13.6) (or its alternative form (13.14) for making an estimate of Y from each value of X in Table 13.3, we would find that these estimates differed somewhat from the actual values of Y. For instance, the estimated value of Y corresponding to X = 50, obtained by putting X = 50 in (13.6), is -6.1215 + .8394(50) = 35.85. The actual value of Y is 44. The difference or residual is (44-35.85) = +8.15. Similarly, we can find such a difference for each of the 17 pairs of measurements. While there are 17 such residuals, there are only 15 degrees of freedom; this means that if we know the values of 15 of the residuals, the remaining 2 are automatically determined. The sum of the squares of these residuals, divided by the number of degrees of freedom, is called the variance of the estimates of Y from X, or more briefly, the variance of estimate, and will be denoted by s_E^2. s_E is called the standard error of estimate.

Graphically speaking, the residuals we are talking about are the dotted vertical segments in Figure 13.2. Those lying above the regression line are positive and those below are negative. The variance of estimate s_E^2 is

simply the sum of the squares of these segments divided by 15.

We may therefore regard s_E^2 as an index of the amount of scatter of points around the regression line. The smaller the value of s_E^2, the smaller the scatter.

Frequently we have problems in which the scatter diagram is a roughly elliptical cluster of points. In such cases we find that these residuals can be reasonably well fitted by a normal distribution with mean 0 and standard deviation s_E. This means that if we were to draw a line parallel to the regression line on each side of the regression line so that the distance measured vertically between the regression line and each parallel line is s_E, then approximately 68% of the points in the scatter diagram would lie between the two parallel lines (see Figure 13.4). If the lines are placed at a distance $2s_E$ on each side of the regression line, about 95% of the points will lie between the two parallel lines. If they are placed at a distance of $3s_E$ on each side of the regression line, about 99.7% of the points will lie between the two parallel lines.

What we need now is a simpler way of calculating s_E^2 than evaluating, one-by-one, the differences between the estimated values of Y and the actual values of Y. To do this, let us consider the general form of the regression equation (13.9) and any given pair of measurements in the sample, say (X_j, Y_j). The estimate of Y_j from (13.9) is found by putting $X = X_j$ and solving for Y. We get

$$\underline{\text{Estimate}} \text{ of } Y_j = \overline{Y} + \hat{b}(X_j - \overline{X}) .$$

The <u>actual value</u> of Y_j is Y_j. The <u>difference</u> E_j between actual value and estimate is

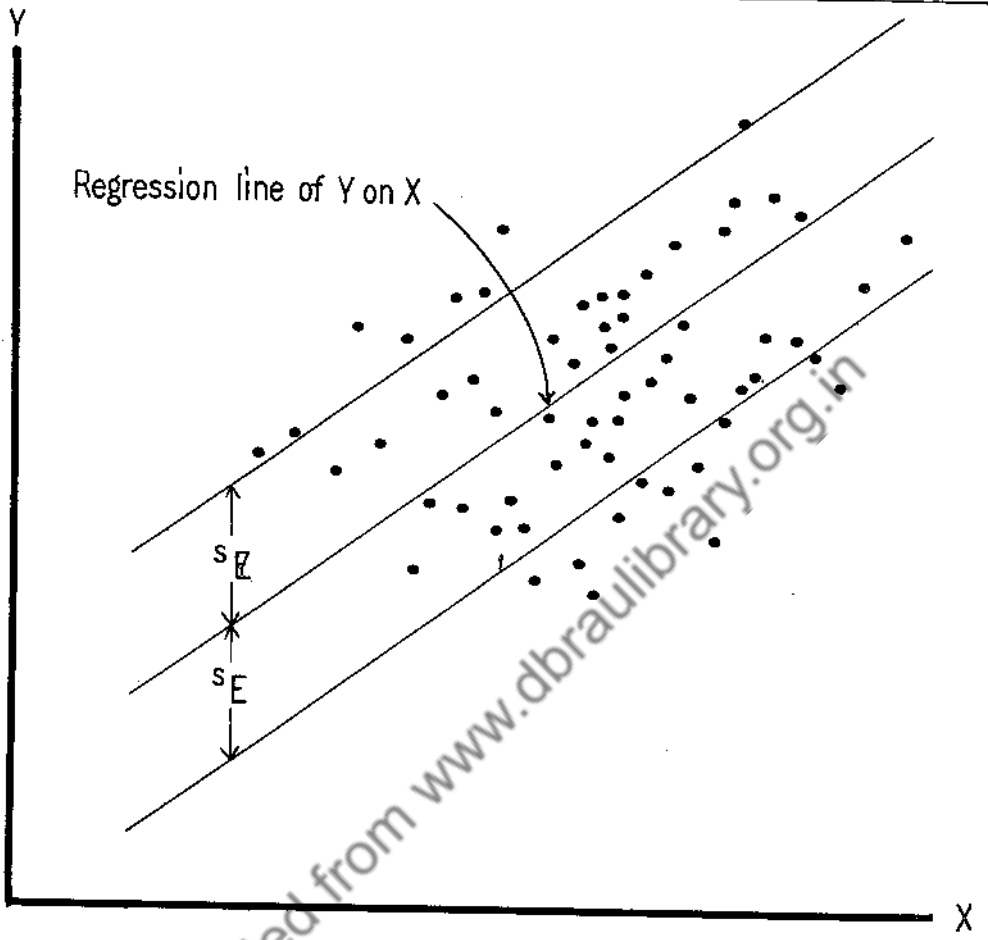
$$E_j = [(Y_j - \overline{Y}) - \hat{b}(X_j - \overline{X})] . \tag{13.15}$$

By definition, s_E^2 is the square of this difference summed over all pairs of measurements in the sample and divided by n-2, i.e.,

$$s_E^2 = \frac{1}{n-2} \sum_{j=1}^{n} [(Y_j - \overline{Y}) - \hat{b}(X_j - \overline{X})]^2 . \tag{13.16}$$

Squaring the quantity in [ ], we have

$$s_E^2 = \frac{1}{n-2} \sum_{j=1}^{n} [(Y_j - \overline{Y})^2 - 2\hat{b}(X_j - \overline{X})(Y_j - \overline{Y}) + \hat{b}^2(X_j - \overline{X})^2] . \tag{13.17}$$



Graphical Representation of Standard Error of Estimate s_E

Figure 13.4

Now

$$\sum_{j=1}^{n} (Y_j - \bar{Y})^2 = (n-1) \cdot s_Y^2 , \quad \sum_{j=1}^{n} (X_j - \bar{X})^2 = (n-1) \cdot s_X^2$$

and

$$\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y}) = S[(X-\bar{X})(Y-\bar{Y})] = (n-1) \operatorname{cov}(X,Y)$$

as will be seen from (13.10). Hence (13.17) may be simplified and expressed as follows:

$$s_E^2 = \frac{n-1}{n-2}\left[s_Y^2 - \hat{b}^2 s_X^2\right] . \tag{13.18}$$

By making use of (13.11) we may express s_E^2 alternatively as follows:

$$s_E^2 = \frac{n-1}{n-2} s_Y^2 (1 - r^2) . \tag{13.18a}$$

Either of the formulas (13.18) or (13.18a) enables us to calculate the variance of estimate without using all the individual differences between estimated and actual values of Y.

It should be noted from (13.18) that when $\hat{b} = 0$ (or from (13.18a) that when $r = 0$) we have $s_Y^2 = s_E^2$, which means that under these conditions the use of X is of no value in estimating values of Y.

Returning to the example of 17 pairs of measurements, we have for the variance of estimate:

$$s_E^2 = \frac{16}{15}(181.0074)(1-(.7306)^2 = 90.0159$$

and for the standard error of estimate

$$s_E = 9.488 .$$

13.24 Remarks on the sampling variability of regression lines.

The regression line of Y on X, as determined by the method of least squares, provides us with an objective method for making an estimate of the value of Y for an additional pair of measurements in which the value of X is given. In practice, such a regression line is determined from a sample of observed pairs of measurements. If we imagine a population (either finite or indefinitely large) of pairs of measurements, we can imagine a regression line of Y on X for this population.

It is clear that if the sample regression coefficient $\hat{b}$ is zero, then the regression line reduces to the form $Y - \bar{Y} = 0$. This means that the regression line is parallel to the X axis and that we would get the same estimate of Y no matter what value of X we have. In practice, we would rarely find a sample regression coefficient equal to zero, although we may get values

close to zero sometimes. If we consider many samples from a population with a given regression coefficient, b, the sample regression coefficient will have values clustering around the value of b in the population, i.e., they will have their own theoretical sampling distribution. This theoretical sampling distribution is known under the assumption that the population is indefinitely large and is such that if we consider the pairs of measurements having a specific value of X, then Y is a chance quantity having a normal distribution with mean $a+bX$ and variance σ^2. This is the assumption usually made when we consider sampling fluctuations of $\hat{b}$. In fact, if b is the value of the regression coefficient in such a population and if $\hat{b}$ is the value of the regression coefficient in a sample, then the quantity

$$\frac{(\hat{b} - b)\sqrt{n-1}\ s_X}{s_E} \tag{13.19}$$

will have the Student t distribution (see Section 10.4) with n-2 degrees of freedom. This means that if we choose a confidence coefficient α, we can say that

$$\Pr\left(-t_\alpha < \frac{(\hat{b} - b)\sqrt{n-1}\ s_X}{s_E} < t_\alpha\right) = \alpha\ , \tag{13.20}$$

where the value of t_α is found from Table 10.2 for any specified number of degrees of freedom. Expression (13.20) may be rewritten as

$$\Pr\left(\hat{b} - \frac{s_E\, t_\alpha}{\sqrt{n-1}\ s_X} < b < \hat{b} + \frac{s_E\, t_\alpha}{\sqrt{n-1}\ s_X}\right) = \alpha\ , \tag{13.21}$$

which means that the following expressions are 100α % confidence limits of the value of the regression coefficient b in the population:

$$\hat{b} \pm \frac{s_E\, t_\alpha}{\sqrt{n-1}\ s_X}\ . \tag{13.22}$$

The assumptions under which the confidence limits (13.22) are exact should be kept in mind. In practice, such assumptions are only approximately satisfied and (13.22) would only be considered as approximate 100α% confidence limits.

In the example of 17 pairs of scores, suppose we consider this as a sample of 17 drawn at random from a population of students taking the particular

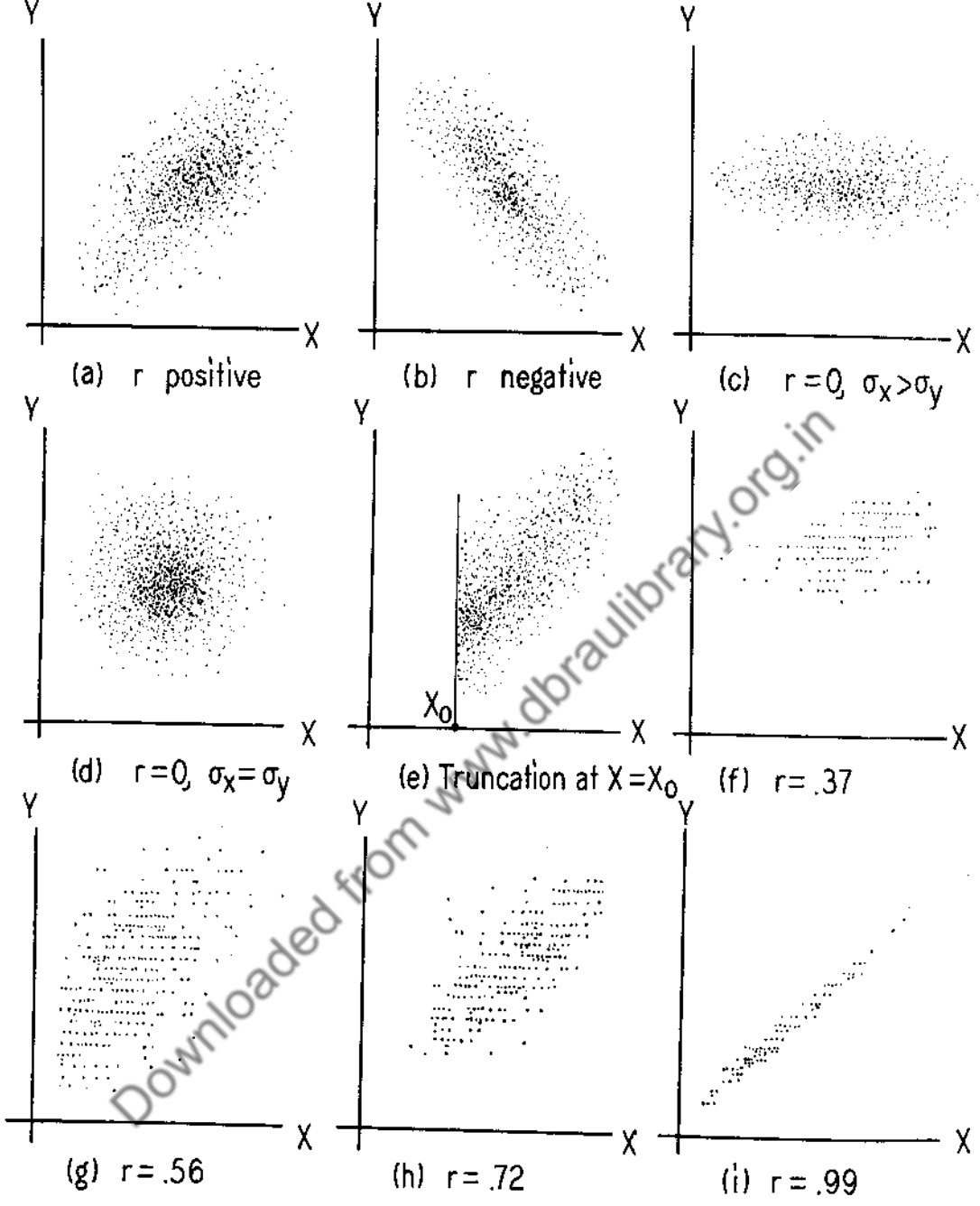
mathematics course involved in the example. Substituting in (13.22) the numerical values of $\hat{b}$, s_E, s_X and r which we have found and substituting $t_a = 2.131$ (from Table 10.2 for $a = .95$ and 15 degrees of freedom) we find the following 95% confidence limits of b :

$$.839 \pm \frac{(9.488)(2.131)}{\sqrt{16}\,(11.71} = .839 \pm .432 .$$

Similarly, we can establish confidence limits for the population value of the intercept a, or for any quantity such as $a + bX_o$, (the mean of Y in the population of pairs of measurements having $X = X_o$). These confidence limits are a little more complex and will not be given here.

13.25 Remarks on the correlation coefficient.

The correlation coefficient r, defined by (13.10), is useful in certain kinds of statistical problems as a simple index for expressing the degree of relationship between two variables X and Y when the scatter diagram for the sample consists of a swarm of points which is roughly elliptical in shape, and when we wish to consider neither variable as a predictor for the other as in the case of regression analysis. The value of r lies between -1 and +1. Positive values of r indicate positive relationship, the long direction of the cluster running from lower left to upper right, so as X increases, Y increases (see (a) in Figure 13.5). Negative values of r indicate a negative relationship, the long direction of the elliptical cluster running from upper left to lower right so that as X increases, Y decreases (see (b) in Figure 13.5 . If $r = +1$, then all points on the scatter diagram will fall on a straight line with a positive slope. Similarly, if $r = -1$, the points of the scatter diagram fall along a straight line with negative slope. In either of these cases the relationship between X and Y will be a perfect linear relationship. If $r = 0$, the scatter diagram is such that the two regression lines are parallel to the X and Y axes respectively, and (considering scatter diagrams which are roughly elliptical in shape) we consider there to be no relationship between X and Y. If the cluster of points is something like that in (c) of Figure 13.5, the standard deviation of X exceeds that of Y and r will be 0 (or nearly so). If the scatter diagram is circular, like (d) in Figure 13.5, the two standard deviations will be equal (or nearly so) and r will be 0 (or nearly so). If one should eliminate all



Scatter Diagrams with Various Correlation Coefficients (Reproduced by by permission of the College Entrance Examination Board)

Figure 13.5

points in the scatter diagram for which X is less than X_0 we would get a truncated scatter diagram (see (d) in Figure 13.5). The effect of such a truncation is, in general, to lower the value of r. To get some notion of the size of r for actual scatter diagrams showing various degrees of relationship between X and Y, the values of r for (f), (g), (h) and (i) of Figure 13.5 are .37, .56, .72 and .99, respectively.

If a large number of samples is considered as being drawn at random from a population of pairs of measurements having a correlation coefficient ρ, the values of the correlation coefficient in these samples will have a sampling distribution. The theoretical sampling distribution of a correlation coefficient for a sample of a given size is very complicated. The mathematical formula for it is known in case the population is indefinitely large and has a two-variable normal or Gaussian distribution having a correlation coefficient ρ. To attempt to discuss such a population would be beyond the scope of this course. It is sufficient to say that such a distribution serves as a satisfactory model for describing the way in which plotted points will be distributed in large-sample scatter diagrams which occur in plotting pairs of examination scores for a large number of students, heights and weights of a large number of men, length and breadth of leaves of a given tree, and so on.

In spite of the complication of the theoretical sampling distribution of the correlation coefficient, diagrams have been worked out for determining from a sample correlation coefficient r confidence limits of the correlation coefficient ρ of the population from which the sample is drawn. Figure 13.6 gives such a diagram for a confidence coefficient of 0.95.

As an example illustrating the use of Figure 13.6, suppose a sample of 25 pairs of measurements from a two-variable normal population yields a correlation coefficient equal to .60. What can we consider to be the 95% confidence limits in the population? The answer is found by entering the scale of r (sample correlation coefficient) at r = +.60, and noticing where the vertical line through r = +.60 cuts the two curves marked 25. Reading the two cuts on the scale of ρ (population correlation coefficient), we find the 95% confidence limits of ρ to be .27 and .79. Samples of only 25 pairs of measurements do not determine very close confidence limits for ρ. If the sample had r = +.60 and n = 400, the 95% confidence limits would be .53 and .66 — which are much closer together.

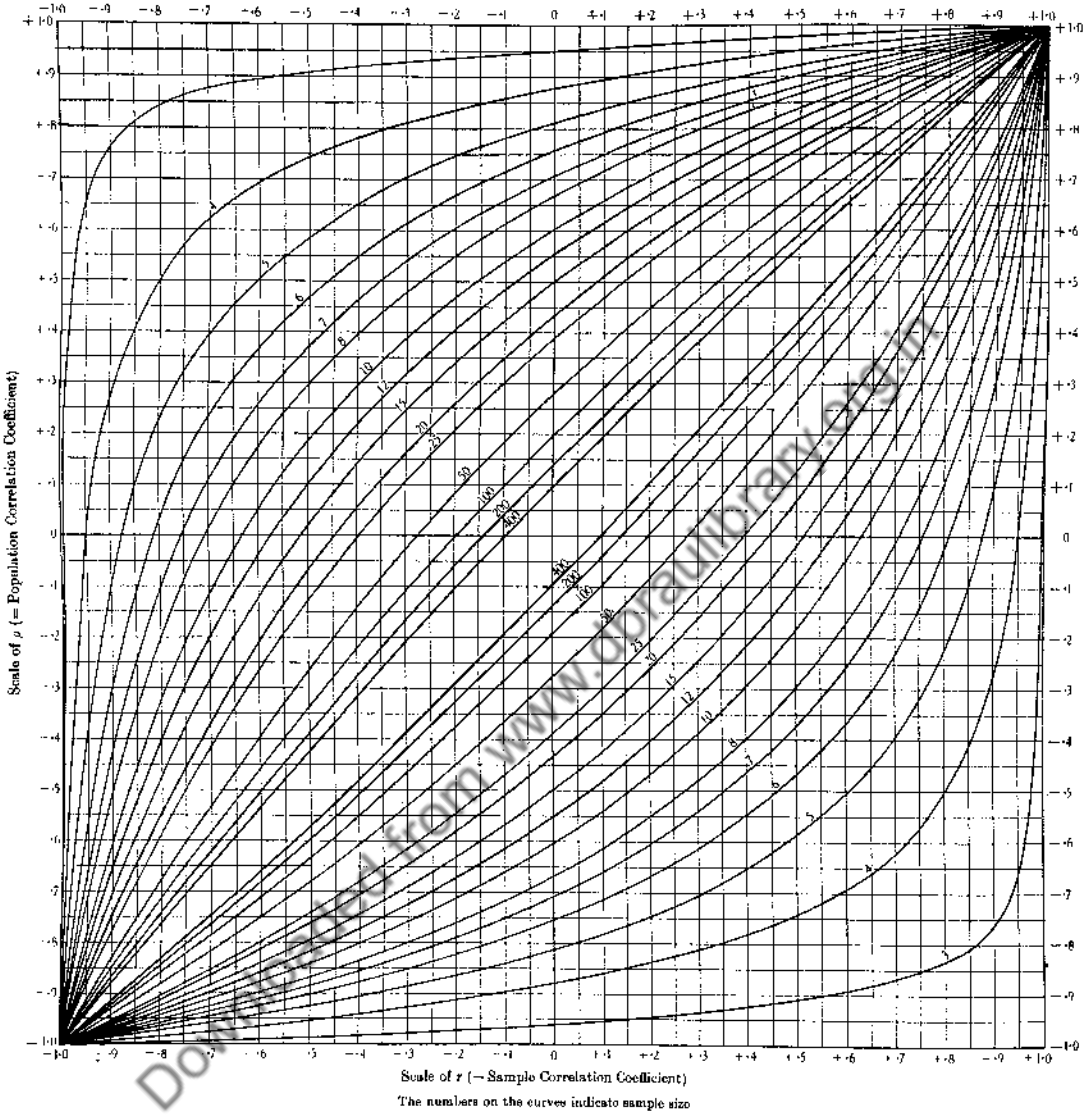


Diagram for Determining 95% Confidence Limits of the Correlation Coefficient for Sample Sizes Ranging from 3 to 400

(Reproduced by courtesy of the author, F. N. David, and the publisher, the Biometrika Office)

Figure 13.6

Exercise 13.2.

1. The following table shows the scores made on a placement mathematics test by 10 students and also the final groups made by these men in a freshman mathematics course at Princeton.

Placement Test Score X	Final Group Y
60	7
72	2
67	4
77	4
69	3
87	1
75	3
81	1
81	1
61	4

(a) Make a scatter diagram.

(b) Find the means, standard deviations, correlation coefficient for X and Y, the regression coefficient of Y on X, and the standard error of estimate s_E.

(c) Determine the regression equation of Y on X and graph it on the scatter diagram.

(d) What Final Group would you predict for a student who made 66 on the placement test?

2. The percent carotene content of wheat (X) and the percent carotene content of flour (Y) for 10 varieties of wheat were found to be as follows (from Goulden's data):

X	Y
1.16	2.39
2.13	3.11
1.41	2.15
1.42	1.96
1.50	2.02
1.25	1.76
1.65	2.10
1.24	2.12
1.48	2.28
1.35	1.86

(a) Make a scatter diagram.

(b) Find the means, standard deviations, correlation coefficient for X and Y, regression coefficient of Y on X, and the standard error of estimate s_E.

(c) Determine the regression line of Y on X and graph it on the scatter diagram.

(d) If a new variety of wheat is found to have 1.90% carotene, what estimate would you make for the carotene content of flour made from this variety?

3. Tensile strength (Y) in 1000 pounds per square inch and hardness (X) in Rockwell's E for each of 10 specimens of aluminum die castings were found to be as follows (from Shewhart's data):

X	Y
53.0	29.3
70.2	34.9
84.3	36.8
55.3	30.1
78.5	34.0
63.5	30.8
71.4	35.4
53.4	31.5
82.5	32.2
67.3	33.4

(a) Make a scatter diagram.

(b) Find the means, standard deviations, correlation coefficient for X and Y, regression coefficient of Y on X, and the standard error of estimate s_E.

(c) Determine the regression line of Y on X and graph it on the scatter diagram.

(d) If a die-casting is found to have a Rockwell hardness of 35, what would you estimate its tensile strength to be?

(e) Find 95% confidence limits of the regression coefficient b of the population regression line. Interpret these confidence limits.

4. Find the means, standard deviations, correlation coefficient and standard error of estimate s_E from Problem No. 2 of Exercise 13.1. Determine the regression line of Y on X from these quantities and plot the regression line. If the Rapid Method yielded 55% fat in a new brand of mayonnaise, what estimate would you make for the fat content that would be found by the A.O.A.C. method?

5. For a certain group of 200 college entrance students, suppose the regression line of French achievement score (Y) on verbal aptitude score (X) is

$$Y = .72\ X + 141.2\ .$$

If it is known that the mean and standard deviation of the French achievement test for this group are 512 and 98 respectively, and that the standard deviation of the verbal aptitude scores for the group is 100, what is the mean of the verbal aptitude scores and the correlation coefficient between X and Y? Write down the equation of the regression line of X on Y. If the French achievement score of a student from this group is 580, what estimate would you make for his verbal aptitude score?

6. Suppose the correlation coefficient between the verbal aptitude score and mathematical aptitude score of a random sample of 50 students from the class of 1952 turns out to be .58. Assuming Figure 13.6 to be applicable to this problem for all practical purposes, find approximately the 95% confidence limits of the correlation coefficient between these scores for the entire class of students.

7. Suppose a sample of 50 pairs of measurements from a population having approximately a two-variable normal distribution yields a correlation coefficient of .20. Is this significantly different from 0 at the 95% probability level? (Make use of Figure 13.6.)

8. Suppose the average height of male students at College A and that of their fathers are 68 inches and that the standard deviations of heights of sons and fathers are equal. Suppose the heights of sons (eldest if there is more than one son at College A) and fathers have a correlation coefficient of .50. Graph the regression line of sons' height on fathers' height. What can you say about the average height of sons of fathers of a given height above average height for fathers? Similarly for sons of fathers of a height below average?

13.3 Simplified Computation of Coefficients for Regression Line.

The computations involved in Table 13.3 are straightforward but cumbersome. Simplified, but less direct, computational schemes can be devised which will lighten the arithmetic. We shall consider two schemes; the first

is a scheme making use of a working origin which is useful for less than 40 or 50 pairs of measurements, and the second is a fully coded scheme useful for more than this number of pairs of measurements.

13.31 Computation by using a working origin.

Let us return to the example of 17 pairs of scores. We choose a convenient working origin for the X measurement near the middle of the set of values of X in the sample. 50 will be satisfactory. Similarly, let us choose 40 as a working origin for the Y measurement. Denote the deviations X-50 and Y-40 by X' and Y' respectively. Then

$$X = X' + 50$$
$$Y = Y' + 40 .$$

Using the results of Section 3.41, we have

$$\bar{X} = \bar{X}' + 50$$
$$\bar{Y} = \bar{Y}' + 40 \tag{13.23}$$
$$s_X^2 = s_{X'}^2, \quad s_Y^2 = s_{Y'}^2 .$$

Also

$$S[(X-\bar{X})(Y-\bar{Y})] = S[(X'-\bar{X}')(Y'-\bar{Y}')]$$

i.e., cov (X,Y) = cov (X',Y'). Therefore, the correlation coefficient between X and Y is equal to the correlation coefficient between X' and Y', which means we can compute r by using the formula

$$r = \frac{\text{cov}(X',Y')}{s_{X'}\, s_{Y'}} . \tag{13.24}$$

We can evaluate S(X'), S(Y'), S(X'Y'), $S(X'^2)$, $S(Y'^2)$ by constructing Table 13.4, and from these we determine the values of $\bar{X}$, $\bar{Y}$, s_X, s_Y, r, as shown below Table 13.4.

In general, suppose we have a sample of n pairs of measurements (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). Now let X_o be the working origin of X measurements, and Y_o the origin of Y measurements. Then let X' and Y' be defined as

$$X' = X - X_o$$

$$Y' = Y - Y_o .$$

Following methods very similar to those of Section 3.4 we may express the means, variances, regression coefficient, and correlation coefficient, in terms of the similar quantities for X' and Y' as follows:

$$\bar{X} = X_o + \bar{X}', \quad \bar{Y} = Y_o + \bar{Y}'$$

$$s_X^2 = s_{X'}^2 , \quad s_Y^2 = s_{Y'}^2$$

$$\mathrm{cov}(X,Y) = \mathrm{cov}(X',Y')$$

(13.25)
$$\hat{b} = \frac{\mathrm{cov}(X,Y)}{s_X^2} = \frac{\mathrm{cov}(X',Y')}{s_{X'}^2}$$

$$r = \frac{\mathrm{cov}(X,Y)}{s_X s_Y} = \frac{\mathrm{cov}(X',Y')}{s_{X'} s_{Y'}} .$$

TABLE 13.4

Table for Calculating Means, Variances and the Correlation Coefficient by Use of Working Origin

	X	Y	X'	Y'	X'^2	Y'^2	X'Y'
	50	44	0	4	0	16	0
	44	25	-6	-15	36	225	90
	54	44	4	4	16	16	16
	46	26	-4	-14	16	196	56
	64	41	14	1	196	1	14
	37	21	-13	-19	169	361	247
	77	60	27	20	729	400	540
	62	52	12	12	144	144	144
	45	15	-5	-25	25	625	125
	69	55	19	15	361	225	285
	44	42	-6	2	36	4	-12
	57	58	7	18	49	324	126
	63	37	13	-3	169	9	-39
	68	36	18	-4	324	16	-72
	45	40	-5	0	25	0	0
	71	58	21	18	441	324	378
	50	36	0	-4	0	16	0
Total			135 -39 96	94 -84 10	2736	2902	2021 -123 1898

$$S(X') = 96 , \quad \bar{X}' = \frac{96}{17} = 5.647 , \quad \bar{X} = \bar{X}' + 50 = 55.65$$

$$S(Y') = 10 , \quad \bar{Y}' = \frac{10}{17} = .588 , \quad \bar{Y} = \bar{Y}' + 40 = 40.59$$

$$S(X'^2) = 2736 , \quad s^2_{X'} = \frac{1}{16}\left[2736 - \frac{(96)^2}{17}\right] = 137.1176 , \quad s_X = s_{X'} = 11.71$$

$$S(Y'^2) = 2902 , \quad s^2_{Y'} = \frac{1}{16}\left[2902 - \frac{(10)^2}{17}\right] = 181.0074 , \quad s_Y = s_{Y'} = 13.45$$

$$S(X'Y') = 1898, \quad \mathrm{cov}(X',Y') = \frac{1}{16}\left[1898 - \frac{(10)(96)}{17}\right] = 115.0956$$

$$r = \frac{115.0956}{\sqrt{(137.1176)(181.0074)}} = .7306 .$$

Once the values of $\bar{X}$, $\bar{Y}$, s_X, s_Y, and r are calculated from these formulas, we can then proceed to substitute these values in formulas such as (13.11), (13.12) and (13.18).

13.32 Computation by using a fully coded scheme.

The method described in Section 13.31 is useful for computing the means, variances and the covariance when the number of pairs of measurements does not exceed 40 or 50. If there is a larger number of measurements, it is worthwhile to consider a scheme which involves grouping the observations with respect to X and also with respect to Y. We determine cell lengths, cell boundaries and cell midpoints for X and for Y by following procedures similar to those set up in Chapter 3. We also introduce new units of measurement and new origins for both variables. We use the cell lengths of the X distribution and of the Y distribution as new units of measurement.

If X_o and Y_o are the arbitrary origins for X and Y respectively, if the new units for the X and Y measurements are h and k, and if the new variables used for measuring X and Y are Z and W, we may then write

(13.26)
$$X = X_o + hZ$$
$$Y = Y_o + kW .$$

Now, we would like to compute the means, variances and covariance for the X and Y measurements in terms of X_o, Y_o, k, h and the means, variances and covariance

for Z and W.

Consider a sample of pairs of measurements (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). For any one of these pairs of measurements, say (X_j, Y_j), there is a pair of measurements (Z_j, W_j) which we can determine by substituting $Y = Y_j$ in (13.26) and solving for Z and W. From Section 3.42 we may write down at once the following formulas:

$$\bar{X} = X_o + h\bar{Z} \ , \quad \bar{Y} = Y_o + k\bar{W}$$

(13.27)

$$s_X^2 = h^2 \cdot s_Z^2 \ , \quad s_Y^2 = k^2 \cdot s_W^2$$

We must now find a formula for calculating the covariance. We have

$$\begin{aligned} \text{cov}(X,Y) &= S[(X-\bar{X})(Y-\bar{Y})] \\ &= S[(hZ-h\bar{Z})(kW-k\bar{W})] \\ &= hkS[(Z-\bar{Z})(W-\bar{W})] \ , \end{aligned}$$

since h and k are constants.

$$\text{cov}(X,Y) = hk \ \text{cov}(Z,W) \ . \tag{13.28}$$

Therefore, since $s_X = hs_Z$ and $s_Y = ks_W$, we have for the regression coefficient and correlation coefficient

$$\hat{b} = \frac{\text{cov}(X,Y)}{s_X^2} = \frac{h\,k\,\text{cov}(Z,W)}{h^2\,s_Z^2} = \frac{k\,\text{cov}(Z,W)}{h\,s_Z^2} \ ,$$

(13.29)

$$r = \frac{\text{cov}(X,Y)}{s_X\,s_Y} = \frac{h\,k\,\text{cov}(Z,W)}{(h\,s_Z)(k\,s_W)} = \frac{\text{cov}(Z,W)}{s_Z \cdot s_W} \ .$$

Thus, the correlation coefficient between X and Y has the same value as that between Z and W. This is another way of saying that the correlation coefficient for a scatter diagram remains the same, no matter where the origin is located, and no matter what units are used to describe the measurements.

Formulas (13.26), (13.27), (13.28) and (13.29) are the formulas for the fully coded computation of the means, variances, covariance, regression coefficient and correlation coefficient.

Let us illustrate the computations by an example. Table 13.5 (data

by Milbourn) shows measurements of initial thickness (X) and final thickness (Y) in .0001" units at 88 positions on a coil of sheet metal, the X measurement being obtained before and the Y measurement after a rolling operation.

The first thing we do with the measurements in Table 13.5 is to make a two-way frequency table. It is convenient to select cell lengths of 5 for both variables, i.e.,

$$h = k = 5 \ .$$

The following cell midpoints are convenient:

for X : 553, 558, ..., 623
for Y : 398, 403, ..., 453 .

TABLE 13.5

Thickness before Rolling (X) and Thickness after Rolling (Y) at 88 Positions of a Coil of Sheet Metal

X	Y	X	Y	X	Y	X	Y
577	408	581	427	610	439	607	442
568	397	601	432	604	437	602	431
568	409	597	432	605	448	608	437
589	406	592	439	609	435	622	451
590	413	588	423	602	442	614	437
575	412	595	428	606	435	588	432
571	398	599	438	610	439	598	431
572	410	594	428	612	450	608	431
576	409	591	438	604	431	612	446
583	419	597	430	598	437	608	431
580	417	601	432	598	432	602	433
569	402	601	437	615	437	601	434
578	415	596	437	612	450	599	426
581	412	593	438	608	432	602	440
589	415	605	438	603	438	596	426
577	421	605	438	608	444	592	423
579	407	594	442	608	437	598	429
574	425	598	430	611	448	603	422
592	419	601	439	605	433	593	427
594	421	608	444	602	434	567	404
583	428	610	438	608	443	566	406
572	409	598	443	612	438	553	404

We now construct the two-way frequency table as shown bordered by the heavy lines in Table 13.6, tallying the frequencies with which the various

TABLE 13.6

Scheme for Fully Coded Computation of Means, Variances, and Covariance for Two Variables

	y_j / x_i	553	558	563	568	573	578	583	588	593	598	603	608	613	618	623	w_j	g_j	$f_j w_j$	$g_j w_j^2$	v_j	$v_j w_j$
	453															1	6	1	6	36	7	42
	448										1		4				5	5	25	125	23	115
	443									1	1	1	4				4	7	28	112	22	88
	438									3	3	7	5	3			3	21	63	189	65	195
	433								1		3	8	5				2	17	34	68	50	100
	428							2		3	5						1	10	10	10	11	11
	423						1	1		1	2		1				0	6	0	0	0	0
	418							1	1		1						-1	3	-3	3	-2	2
	413						1	1	1	2							-2	5	-10	20	-6	12
	408					2	2	3		1							-3	8	-24	72	-20	60
	403	1				2											-4	3	-12	48	-15	60
	398					1	1										-5	2	-10	50	-7	35
(g)																		(h)	(i)	(j)	(k)	(l)
(f)	$u_i z_i$	28	0	0	76	39	24	1	0	15	48	138	164	145	0	42	720=S(ZW)					
(e)	u_i	-4	0	0	-19	-13	-12	-1	-5	15	24	46	41	29	0	6						
(d)	$f_i z_i^2$	49	0	0	80	45	24	4	0	10	48	162	224	175	0	49	870=S(Z^2)					
(c)	$f_i z_i$	-7	0	0	-20	-15	-12	-4	0	10	24	54	56	35	0	7	128=S(Z)					
(b)	f_i	1	0	0	5	5	6	4	5	10	12	18	14	7	0	1	88 = n	88	S(W) = 107	$S(W^2)$ = 733		S(ZW) = 720
(a)	z_i	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7						

combinations of values of X and Y fall into the various cells. In an actual situation, it is sufficient to put the tally marks in light pencil in the upper left-hand corner of each cell and then erase them when they are counted. Hence, we shall show the total frequency in each cell and not the tally marks. We shall record only cell midpoints - not cell boundaries.

Looking at Table 13.6, you will see that rows (a), (b), (c), (d), (e), (f) and columns (g), (h), (i), (j), (k), (l) have been added for computational purposes. Row (a) is for the values (integral values) of Z, the coded variable for measurement X. Row (b) is the frequency distribution of values of Z; the frequencies are obtained by adding the frequencies in the two-way frequency table by columns. The sum of the entries in row (b) is n, the number of pairs of measurements in the sample. Row (c) is for the product of values of Z and the frequency and the sum of its entries gives $S(Z)$. Row (d) is for the product of Z^2 and the frequency, and the sum of its entries gives the value of $S(Z^2)$. Any entry in row (e) is obtained by multiplying the frequencies in the column of the two-way frequency table corresponding to that entry by their W values and adding. For instance, we obtain the entry 15 as follows:

$$(4)(1) + (3)(3) + (2)(0) + (1)(3) + (0)(2) + (-1)(1)$$

$$= 4 + 9 + 0 + 3 + 0 - 1 = 15 .$$

The entries in row (f) are obtained by multiplying corresponding entries in row (e) and row (a). The sum of the entries in row (f) is the sum of the products ZW for each entry in the two-way table, i.e., $S(ZW)$. Rows (g), (h), (i), (j), (k), (l) give similar results for the Y and W measurements and their distribution. Since the total of row (f) and the total of column (l) are each equal to $S(ZW)$, we have a convenient computational check. Applying formulas (13.26), (13.27), (13.28), (13.29), the computations of the means, variances, covariance, regression coefficient and correlation coefficient for X and Y from the material in Table 13.6 are as follows:

$$X_o = 588 , \quad Y_o = 423 , \quad h = k = 5 ,$$

$$\overline{Z} = \frac{128}{88} = 1.4545 , \quad \overline{W} = \frac{107}{88} = 1.2159$$

$$\bar{X} = 595.27 \ , \ \bar{Y} = 429.08$$

$$s_Z^2 = \frac{1}{87}\left[870 - \frac{(128)^2}{88}\right] = 7.8600 \ , \ s_X = 5\sqrt{7.8600} = 14.02$$

$$s_W^2 = \frac{1}{87}\left[733 - \frac{(107)^2}{88}\right] = 6.9299 \ , \ s_Y = 5\sqrt{6.9299} = 13.16$$

$$\text{cov}(Z,W) = \frac{1}{87}\left[720 - \frac{(128)(107)}{88}\right] = 6.4869$$

$$\hat{b} = \frac{5(6.4869)}{5(7.8600)} = .83$$

$$r = \frac{6.4869}{\sqrt{(7.8600)\ (6.9299)}} = .88$$

The equation of the regression line of Y on X is

$$(Y - 429.08) = .83\ (X - 595.27) \ .$$

A general table similar to the illustrative Table 13.6 could be constructed and discussed, but this seems unnecessary. You will be able to see the full generality of the fully coded computational scheme from the example we have given.

There are other fully coded computational schemes for calculating means, variances, and the covariance, but we shall not consider them here.

Exercise 13.3.

1. Each of fifteen expert riflemen fired a set of rounds in a kneeling position and a set of rounds in a standing position. Each man obtained a score X for his firings from the kneeling position, and a score Y for his firings from the standing position. The pairs of scores for the 15 men were as follows (from Scarborough and Wagner):

X	Y	X	Y
91	78	93	82
93	85	88	71
91	82	88	83
89	79	92	89
90	75	91	84
91	87	92	79
95	86	94	86
		94	85

(a) Make a scatter diagram of these points.

(b) Find the means, standard deviations, and correlation coefficient for X and Y, making use of a working origin.

(c) Find the regression lines of Y on X and of X on Y and graph them on the scatter diagram.

(d) Find the standard error of estimate of Y from X. Also the standard error of estimate of X from Y.

(e) If an expert rifleman makes a score of 96 from a kneeling position, what score would you estimate him to make if he fires from a standing position? If such a rifleman makes a score of 70 from a standing position, what score would you estimate him to make from a kneeling position?

2. Thirty prepared specimens of a synthetic rubber (Neoprene GN) were tested for abrasion loss in cc. per H.P. hour (Y) and hardness in degrees Shore (X). The following data (from Buist and Davies) were obtained:

X	Y	X	Y	X	Y
45	372	64	164	71	219
55	206	68	113	80	186
61	175	79	182	82	155
66	154	81	32	89	114
71	136	56	228	51	341
71	112	68	196	59	340
81	55	75	128	65	283
86	45	83	97	74	267
53	221	88	84	81	215
60	166	59	249	86	148

(a) Make a scatter diagram.

(b) Find the regression line of Y on X and plot it on the scatter diagram.

(c) Find the standard error of estimate s_E and draw a line on each side of the regression line at a vertical distance of s_E and parallel to the regression line.

(d) Find 95% confidence limits of $\hat{b}$ and r.

(e) If a specimen of the rubber should have a hardness of 78, what estimate would you make for abrasion loss?

3. The 67 Economics Departmental students of the Princeton class of 1938 had the following Verbal Scores (X) and Mathematical Scores (Y) on the Scholastic Aptitude Test (S.A.T.):

X	Y	X	Y	X	Y
345	577	523	550	585	537
395	569	556	550	593	717
563	608	479	678	417	496
543	505	629	614	486	582
402	705	490	640	604	647
472	531	730	556	515	620
691	577	611	730	523	629
624	556	468	614	545	511
523	634	574	453	505	756
461	357	420	621	527	621
490	589	596	614	384	527
530	672	585	498	431	524
516	640	354	569	574	627
444	499	494	698	494	543
604	543	439	595	560	466
406	473	446	543	464	589
475	569	505	511	549	548
585	679	585	672	541	705
523	608	468	653	468	582
582	556	578	466	629	595
575	549	603	634	607	563
439	350	417	666	490	498
				549	537

(a) Make a scatter diagram for these 67 pairs of scores.

(b) The mean Verbal Score and mean Mathematical Score of all College Board candidates taking the S.A.T., is 500 in each case. From the scatter diagram of the 67 pairs of scores, determine how many of the Economics Departmental Students have scores: (i) above average on both the Verbal and the Mathematical parts of the S.A.T., (ii) below average on both, (iii) above average on the Mathematical part and below average on the Verbal part, and (iv) below average on the Verbal part and above average on the Mathematical part.

(c) By using one of the simplified computational schemes, calculate the means, standard deviations and correlation coefficient for X and Y.

(d) Write down the regression equation of Y on X and plot it on the scatter diagram.

(e) Find the standard error of estimate s_E.

4. A mathematics test consisting of 5 subtests was given to a group of freshmen students at a certain military academy. A study was made of the relationship between the Part I Score (X) and the Total Score (Y) of the test. The pairs of scores for the 139 students are given in the following two-way frequency table, in which the cell midpoints are given at the left and below.

y_j / x_i	3	4	5	6	7	8	9	10	11	12	13	14	15	16
137														1
132													2	1
127												1	1	1
122													1	
117										1	2		1	
112									3		3		1	1
107								1			4	1		1
102					1	1		1		3		3	1	1
97								1	2	1	1	1		
92							1	4	1					
87				1	1		1	2	2	3				
82						6	2	4	4		2			
77			1		1		2	2	1	1				
72		1	1	1	2	4	1	2						
67				1	3	3		3		2				
62				2	4	2								
57				2	2	1		1						
52	1		1	1		2								
47		2		1										
42	1	2	1		1									
37			1											
32														
27		1												

(a) By use of the fully coded scheme, determine the means, standard deviations, and correlation coefficient.

(b) Find the equation of the regression line of Y on X.

(c) Find the standard error of estimate s_E.

(d) What estimate would you make for the Total Score of a student who took only Part I and made a score of 14 on it?

13.4 Generality of the Method of Least Squares.

In Section 13.2 we discussed the problem of fitting a regression line

of the form $Y = a + bX$ to a sample of n points by the method of least squares. While this is one of the most important cases which arises in elementary statistical analysis, it is to be emphasized that the method of least squares is used for fitting other kinds of regression lines and curves.

13.41 Fitting a line through the origin by least squares.

There are problems in which it may be sufficient to fit a straight line of the form $Y = bX$, i.e., cases in which we may safely put $a = 0$ at the outset. The example considered in Section 13.1 is such a case. In that case it will be sufficient for routine purposes to fit the very simple formula $Y = bX$, from which accurate estimates of concentration of alpha-resin can be made from colorimeter readings by a simple multiplication. To fit the line $Y = bX$ to the data of Table 13.1, we proceed as we did in Section 13.2 for fitting the line $Y = a + bX$ and consider the sum of squares of differences between actual values of Y and estimated values of Y, i.e., $S = (.12 - 8b)^2 + (.71 - 50b)^2 + \ldots + (2.50 - 181b)^2$. The value of b which minimizes S is that for which $\frac{dS}{db} = 0$. This gives

$$-2(8)(.12-8b)-2(50)(.71-50b) - \ldots -2(181)(2.50-181b) = 0 .$$

Dividing by 2 and collecting terms we get

$$-991.01 + 71890b = 0$$

from which

$$\hat{b} = .0138 .$$

The least squares fitted line is therefore

$$Y = .0138X .$$

If this line is graphed on Figure 13.1, it does not differ perceptibly from the line already plotted "by eye", thus indicating that when there is a very close straight line relationship between two variables, little, if anything, is to be gained by using least squares.

13.42 Fitting parabolas and higher degree polynomials.

There are situations in which we may wish to fit a parabola of the form

$$Y = a + bX + cX^2 , \tag{13.30}$$

a cubic of the form

$$Y = a + bX + cX^2 + dX^3 ,$$

or a polynomial of higher degree.

The procedure here is just as before: we take the difference between each actual Y and estimated Y, square the difference and form S, the sum of the squared differences. S will involve the undetermined constants a, b, c, etc. The values of a, b, c, etc., which minimize S are those which satisfy the simultaneous equations obtained by setting $\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0, \frac{\partial S}{\partial c} = 0$, etc. These equations are called normal equations. There are as many normal equations as constarts to be determined.

To illustrate the case for 3 constants, let us fit $Y = a + bX + cX^2$ to the following data:

X	0	1	2	3
Y	-.1	1.1	3.9	9.2

For S we have

$$S = (-.1-a)^2 + (1.1-a-b-c)^2 + (3.9-a-2b-4c)^2 + (9.2-a-3b-9c)^2 .$$

The normal equations are

$$-2(-.1-a)-2(1.1-a-b-c)-2(3.9-a-2b-4c)-2(9.2-a-3b-9c) = 0$$

$$-2(1.1-a-b-c)-2(2)(3.9-a-2b-4c)-2(3)(9.2-a-3b-9c) = 0$$

$$-2(1.1-a-b-c)-2(4)(3.9-a-2b-4c)-2(9)(9.2-a-3b-9c) = 0 .$$

Dividing by 2, and simplifying, we have

$$-14.1 + 4a + 6b + 14c = 0$$

$$-36.5 + 6a + 14b + 36c = 0$$

$$-99.5 + 14a + 36b + 98c = 0 .$$

Solving these equations we get

$$\hat{a} = -.055, \hat{b} = -.005, \hat{c} = 1.025 ,$$

and hence the equation of the regression curve of Y on X is

$$Y = -.055 - .005X + 1.025X^2 .$$

In general, for n pairs of measurements (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), we have

$$S = \sum_{j=1}^{n} (Y_j - a - bX_j - cX_j^2)^2 ,$$

or more briefly

$$S = S(Y - a - bX - cX^2)^2 . \tag{13.31}$$

The normal equations are

$$-2 \cdot S(Y - a - bX - cX^2) = 0$$
$$-2 \cdot S[X(Y - a - bX - cX^2)] = 0$$
$$-2 \cdot S[X^2(Y - a - bX - cX^2)] = 0 .$$

Dividing by 2 and writing the coefficients of these equations out explicitly, we have

$$\begin{aligned} -S(Y) + aS(1) + bS(X) + cS(X^2) &= 0 \\ -S(XY) + aS(X) + bS(X^2) + cS(X^3) &= 0 \\ -S(X^2Y) + aS(X^2) + bS(X^3) + cS(X^4) &= 0 , \end{aligned} \tag{13.32}$$

which shows the general structure of the normal equations, and also indicates how one would set up normal equations for determining 4, 5 or any number of constants. In any particular example there may be various shortcuts to solving the normal equations. The most systematic method is successive elimination, but we shall not go into this method here.

The job of finding the variance of estimate s_E^2 of Y from X using the regression equation (13.30) amounts to inserting the solutions $\hat{a}$, $\hat{b}$, and $\hat{c}$ of the equations (13.32) into (13.31), finding the sum of the squares and dividing by n-3, the number of degrees of freedom. If there are not many pairs of measurements, say less than 20, the simplest procedure is to evaluate each difference in the sum of squares, square it and add the squares. For more pairs of measurements, we can express the sum of squares in terms of $\hat{a}$, $\hat{b}$, $\hat{c}$, and sums of various powers and products of X and Y. For squaring $(Y - \hat{a} - \hat{b}X - \hat{c}X^2)$ and summing over all n pairs of measurements, we have

$$S(Y - \hat{a} - \hat{b}X - \hat{c}X^2)^2 = S(Y^2) + \hat{a}^2S(1) + \hat{b}^2S(X^2) + \hat{c}^2S(X^4) \tag{13.33}$$

$$+ 2\hat{a}\hat{b}S(X) + 2\hat{a}\hat{c}S(X^2) + 2\hat{b}\hat{c}S(X^3) - 2\hat{a}S(Y) - 2\hat{b}S(XY) - 2\hat{c}S(X^2Y).$$

Evaluating all the terms on the right-hand side of (13.33) from the data, inserting their values and dividing the result by n-3, gives the variance of estimate.

13.43 Fitting exponential functions.

In growth and other types of problems, we are often faced with the problem of fitting a curve of the form

$$U = Ae^{bX} \tag{13.34}$$

to n pairs of measurements (U_1, X_1), (U_2, X_2), ..., (U_n, X_n) where A and b are constants and e = 2.71828...., the base of natural logarithms.

If we take natural logarithms (i.e., $\log_e$) of both sides of (13.34) we have

$$\log_e U = \log_e A + bX . \tag{13.35}$$

Putting $\log_e U = Y$, and $\log_e A = a$, equation (13.35) may be written as

$$Y = a + bX ,$$

which can be fitted to the measurements (Y_1, X_1), (X_2, Y_2), ..., (X_n, Y_n) [i.e., to the pairs of measurements $(X_1, \log_e U_1)$, $(X_2, \log_e U_2)$, ..., $(X_n, \log_e U_n)$] according to the procedures of Section 13.22. When $\hat{a}$ and $\hat{b}$ are found, we then find $\hat{A}$ from the equation $\log_e \hat{A} = \hat{a}$, or $\hat{A} = e^{\hat{a}}$, and insert $\hat{a}$ and $\hat{b}$ in (13.34), thereby obtaining the fitted curve. If the fitted form of (13.34) is plotted on semi-log graph paper, with U measured on the log scale, the resulting graph will be a straight line.

Sometimes, we have to fit an equation of the form

$$V = AU^b , \tag{13.36}$$

where A and b are constants, to n pairs of points (U_1, V_1), (U_2, V_2), ..., (U_n, V_n). By taking logarithms (to the base 10, say) of both sides of (13.36), we have

$$\log V = \log A + b \log U .$$

In this case we put log V = Y, log A = a, log U = X, and the equation becomes

$$Y = a + bX$$

to be fitted by least squares to the n pairs of points (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) [i.e., to the pairs of measurements $(\log U_1, \log V_1)$, $(\log U_2, \log V_2)$, ..., $(\log U_n, \log V_n)$] by the procedures of Section 13.22. When $\hat{a}$ and $\hat{b}$ are determined, the fitted form of (13.36) is

$$V = \hat{A}\, U^{\hat{b}}$$

where $\hat{A} = 10^{\hat{a}}$.

If the fitted form of (13.36) is plotted on log-log graph paper, the resulting graph is a straight line.

13.44 Multiple linear regression.

The ideas discussed in the foregoing paragraphs can be extended to situations in which we may wish to estimate a measurement from several other correlated measurements rather than one. For example, suppose we have n triples of measurements (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), ..., (X_n, Y_n, Z_n), and that we wish to find a regression function of X and Y for estimating Z. The simplest type of a regression function is a linear one of the form

(13.37) $$Z = a + bX + cY .$$

The constants a, b, c can be determined by least squares, just as a, b, c were determined in fitting the form $Y = a + bX + cX^2$ in Section 13.42. In fact, if we replace X^2 by Y and Y by Z in $Y = a + bX + cX^2$, we have (13.37). We shall not carry out the details. The procedure extends to sets of 4, 5, 6 or any number of measurements, where one wishes to estimate one of them from the remaining measurements. In the case of 3 measurements, we may think of our n triples of measurements as n points plotted in 3 dimensions—thereby obtaining a 3-dimensional scatter diagram. When we fit (13.37) by least squares, we are fitting a plane to the 3-dimensional scatter so as to be able to do the best job of estimating Z for any triple of measurements, knowing X and Y.

Such regression analysis is called multiple linear regression analysis and has many technical ramifications which are beyond the scope of this course. The basic principle of fitting such regression functions is simple: it is

based on least squares. Such analysis is used extensively in the analysis of psychological tests, economic analysis, etc. The main problem in fitting such functions is to make sure that the 3-and higher dimensional scatter diagrams are not curved or twisted around so that they cannot be fitted by such functions. This problem is not very serious in some of the fields referred to, particularly psychological tests.

Exercise 13.4.

1. In determining volume of red cells in defibrinated beef blood by a certain method, the following results were obtained for various dilutions of the blood in its own serum, tests being made on three specimens of blood at each dilution (McLain data):

Percent of whole blood in mixture X	Percent of volume occupied by red cells Y
100	40.5
100	41.5
100	46.0
90	35.5
90	40.5
90	41.0
80	32.0
80	36.0
80	36.5
50	20.0
50	22.0
50	23.0

Fit the line $Y = bX$ to this data by least squares. Graph the data and fitted line . What is the interpretation of b?

2. Fit a regression equation of the form $Y = a + bX + cX^2$ to the data of Problem 3 in Exercise 13.1.

3. Suppose measurements $Y_1, Y_2, \ldots, Y_n$ are taken from a population made at times $X_1, X_2, \ldots, X_n$. We may then think of $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ as pairs of measurements. If the regression line $Y = a$ is fitted to these points by least squares, show that $\hat{a} = \bar{Y}$ and that the variance of the estimate s_E^2 of Y from X using this regression line is s_Y^2.

4. The dry weights (in grams) of chick embryos from ages 6 to 16 days were found to be as follows (data from Snedecor):

Age in days X	Dry weight of embryo U
6	0.029
7	0.052
8	0.079
9	0.125
10	0.181
11	0.261
12	0.425
13	0.738
14	1.130
15	1.882
16	2.812

Fit a regression function of the form $U = Ae^{bX}$ by least squares.

5. In a certain experiment, the following values were obtained for the volume of a gas for various pressures:

Pressure in kg/cm^2 P	Volume in litres V
0.5	1.62
1.0	1.00
1.5	0.75
2.0	0.62
2.5	0.52
3.0	0.46

Fit a regression function of the form $P = AV^b$ to this data by least squares.

6. Eight students made the following scores on Test I, Test II and the Final Examination in a certain mathematics course:

Test I X	Test II Y	Final Examination Z
43	22	66
38	29	38
27	23	55
28	33	63
35	20	25
21	8	17
19	17	33
13	19	18

Fit the linear regression equation $Z = a + bX + cY$ to this data by least squares. What estimate would you make for the Final Examination score for a student who made 40 on Test I and 25 on Test II?

7. The following measurements were made on 10 aluminum die castings:

Hardness in Rockwell's E X	Density in $gm./cm^3$ Y	Tensile strength in 1000 lb/sq.in. Z
53.0	2.67	29.3
70.2	2.71	34.9
84.3	2.87	36.8
55.3	2.63	30.1
78.5	2.58	34.0
63.5	2.63	30.8
71.4	2.67	35.4
53.4	2.67	31.3
82.5	2.72	32.2
87.3	2.61	33.4

Fit a regression function of the form $Z = a + bX + cY$ to this data by least squares. What tensile strength would you estimate for a casting with $X = 80.0$ and $Y = 2.65$?

INDEX