SPECIAL TOPICS IN ELEMENTARY MATHEMATICAL STATISTICS

By

WAYNE F. HAYES

Bachelor of Science
Southwestern State College
Weatherford, Oklahoma
1960

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1965

Submitted to the faculty of the Graduate College of the Oklahoma State University in partial fulfillment of the requirements for the degree of
DOCTOR OF EDUCATION
May, 1967

SPECIAL TOPICS IN ELEMENTARY

MATHEMATICAL STATISTICS

Thesis Approved:

H S Mendenhall

Thesis Adviser

Milton E Berg

W Ware Marsden

Jeanne Agnew

A D Durham

Dean of the Graduate School

ACKNOWLEDGMENT

I am extremely grateful for the time, patience, and many considerations given me by the chairman of my advisory committee, Professor H. S. Mendenhall, and the other members, consisting of Professors Milton Berg, Jeanne Agnew, Kenneth Browne, and especially to Professor Robert Hultquist for his encouragement and many constructive comments.

Additionally, I am deeply aware of, and most appreciative for, the many sacrifices my family has made in order for me to have the opportunity to prepare this dissertation.

TABLE OF CONTENTS

CHAPTER I

PROBABILITY AND BASIC CONCEPTS

1.1 Introduction

An individual's approach to probability depends to a great extent upon his interest in the subject. A pure mathematician might rely heavily upon the axiomatic approach, while the applied statistician may prefer to take the intuitive approach. The latter might attempt to consider probability as the proportion of times that a certain event will occur if the experiment related to the event is repeated indefinitely. The approach which we shall pursue here is a blend of these two points of view. We shall attempt to present the basic concepts of probability intuitively through example, but we shall also strive to introduce certain topics without the loss of mathematical rigor.

In most scientific studies the inductive procedures are employed to a great extent. By the inductive procedure, we mean studying particular cases and trying to draw generalizations from them. An example of inductive reasoning is: All sheep which I have seen are white; hence, all sheep are white. Although we see a great number of faults in this process, one will soon discover that the structure of this

kind of thinking is basic to all scientific thought. Scientific knowledge consists of generalizations which are based on observation and experimentation. We can see how limited we would be if we did not employ the inductive procedure. We might be able to report what we have observed and measured, but we would never be able to put this information to work. Hence, if we are to learn from experimentation and use our knowledge of the past for predictions, we must face the gamble which is intrinsic in inductive statistics and the scientific method.

After considering the preceding remarks we see that associated with any inductive process there is almost certainly a degree of risk. Our purpose is to develop methods which will help to minimize this uncertainty. This is where probability will come to our aid. By using probability we will be taking a calculated risk rather than trying to play the role of a fortune teller.

Realizing it is almost impossible to understand the factors involved in chance, which form the basis of inductive statistics, without a sound understanding of the concept of probability, we shall devote a considerable amount of the following discussion to the meaning of probability.

1.2 Probability

When any discussion of probability is endeavored, the most annoying obstacle is the multiplicity and vagueness of meaning which everyday language has given to such words as

"possible", "probable", "likely", and "chance". As long as we are engaged in conversational language it may not matter whether we use these vague terms which may at times add a certain degree of color to our discussion. However, if we are in need of precise statements which we encounter in almost any study of mathematics, we must limit ourselves to well defined terms. It should also be remembered that the names we use for concepts are really irrelevant, the paramount thing being that we have a true understanding of the concepts or ideas for which they stand.

The term "probability" is used for the important concept of relative frequency, or more precisely, the limit of a relative frequency. The usual definition of relative frequency is given as follows: If an event can lead to the occurrence of N equally likely results of which S are denoted as successes, the probability of a success is given by the ratio $\frac{S}{N}$. This so-called definition has some very obvious shortcomings, since the term "equally likely" is also defined in terms of probability. If two events are said to be equally likely, this is usually meant to imply that they are equally probable (they have the same probability), and consequently we are using in this definition the word which we are trying to define. Even though we can not accept this as a definition, it does help us a great deal in calculating probabilities once we know or have assumed the various alternatives are equally likely.

While considering the discussion given above, it is

evident that the proportion of successes can never be negative, and since it can also never exceed unity, the probability of an event is between 0 and 1, inclusive. A probability of 0 does not mean that the event is beyond the realm of possibility, we usually understand this to mean it is extremely unlikely. For example, it can be shown the probability is 0 that a point selected in a random fashion from the interval from 0 to 1 will represent a rational number, even though it is conceivable that the point could represent a rational. Similarly, the probability of 1 would be attached to the event of selecting an irrational even though it is not beyond the realm of possibility that the point would represent a rational number.

Probability can be considered as a substitute for certainty and truth. By a substitute for truth we mean that using probability we can not generally make statements which are always true, we can only make statements which are usually true. Let us illustrate this concept with an example. Let us suppose that you are a farmer and you have planted a certain crop. You decide to consult an agricultural expert about your chances of its being a success. If the expert could tell you for sure that your crop will be a success, you would know exactly where you stand. Similarly, a negative reply would also tell you exactly what to expect. However, the kind of reply which you will receive will merely tell you that your chances for a success may be pretty good, average, or fairly poor. Now,

since these terms are fairly vague, you may decide to ask him to try to explain this in terms of probability. Let us suppose, for instance, that he tells us the probability that you will have a successful crop is .90. Now where do you stand? Would this entitle us to say that the expert was right if the crop succeeds, wrong if the crop fails?

In order to answer this question, let us try to determine what the expert meant by a probability of .90. According to the discussion given above, a probability of .90 means that in the long run something can be expected to happen about 90 per cent of the time. Consequently, when the expert told us that the probability of a successful crop was .90, he meant to say that among a large number of similar crops and under similar conditions such as weather, etc., about 90 per cent can be expected to succeed.

The implication given above is that when we are discussing the probability of a certain event, we must refer to what will happen in the long run in a large number of similar events. There is some objection to this concept in the absence of absolute certainty or absolute truths. But if we want to be scientific, i.e., if we want to obtain knowledge from observations and experiments, we must resign ourselves to the fact that almost all scientific predictions are of this type.

One might get the impression from the above that it is perfectly safe to make scientific predictions, since no one can prove us right or wrong on the basis of a single event.

When we are told that there is a 90 per cent chance of an event taking place, this simply says that under similar conditions the event will occur 9 out of 10 times. It says nothing about what will happen for any given event. This does not imply that we should go about making wild predictions about a single event. It should be kept in mind that it is important in everyday life as well as in the scientific realm to make correct decisions as often as possible. It is necessary in each case to know the proper odds, i.e., which are the correct probabilities. For example, suppose we knew that the probability was .30 that we would catch a cold if we were to go quail hunting on a very cold day while not properly dressed, but the probability was .10 that we would catch a cold if we wore the proper attire. We would probably play the odds and wear the proper dress. It should be evident that this would not protect us from catching a cold, but in the long run we would catch fewer colds if we wore the proper attire. Thus we would expect a larger per cent of success if we were to rely upon the odds. From this discussion it is evident that even though probability does not guarantee success, it acts as a very important guide in life to help in the long run to enjoy more success.

In the last few paragraphs we have mentioned prediction and estimation, etc. Let us consider an example of how we might use the idea of prediction. Suppose you are attending a large school and the administration has come to

the realization that they should let the student body decide upon a certain proposal by election. You have a statistician friend who lives in the same city. So you decide to try to predict the outcome of the election. You confer with your friend and he explains how to proceed. You follow his instructions and take a sample of opinions as suggested, and find 60 per cent of the people you ask favor the proposal. You then give the information to your statistician friend. He studies the data and proceeds to analyze it in a statistical manner. He reports that he is 95 per cent confident that the proposal will carry. Now your friend has assigned a probability estimate of .95. He is telling you that using his procedure he can expect to be successful about 95 per cent of the time. In other words, we discuss the accuracy of our results by giving the success ratio of the methods which we have employed.

In remaining chapters dealing with estimation, the "goodness" of decisions which is based on information we have at our disposal, usually a sample, shall be expressed in terms of the success ratio of the statistical techniques which we have employed. For example, suppose that two students take the entrance examination at a certain college. If from the results of this examination we could be 90 per cent confident that one student was better than the other, then we again imply that the statistical method employed promises to provide correct decisions about 90 per cent of the time. Hence from now on the probabilities which shall

be assigned to the results of predictions and estimators will therefore always express the goodness of the methods employed. Another way of expressing this is that the probabilities will stand for the proportion of times that we can expect these methods to present us with the correct results if the methods are employed a large number of times.

In the preceding discussion we have tried to introduce probability intuitively. Let us now consider probability from a somewhat more mathematical point of view. Suppose that you are fairly proficient in the shooting of a shotgun and you decide to enter a trap-shoot sponsored by the local gun club. Suppose also that each time you shoot you receive a "1" marked on the score card if you hit, and a "0" if you miss. Let us consider this as an experiment and represent the point 1 on the x axis if the clay pigeon is hit and 0 for a miss. We now ask what are possible outcomes of the experiment? It is easily seen that the only possible outcome is a zero or a one. These outcomes of an experiment are called the sample space. We now formalize the definition.

Definition 1.1 The set of points representing the possible outcome of an experiment is called the sample space, or the event space, of the experiment.

Let us now consider some of the basic rules of probability which will be useful in the following chapters.

From the discussions given earlier, it should be evident that if we denote the probability of the occurrence of an event A by P(A) then $0 \leq P(A) \leq 1$ which expresses the fact that probabilities must be between 0 and 1, inclusive. Another basic rule of probability which is immediate is that if the probability of an event A is P(A) then the probability that A does not occur is 1-P(A). This means that the probability that A will happen plus the probability A will not happen is one. For example, if we are 90 per cent sure of passing a test then the probability we will fail is 10 per cent; if the probability is .60 that a team will win a certain game then the probability the team will lose is .40.

Let us now turn our attention to problems which arise in studying probability where more than one event can occur simultaneously. When studying the occurrence of more than one event we must consider the concept of events being mutually exclusive. Two or more events are said to be mutually exclusive if they cannot occur at the same time. For example, in our illustration you would either hit a clay pigeon or would not hit it when you shoot, you cannot hit it and also miss it simultaneously. Hence, the events of hitting or missing the target are mutually exclusive. Of course not all events are mutually exclusive. Suppose you are in the process of buying a new car, and you must decide between a Ford or a Chevrolet. Say the event is your getting a car. Since you could conceivably buy both, these

events are not mutually exclusive.

Continuing our discussion of mutually exclusive events, suppose the probability that a person will enroll at Oklahoma State University is known to be .40, while the probability is known to be .30 that he will enroll at the University of Texas. It seems reasonable that he cannot enroll in both universities, hence these events are mutually exclusive. Also the probability that the person will enroll at Oklahoma State University or the University of Texas is the sum of the individual probabilities. This type of reasoning leads us to the next useful concept. If two events, A and B, are mutually exclusive, the probability of A or B, written P(A or B), is equal to the sum of the individual probabilities, i.e., P(A or B) = P(A)+P(B).

Our discussion thus far has been primarily centered around the concept of mutually exclusive events. Let us now turn our attention to events which are not mutually exclusive, such as the events of wearing the proper clothing in very cold weather and of catching a cold, or the events of cold weather and of snowfall. We see at once that these events are not mutually exclusive. In fact, it seems conceivable that the events are very much dependent on one another. Let us illustrate the meaning of dependence and independence of two events before we give a formal definition. An event B is independent of another event A if the probability of the occurrence of B is the same regardless of whether A has previously occurred or is occurring

at the same time. If, on the other hand, the probability of B is in any way affected by the results of what happened in A, the two events are said to be dependent.

The probability that a student will make an A on a test is surely very much dependent on the amount of preparation he has made for the test. However, the probability that he passed the test is not dependent upon whether he uses cream in his coffee the morning before the test. In real life it is usually very hard to find events which are completely independent of every other possible event. For in the example of independence just stated, if someone had put poison in the cream than it would effect his passing the test that day. Therefore, when we speak of events being independent we are assuming that the type of phenomena as mentioned above will not happen. Let us consider another example to illustrate dependence. Suppose we have a hat containing four pieces of paper, two labeled with the letter "a" and two with the letter "b". Event A is the drawing of a slip of paper and looking at it and not replacing it in the hat. Suppose we get an a, then if P(B) is the probability of drawing a b, we see that the probability is different before we drew the first piece of paper. Hence the event B is dependent upon the event A. We shall now formalize the definition of the independence of only two events, which could be extended to any number of events.

Definition 1.2 The two events A and B are independent if, and only if, the joint probability, written P(A and B) or P(A,B), is equal to the product of the individual probabilities, i.e., $P(A,B) = P(A)\ P(A)$.

So if we know two events are independent and we want to find the probability of the occurrence of both events, simultaneously, we find the product of the individual probabilities. For example, returning to the discussion of the man shooting clay pigeons, we can answer the question, "What is the probability that he will hit two clay pigeons in a succession?", assuming the probability of his hitting a pigeon is 9/10. The event of his hitting the second time would not depend upon his hitting the first time, i.e., the events would be independent. Hence, employing the above definition we see the probability of hitting two clay pigeons consecutively is (9/10)(9/10) which equals 81/100.

We must be careful and not apply the above definition to events which are not independent. Let us consider the two events, A and B, where A is the event it will snow today, and B that the highways will be slick. It is obvious that the two events are dependent, i.e., if it snows the probability that the roads will be slick is much higher than if it does not snow. This type of problem leads us to the introduction of the concept of conditional probability. The probability that the event B will happen provided that A has taken place, is called the conditional probability of B

relative to A and is denoted by $P(B \mid A)$. Using this concept we can define the joint probability of A and B as follows:

Definition 1.3 If A and B are two events then the joint probability of A and B is given by $P(A \text{ and } B) = P(A)P(B|A)$.

It should be noted that this holds also if A and B are independent, since if A and B are independent $P(B|A) = P(B)$. For instance, we are given the event A that a man has received 5 traffic citations in the past 6 months and B the event that he will receive a citation in the next 6 months, assuming he continues driving. It seems conceivable that the event B, i.e., $P(B|A)$ would be much higher than $P(B)$ where the man was just an ordinary driver. The latter makes no reference to the man's past driving record.

Before concluding this brief discussion on probability, we shall illustrate another use of a relationship which exists between two events A and B. To help illustrate the relation to be given, suppose $P(A)$ and $P(B)$ represent the probability that a certain student will make the baseball and basketball teams, respectively, in a small school. It seems conceivable these two events are dependent, i.e., if a person is a good athlete he would probably excel in several sports. The question might arise, "What is the probability that he will make at least one of the teams, i.e., what is $P(A \text{ or } B)$?". Before answering this question, let us prove the following theorem.

Theorem 1.1 If A and B are any two events in the sample space S, then

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

Proof:

Now by A or B we mean $A \cup B$ and $A \cup B = A \cup (\bar{A} \cap B)$, where $\bar{A}$ is the set of points in S which are not in A. However, A and $\bar{A} \cap B$ are disjoint so we have $P(A \cup B) = P\left[A \cup (\bar{A} \cap B)\right] = P(A) + P(\bar{A} \cap B)$. Now $B = (A \cap B) \cup (\bar{A} \cap B)$, and the two sets $(A \cap B)$ and $(\bar{A} \cap B)$ are disjoint. Hence, we have $P(B) = P\left[(A \cap B) \cup (\bar{A} \cap B)\right] = P(A \cap B) + P(\bar{A} \cap B)$ or $P(\bar{A} \cap B) = P(B) - P(A \cap B)$. Substituting $P(A \cap B)$ into the equation above gives $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. But $P(A \cap B) = P(A,B)$. Thus we have the desired results, i.e., $P(A \text{ or } B) = P(A) + P(B) - P(A,B)$.

To illustrate this theorem suppose in the preceding example we have $P(A) = .70$, $P(B) = .80$, and $P(A,B) = .63$. From the theorem above we have $P(A \text{ or } B) = P(A) + P(B) - P(A,B)$. For the example given, the probability $P(A \text{ or } B) = .70 + .80 - .63 = .87$.

1.3 Finite Sums and Products

Many times in the following chapters we shall be in need of an expression for the sum of a certain quantity. For example, suppose we are given two hundred numbers $N_1, N_2, \ldots, N_{200}$ and we would like to express their sum. We shall designate the sum of these two hundred numbers by

$\sum_{i=1}^{200} N_i$. Here, Σ is the Greek capital letter sigma, and in this connection it is often called the summation sign. The letter i is called the summation index, while the term following the Σ is called the summand. The 1 below Σ indicates that the first term of the sum is obtained by putting i=1 in the summand. The 200 above the Σ indicates that the last term of the sum is obtained by putting i=200 in the summand. The other terms of the sum are obtained by giving i the integral values between 1 and 200. We can see how this notation can save time and space when we are concerned with writing the sum of a finite collection of terms. We should also note that one of the properties of a finite sum is $\sum_{i=1}^{n} (N_i+M_i) = \sum_{i=1}^{n} N_i + \sum_{i=1}^{n} M_i$, i.e., we can distribute the Σ over the finite sum.

Many times we are concerned with sums over sets which are not finite. Generally, we will be summing over a countable set, that is, a set which can be put in a one to one correspondence with the set of positive integers. We shall assume $\sum_n (a_n + b_n) = \sum_n a_n + \sum_n b_n$, since in all cases we will be dealing with the a_n and b_n will be greater than or equal to zero and the condition which must be placed upon $\sum_n a_n$ and $\sum_n b_n$ is that $\sum_n a_n$ and $\sum_n b_n$ must converge, that is, the sum over n must be some finite number. These conditions will generally be satisfied in developing the

following theory since we will usually be summing over n where a_n represents a function whose sum over n is equal to one.

In several situations, to save space, we shall choose an analogous notation for a product by using the capital Greek letter $\prod$ instead of Σ in the sums. In this case the terms resulting from substituting the integers for the index are multiplied instead of added. For example,

$$\prod_{i=1}^{6} a_i = a_1 a_2 a_3 a_4 a_5 a_6.$$

1.4 Random Variable and Probability Functions

In several examples we have used thus far, we usually associated a number with the outcome of an experiment, as in the example of the trap-shoot, where we associated the real number 1 with a success and the real number 0 with a failure. Let us try to find a function that will give some relation between the outcome and the probability of one of the events occurring. Suppose we know that the probability of a hit is, say p = .9, while the probability of a miss is 1 - p = .1. We need to find a function of x and p which will give the probability p of a given x happening. Consider $f(x) = p^x (1-p)^{1-x}$, x=0, 1. Now $f(1) = p^1(1-p)^0 = p$, hence, the probability that x = 1 is given by p(x = 1) = f(1), and p(x = 0) = f(0) = 1 - p. Thus this function gives us the desired probabilities

which we mentioned at the outset we wanted. The function $f(x) = p^x(1-p)^{1-x}$ is called the probability function of X if certain conditions hold. Let us give formal definition of these concepts.

Definition 1.4 Let S be a sample space and X a real valued function defined on S. Then X is called a random variable. X is a discrete random variable if it assumes a finite or countable number of points. X is a continuous random variable if it assumes an uncountable number of points.

After a random variable X has been defined on a sample space, interest usually centers on determining the probability that X will assume specified values in its range. The relationship between the value of X and its probability is expressed by means of a function called the frequency or probability function, which is defined as follows:

Definition 1.5 A function f(x) that yields the probability that the discrete random variable X will assume any particular value or set of values in its range is called the frequency (probability) function of the random variable X. If X is a continuous random variable then f(x) is referred to as a density function.

Thus far in our discussion we have considered only one-dimensional probability functions. Many experiments involve several random variables rather than just one such

variable. The definitions concerning probability functions which involve more than one random variable are a straight forward extension of those of a one-dimensional random variable.

Definition 1.6 A function $f(x_1,x_2,\ldots,x_n)$ that yields the probability that the random variables $X_1,X_2,\ldots,X_n$ will assume any particular value or set of values in their range is called the joint density (probability) function of the random variables $X_1,\ldots,X_n$.

Also since we will be dealing with the joint probability functions which contain more than one random variable it behooves us to define the joint probability function of n independent random variables. This follows analogously to the probability of independent events.

Definition 1.7 If the joint probability function $f(x_1,x_2,\ldots,x_n)$ can be factored in the form $f(x_1,x_2,\ldots,x_n) = f_1(x_1)\ f_2(x_2),\ldots,f_n(x_n)$, where $f_i(x_i)$ is the probability function of X_i, then the random variables $X_1,X_2,\ldots,X_n$ are said to be independently distributed.

A function closely related to the probability function $f(x)$ is the cumulative distribution function. Since in the case of discrete variates the probabilities are given by sums, it often is convenient to deal with the sums of the

probability functions rather than the probability functions themselves. Suppose for example that X is the number of tosses required to obtain a head with an ideal coin. Then the probability function is $f(x) = (\frac{1}{2})^x$, $x = 1,2,\ldots$. Now the $p(1 \leq X \leq 2) = \sum_{x_i=1}^{2} (\frac{1}{2})^{x_i} = (\frac{1}{2})^1 + (\frac{1}{2})^2 = 3/4$. Now consider the $p(X \leq x) = \sum_{x_i=1} (\frac{1}{2})^{x_i} = F(x)$.

$F(x)$ is the probability that the value of the random variable will be less than or equal to x. F is called the cumulative distribution function of X. F is defined similarly for a continuous random variable except in terms of integrals. A useful property of F is as follows:

Theorem 1.2 If F is a cumulative distribution function of a random variable then (1) F is non-decreasing

(2) $F(-\infty) = 0$

(3) $F(\infty) = 1$

The concept of a random variable is employed by the statistician in a manner very similar to that in which a mathematician uses the concept of a mathematical variable. Suppose for example we are flipping a coin and we record a 1 if the coin lands with the head up and record a 0 if we get a tail. We have assigned a real number to all the possible outcomes of the given experiment.

In the example above if we let X denote the outcome when the coin was tossed we see that X is a real valued

function defined on the sample space, i.e., X (heads) = 1, and X (tails) = 0 . To consider another example to further illustrate this concept let the random variable denote the outcome from a cast of a die. Now X can take on the values 1,2,...,6, and is therefore a random variable since it is a real valued function defined on the sample space. It seems obvious that if this die is fair there should exist some probability that we can attach to the event of getting, say a 1, when the die is cast. Now the probability of getting a 1, written $p(X = 1)$ is equal to 1/6.

When such a function $f(x)$ exists such that $f(x) = p(x=X)$ we often say that X is distributed as $f(x)$ and we write $X \sim f(x)$. In the example above concerning the casting of a die we see that $f(x) = 1/6$ $(x=1,2,..,6)$ since $p(X=1) = p(X=2)... = p(X=6) = 1/6 = f(x)$. Therefore we have a function $f(x) = 1/6$ which gives us the probability that the random variable X take on a specific value x.

Throughout the remaining chapters we shall in many instances be given a sample of size n, say $X_1, X_2, .., X_n$ and we will be computing such functions as the mean of the sample $\bar{X}$, the sample variance $s^2 = \frac{\Sigma (X_i - \bar{X})^2}{n}$ and other functions of observed random variables.

Definition 1.8 A function $g(X)$ of observed random variables which contains no unknown parameter is called a statistic.

After considering the basic concepts of random vari-

ables and probability functions, let us now consider an illustration in which we define a random variable and through repetition of an experiment try to determine the probability function for the random variable.

Suppose you are a member of an artillery group in a branch of our armed forces. Suppose further, that your group's assignment is usually attacking convoys of trucks, trains, or regiments of men, i.e., your targets are usually objects which have considerable length. We see that to hit the target we must be accurate vertically, while horizontally we can score a hit even though we are not very close to the center of the target. So it is important to be sure that we have our gun adjusted at a correct elevation. Suppose we conduct an experiment where our target is a certain line at a given distance perpendicular to the gun. Now the experiment will consist of measuring the distance the shell falls from the line, where a shot that falls short of the distance will be a negative number and an over-shot will be a positive number. If we are on the firing range for several days and keep track of how far each shell falls from the line, it is conceivable there would be a large concentration of shots close to the line you were trying to hit. Suppose that you take these measurements for several days and keep track of them. Let us group them so that the ones which are within 5 yards of the given line and those at a distance of between 5 and 10 yards from the line, are together for each 5 yards. We would probably get a histogram

similar to the one in Figure 1, where X is the distance the shells fell from the given line and the vertical axis represents the number of X's appearing in any given interval.

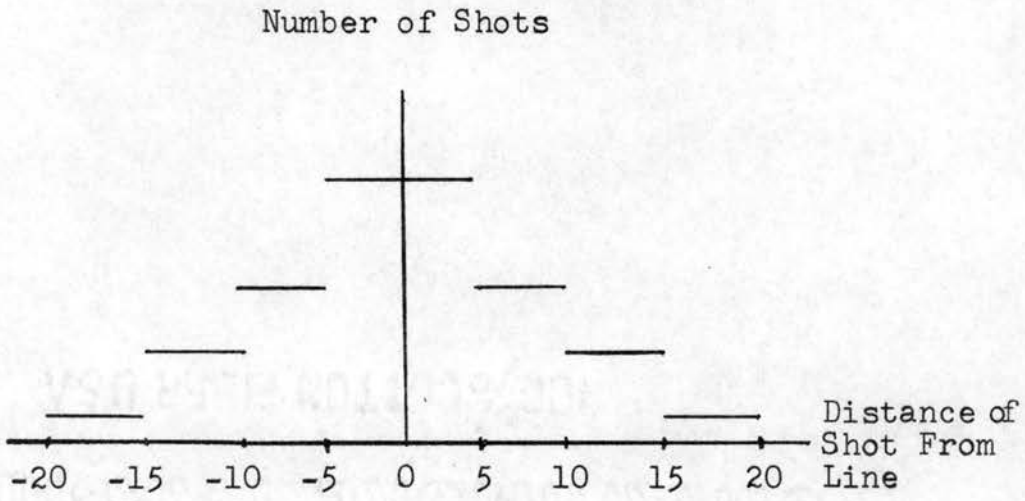

Figure 1

If after each day's shooting you make a histogram similar to the one in Figure 1, we would after a month or so have several histograms. Now if we would plot all of these different histograms on the same graph, more or less on top of each other, we would get a drawing similar to the one in Figure 2, which could be approximated by a bell shaped curve similar to the one shown there.

Now after a large number of similar experiments we could make fairly accurate probability statements, such as, the probability that X is between -5 and 5. The proportion of the shots in any given interval would be the total shots in the interval over the total shots. So the probability

that X is in an interval is related to the area bounded by the curve and the given interval, since the number of observations which appears in a given interval is reflected by the area of the rectangle which represents the frequency. So if we say 60 per cent of the shots fall in -5 to 5, we would be inferring that the area bounded by the rectangle, with base on -5 to 5, has about 60 per cent of the area of all the given rectangles considered together which is 100 per cent or has an area of 1.

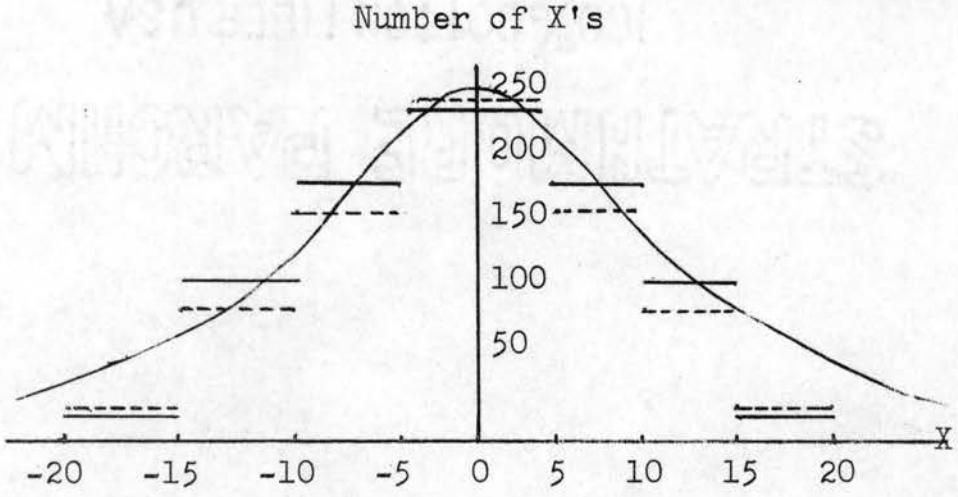

Figure 2

It is conceivable that the random variable X could take on any real number. And the graph of the frequency function for X would approximately be the curve given in Figure 2. Bell shaped curves of this type are called normal curves and the density function, f(x), is given by

$$f(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{\frac{-\frac{1}{2}(x-\mu)^2}{\sigma^2}}$$ where x can be any real

number and σ^2, which is called the variance of X, is non-negative, while μ represents the mean, which can also be any real number. In the example, we were using the mean, the reference line, as 0.

In the paragraph above, we mentioned the variance of X. Let us devote some time to this concept. If, as in the example, you were using very superior shells and your guns were in very good condition, you would expect most of the shells to hit very close together, i.e., there would be very little variation among the X's. However, if the shells were of poor grade, say they had got wet in the process of being shipped to the military post, then one shell might be very good while the next one might go only half the distance to the target. So we see that the values of X would have a large amount of variation. It seems likely that when the variability among the X's is large then the normal bell-shaped curve would be lower since a large number of the X's would be at a greater distance from the mean. The most widely used measure of variation based upon a sample is the so-called standard deviation, which will be denoted by the letter s. This statistic reflects a large difference among the X's and its square is written as

$$s^2 = \frac{\sum_{1}^{n} (X_i - \overline{X})^2}{n-1}$$

where s^2 is an estimate of the variance, and $\overline{X}$ represents the mean of the sample. Now if the mean of the probability

function is known we use for an estimate of the variance the statistic

$$s^2 = \frac{\sum_{1}^{n}(X_i - \mu)^2}{n} .$$

The normal density function is very useful in many problems we shall consider in the following chapters. Many times we shall be concerned with random variables which have a certain probability function which is approximately that of the normal. Hence, in many instances we can apply normal theory in making probability statements which will be fairly accurate if the random variable whose density is approximately that of a normal. We shall consider in the next paragraph a probability function which can be approxmated by the normal.

Suppose you are participating in a trap shoot sponsored by the local gun club. Suppose further that you have been engaging in this sport for several years and you are sure you hit about 50 per cent of the clay pigeons. On this particular day the only type of contest in which you engage is where two clay pigeons are thrown at once and you have two shots to try to break them. Suppose for this particular day you shoot at 200 clay pigeons (100 different sets of 2 clay pigeons each). The expected distribution of the number of hits will be represented by one-quarter, one-half, and one-quarter of 100, respectively.

The expected distribution of the number of hits in 100 shots would be illustrated by the following histogram.

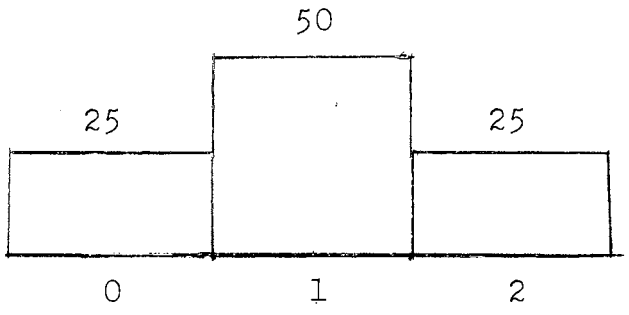

Figure 3

We could have had more clay pigeons thrown at once, but this would complicate our illustration and we could not realistically assume the probability of a hit did not change. Let us suppose that you decide to shoot at 3 pigeons, one at a time. The various events which could happen would be given as follows:

M M M	H H M
H M M	H M H
M H M	M H H
M M H	H H H

We find that we can expect to get 0 hits one-eighth of the time, 1 hit three-eights of the time, 2 hits three-eights of the time, and finally, 3 hits about one-eighth of the time. These probabilities are obtained from the above table by considering favorable outcomes divided by the total number of outcomes. If you shoot 80 times at sets of three clay pigeons, the resulting expected frequency

distribution will be

Number of Hits	Frequency
0	10
1	30
2	30
3	10

which is illustrated by the histogram in Figure 4.

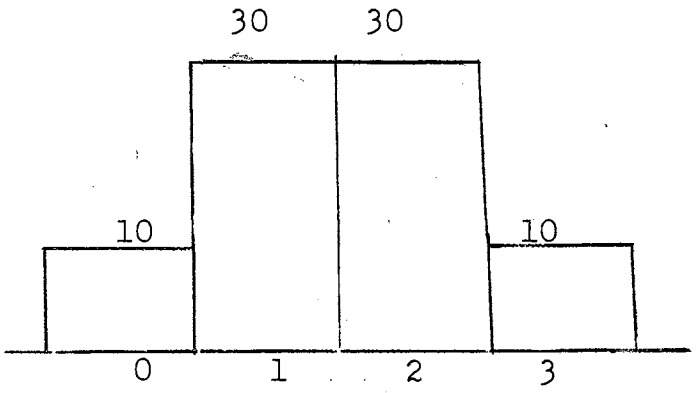

Figure 4

In Figure 4 if X represents the number of successes in a given number of trials then X is said to be distributed as a binomial f(x) given by

$$f(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}, \quad x = 0,1,2,\ldots,n.$$

It seems reasonable that if we repeated this experiment several times that the histogram might be approximated by the normal density with mean of 50. This can be shown by employing some advanced techniques beyond the scope of our

treatment of the subject here. In general it can be shown that the mean of the normal approximation of the binomial is np, and in this case we see the mean is $100(.50) = 50$.

In the following chapters we shall use these results on several occasions to apply normal theory to certain distributions which are approximately normal. It is often much easier to make probability statements concerning a normal variable than probability statements concerning a discrete random variable.

1.5 Sampling

Up to this point we have been concerned with certain aspects of the theory of probability. Let us now give some of the basic concepts and definitions related to the theory of sampling which will be useful in the later development of this theory.

Progress in science is ascribed to experimentation. The research worker performs an experiment and obtains some data. On the basis of the data, certain conclusions are drawn. The conclusions usually go beyond the material and operations of the particular experiment. In other words, the scientist may generalize from a particular experiment to the class of all such experiments. This type of extension from the particular to the general is called inductive inference.

Inductive inferences are well known to be a hazardous process. We must be very careful how we collect the data

we plan to analyze and how we make inferences based on this data. One function of statistics is the provision of techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences. This uncertainty is measured in terms of probability.

Suppose it is desired to estimate the per cent of people who have televisions and who live in cities of one million or more. A person might select a certain city and take a sample of people in the city and determine the per cent of people in that city who own television sets. Before we can draw any valid conclusions we must pick the city and the sample in a certain manner. This line of thinking leads us to the following definition.

Definition 1.9 The totality of elements which are under discussion and about which information is desired will be called the target population.

In the above example the target population consisted of all people that live in cities with population over one million. The problem of inductive inference, from the point of view of statistics, is regarded as follows: the object of an investigation is to find out something about a certain target population. It is generally impossible or impractical to examine the entire population, but one may examine a part, or a sample, of it and, on the basis of this limited investigation, make inferences regarding the target population.

The problem immediately arises as to how the sample of the population should be selected. We can make probability statements about the population if the sample is selected in a certain fashion. Of particular importance is the case where the sample is a random sample, which is defined as follows:

Definition 1.10 Let the random variables $X_1, X_2, \ldots, X_n$ have joint probability (density) function $g(x_1 x_2, \ldots, x_n) = f(x_1) \ldots, f(x_n)$ where the probability function of each X_i is $f(x_i)$. Then $X_1, X_2, \ldots, X_n$ is said to be a random sample of size n from f(x).

Definition 1.11 Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with probability function f(x), then this population is called the sampled population.

Valid probability statements can be made about the sample population based upon a random sample but inferences on the target populations are not always valid.

CHAPTER II

EXPECTED VALUES AND MOMENTS

2.1 Introduction

When attending elementary school, I am sure we were all introduced to the idea of finding the average of a set of numbers. We recall our teacher informed us that to find the average of a set of numbers we would add the numbers and divide by the number of numbers under consideration. Let us investigate this idea in depth and illustrate the connection between the average and the expected value. Suppose we are given the numbers 5,6,4,4,5,6,3,5,7,6,7,3, and 7 and we are asked to find the average. The average of these numbers is given by 5+6+4+4+5+6+3+5+7+6+3+7+7 divided by 13. Let us commute and associate the numbers such that we have all like numbers together. We find the average is equal to

$$\frac{(5+5+5) + (6+6+6) + (4+4) + (3+3) + (7+7+7)}{13}$$

$$= \frac{3(5) + 3(6) + 2(4) + 2(3) + 3(7)}{13}$$

$$= 3/13(5) + 3/13(6) + 2/13(4) + 2/13(3) + 3/13(7).$$

Let us call the coefficients of the numbers we started with $f(X_i)$, and X_i the numbers we started with to find the average. Now if we consider $f(X_i) = n_i/n$ the frequency function of the numbers X_i, then we have an expression for the average of this set of numbers. The average is given by

$$\sum_{i=1}^{13} X_i f(X_i)$$

which we define to be the expected value of X, written E(X).

To help us gain some insight into this concept, suppose a man is engaged in a game of chance; say there are nine cards lying face down and these cards consist of 2 spades, 3 hearts, 3 diamonds and 1 club. Suppose also that these cards are well shuffled and this man is equally likely to draw any one of these cards, i.e., the card will be drawn at random. We see that the probability of drawing a spade is 2/9, the probability of drawing a heart is 3/9, the probability of a diamond is 3/9, and that of a club is 1/9. We could illustrate the relation by the following table where X_1, X_2, X_3, X_4 represent the drawing of a spade, a heart, a diamond, a club, respectively.

X_i	X_1	X_2	X_3	X_4
$p(X_i)$	2/9	3/9	3/9	1/9

Suppose the game consists of the man paying a certain amount of money to get to draw a card and if he draws an X_1 (a spade) he receives \$18, otherwise he receives no prize. The question which we want to answer is,"What is his expected winning or his expectation?". We see his probability of winning is 2/9, hence his expected winnings would be 2/9 (18) + 3/9 (0) + 3/9 (0) + 1/9 (0) = 4, i.e., in the long run he would expect to average winning about \$4 each time he played. We note that the expected value of a random variable need not be any actual value which X can take on.

The concept of expectation is easily extended. If X denotes a discrete random variable which assumes the values $x_1, x_2, \ldots, x_n$ with respective probabilities $f(x_1), f(x_2), \ldots, f(x_n)$ where $f(x_1) + f(x_2) + \ldots, + f(x_n) = 1$, the expected value of X, written E(X), is $x_1 f(x_1) + x_2 f(x_2) + \ldots, + x_n f(x_n)$. We shall now formalize the definition of expected value.

Definition 2.1 Let X be a random variable with probability function f(x). Then the expected value of X, is $E(X) = \sum_x xf(x)$ if X is a discrete random variable. The expected value for a continuous random variable is similarly defined except in terms of integrals.

Many times in practice we are interested in the expectations of some functions of a random variable. The

following discussion will be concerned with an example to help illustrate how we might deal with this problem.

Suppose we are in the business of making bolts. Suppose also that we have a machine which manufactures bolts of one-half inch in length. We know that when we have a number of bolts manufactured by this machine, if we measured these with a very precise instrument, many of these would differ in length from one-half inch by various amounts. Suppose we decide to take a barrel of bolts and divide these into several groups. If the lengths of the bolts were between .49 and .51, they would be in one group, and if their lengths were between .48 and .49 or .51 and .52, and so on. We might expect a large number of these to be in the interval .49 to .51 if the machine is fairly accurate. The histogram might be similar to the following:

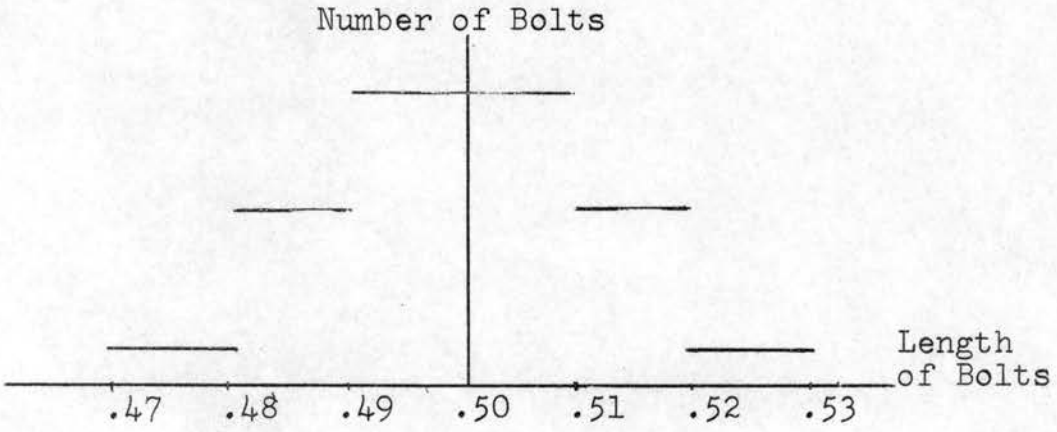

Figure 5

It seems reasonable that this would be approximated by a bell shaped curve with its maximum at the point $x = \frac{1}{2}$.

This type of density function is given by

$$f(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \quad e^{\frac{-\frac{1}{2}(x-\mu)^2}{\sigma^2}}$$

where x can be any real number while σ^2 and μ are parameters, which we shall explain in greater detail later. Returning to the example, it seems reasonable to say that, if we draw a bolt out of a barrel of bolts at random, we would expect to get a bolt of one-half inch in length. Thus we would say that the expected value of X (where X represents the length of a randomly chosen bolt) would be one-half inch.

Now suppose we want to study small differences in the lengths of the bolts. A possible way to do this is to square each value of X_i (Square the length of the bolt). So we take this barrel of bolts and measure each one of them and let $X_i^2 = y_i$. Now this new "barrel of y_i's" has a distribution, that is, we would probably expect a y_i drawn at random to be close to one-fourth. It seems intuitively obvious that $E(X^2)$, i.e., E(y) would be defined very similarly to the expected value of X. We now formalize the definition of the expected value of some function U(X) of the random variable X. This definition is actually redundant since it can be proven by using some advanced techniques. However, it is consistent.

Theorem 2.1 Let X be a random variable with probability function. The expected value of U(X), a function of X, is $E(U(X)) = \sum_x U(x)\, f(x)$ if X is a discrete random variable.

To help understand a practical example, suppose you and some of your friends are engaged in playing a game of monopoly. Suppose you are asked what is the number of spaces you would expect to move on any given roll of the dice. We would probably say 6. Let us find the expected value of X if we are given that X is the number appearing when a pair of dice is cast. To answer this question we must first determine the probability function of X. We shall do this by constructing a table where x_i represents the number of spots appearing on the dice and $f(x_i)$ is the probability of getting x_i on a given cast of the dice.

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

We found $f(x_i)$ by determining how many possible ways we could get the number x_i from the two numbers appearing on the dice divided by the total number of possible outcomes. Now employing definition 2.1, $E(X) = \sum_x xf(x) =$

$$(2)(1/36)+(3)(2/36)+(4)(3/36)+(5)(4/36)+(6)(5/36)+(7)(6/36)$$

$$+(8)(5/36)+(9)(4/36)+(10)(3/36)+(11)(2/36)+(12)(1/36).$$

Hence $E(X) = \frac{2+6+12+20+30+42+40+36+30+22+12}{36} = \frac{252}{36} = 7.$

Thus the expected value of X is 7, not 6 as we had thought it might be. From this we see that on the average we would move 7 spaces each move.

We might also ask,"What is the expected value of 2X?". We are asking the question,"What would we expect 2 times an X value to be?" It appears that it should probably be 2E(X). Let us compute E(2X) using theorem 2.1. $E(2X) = \sum_x 2xf(x) = 2\sum_x xf(x) = 2E(X) = 2(7) = 14$.

Before considering more examples let us consider how we might construct a probability function. Suppose you are watching someone shooting a basketball at the goal. On any given shot he either hits the basket (the ball passes through the goal) or he misses. This operation describes a probability function if we let X = 1 if he hits the basket and X = 0 if he misses. If we watched him shoot at least 100 times each day for one year and each time he hit a basket we threw a small ball with a one on it into a large container, and each time he missed we threw a ball with a zero on it into the container, we would have a population of zeros and ones. If we drew a ball from this large container at random there would be some probability associated with it being a one or a zero. Let p be the probability that X = 1, then the probability of a zero is 1 - p. This probability function is given by f(x) =

$p^x(1-p)^{1-x}$, $x=0,1, 0 \leq p \leq 1$. Let us investigate this probability function further.

Example 2.1 Let the random variable X have the probability function $f(x) = p^x(1-p)^{1-x}$, $x=0,1, 0 \leq p \leq 1$. What is the expected value of X, E(X)? Employing definition 2.1, $E(X) = \sum_x xf(x)$

$$= \sum_0^1 x\, p^x(1-p)^{1-x} = 0 \quad p^0(1-p)^1 + 1p^1(1-p)^0 = 0+p = p.$$

Hence the E(X) for this probability function is p. What is the $E(X^2)$? Using theorem 2.1, we see $E(X^2) = \sum_x x^2 f(x) =$

$$\sum_{x=0}^1 x^2\, p^x(1-p)^{1-x} = 0 \quad p^0(1-p)^1 + 1\, p^1(1-p)^0 = p.$$

A point concerning our terminology is now in order. When considering problems concerning discrete random variables, we discussed a probability function. When dealing with continuous random variables, such as the normal, we shall speak of the probability density function of the random variable. The random variables which will be considered as having a density function are precisely those whose cumulative distribution functions have a derivative at each point, and the derivative of the cumulative distribution function is the density function.

2.2 Moments

Let us return to the example we discussed earlier concerning the manufacturing process which consisted of

making bolts one-half inch in length. Suppose you are in charge of operations at this factory and you have instructed one of your subordinates to take a sample of bolts manufactured by this machine once each day. You have explained to him that he is to find the average lengths of these and report to you if the average deviates more than .1 inch from one-half inch in length. Suppose everything goes well for a month, i.e., we have no report of the average deviating from one-half inch by more than .1. But after our bolts get on the market we have several complaints that they differ in length by a great amount. So we decide to investigate the situation ourselves. We take a sample of bolts and compute the average. To confuse the situation more, we get an average of .495 which is acceptable under our requirements. So you examine a few bolts and find several are three-fourths an inch long and several are about one-fourth inch in length. This leads us to suspect that we need some other criteria for determining whether or not our manufacturing process is acceptable. So we search for a statistic which will serve to help us detect a large amount of variability among the bolts. This leads us to the next statistic we shall consider. We could use a statistic $M=1/n \sum_{i=1}^{n} (X_i-\frac{1}{2})^2$, where X_i represents the length of the ith bolt in the sample of size n. We can see that if M is small (of course, the word "small" would depend upon how large a sample we took and also how much deviation we would allow) then we could conclude that most of the

bolts in the sample are close to one-half inch, while if M were a large number we would conclude that the process should be investigated and possibly be changed. The preceding statistic is called second sample moment about the mean. We shall now formally define some of the concepts introduced.

Definition 2.2 The rth moment of a random variable X, usually denoted by μ_r', is defined as $\mu_r' = E(X^r) = \sum_x x^r f(x)$, where f(x) is the probability function of X. And $\mu_r = E(X-\mu)^r = \sum_x (x-\mu)^r f(x)$ is defined to be the rth moment about the mean.

We note that the first moment, i.e., where r is equal to one, μ_1', is just the expected value of X as defined earlier in this chapter. The concept of variation is of paramount importance in statistics as was indicated in the previous example concerning the manufacturing process. The second moment about the mean is a measure of variation among the random variables and is called the variance of X.

Let us investigate the first and second moments and observe some properties which will be useful in our study. First let us consider the first moment about the mean, usually denoted by μ_1 and given by

$$\mu_1 = E(X-\mu) = \sum_x (x-\mu) f(x)$$

$$= \sum_x xf(x) - \mu \sum_x f(x)$$

$$= E(x) - \mu$$

$$= \mu - \mu$$

$$= 0$$

since $\sum_x f(x)$ is one and $E(X)$ is the mean. Thus we see the first moment about the mean is zero. Also the second moment about the mean is given by

$$\mu_2 = E(X-\mu)^2 = \sum_x (X-\mu)^2 f(x)$$

$$= \sum_x (x^2 - 2x\mu + (\mu)^2) f(x)$$

$$= \sum_x x^2 f(x) - 2\mu \sum_x xf(x) + \mu^2 \sum_x f(x)$$

$$= \mu_2' - 2\mu^2 + \mu^2$$

$$= \mu_2' - \mu^2 \quad .$$

So the second moment about the mean, called the variance of X, is the difference between the second moment and the square of the first moment.

Before continuing our discussion of expected value and moment, we need to prove some basic properties of expected value which will be used through the remaining chapters.

Theorem 2.2 Let X be a random variable with probability function f(x). If C is a constant then the E(C) = C.

Proof:

Applying the definition of expected value we see that the $E(C) = \sum_X Cf(x) = C\sum_X f(x)$. But $\sum_X f(x) = 1$ since f(x) is a probability function. Hence E(C) = C.

Theorem 2.3 Let X be a random variable with probability function f(x). Then the expected value of the sum of two functions of X is the sum of the two expected values, that is, $E[u(X) + v(X)] = E(u(X)) + E(v(X))$.

Proof:

Using the definition of expected value of some function X, i.e., $E[g(X)] = \sum_X g(x)\, f(x)$, if we let g(X) = u(X) + v(X), we have $E[u(X) + v(X)] = E[g(X)] = \sum_X g(x)\, f(x) = \sum_X [u(x) + v(x)]\, f(x) = \sum_X u(x)\, f(x) + \sum_X v(x) f(x) = E[u(X)] + E[v(X)]$. Hence we have $E[u(X) + v(X)] = E[u(X)] + E[v(X)]$.

2.3 Moment Generating Function

In the last section we were concerned with finding the first and second moments of a distribution by applying the definition. But in many cases calculating the moments

by this method becomes very complicated for some probability function. So we search for an alternate way of obtaining the moments of a distribution when the method using only the definition becomes too involved. It turns out that in a great number of cases we can find a function which, when we apply a certain procedure to this function, will give us the moments of the distribution.

Now in our search for this function we want to find some function so that when we apply a certain procedure we will get $\sum_x x^r f(x)$, which is the rth moment of the distribution. It is assumed that the reader is familiar with differentiation of a polynomial and of functions which contain e to powers of a variable. We shall also assume that the reader is familiar with the expansion of e in the form of series, i.e., $e^t = 1 + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots + \frac{t^k}{k!} + \ldots$

Let us consider

$$E(e^{tX}) = E\left[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \ldots + \frac{(tX)^k}{k!} + \ldots\right] =$$

$$E\left[1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ldots + \frac{t^kX^k}{k!} + \ldots\right].$$

The theorem in the last section can also be verified for the infinite case. Hence, assuming the theorem is true for this case after distributing the expected values, we get the following:

$$E(e^{tX}) = E(1) + E(tX) + E\left(\frac{t^2X^2}{2!}\right) + E\left(\frac{t^3X^3}{3!}\right) + \ldots, + E\left(\frac{t^kX^k}{k!}\right) + \ldots =$$

$$1 + t\mu + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \ldots, + \frac{t^k}{k!}\mu_k' + \ldots$$

If we are to get μ_i' out of the above expression, we must apply some procedure to eliminate $\frac{t^r}{r!}$. Suppose we take the derivative r times with respect to t, then we would have an expression as follows:

$$\frac{d^r\, E(e^{tX})}{dt^r} = \mu_r' + (r+1)(r)(r-1)\ldots\, r-(r-2)\; t\mu_{r+1} + \ldots$$

Every term after μ_r' has t as a factor. Now an obvious way to get μ_r' by itself in the expression is to evaluate the expression at t=0. Let us denote $E(e^{Xt})$ by m(t).

Let us now see if $\frac{d^r}{dt^r} m(t)$ evaluated at t=0 is the rth moment as defined earlier, i.e., is $\frac{d^r}{dt^r} m(t) = \sum_x x^r f(x)$ where t=0. Consider $\frac{d}{dt} m(t) = \frac{d}{dt}\left[\sum e^{tx} f(x)\right] = \sum_x x e^{tx} f(x)$. And $\frac{d^r}{dt^r} m(t)$ evaluated at t=0 is given by

$$\frac{d^r}{dt^r}\left[m(t)\right]_{t=0} = \sum_x x^r e^{0(x)} f(x) = \sum_x x^r f(x).$$

So we see a logical choice for a moment generating function for a distribution is $E(e^{tX})$. We shall now give a

formal definition of the moment generating function of a random variable X.

Definition 2.3 Let X be a random variable with probability function f(x). The expected value of e^{tX} is called the moment generating function of X if the expected value exists for every value of t in some interval $-k^2 \leq t \leq k^2$.

The moment generating function is denoted by

$$m(t) = E(e^{tX}) = \sum_x e^{tx} f(x)$$

if X is a discrete random variable. In all the probability functions which we shall be dealing with the moment generating function will exist.

Before leaving moment generating functions, there is one very useful theorem which we shall state. Many times we are confronted with the problem of proving a statistic has a certain distribution. This problem is simplified to a great extent by using the following theorem.

Theorem 2.4 Let X and Y be two random variables with densities f(x) and g(y), respectively. Suppose that the moment generating functions of X and Y both exist and are equal for all t in some interval $-h^2 \leq t \leq h^2$. Then the two densities are equal.

So we see if we know the form of a moment generating function for a variate and we are given another random

variable, if we can show that its moment generating function is of the same form, then by the previous theorem we see that they will have the same probability function. To help us understand the concept of moment generating functions and to illustrate the preceding theorem let us consider another example. We will use the expansion for e^a which was given earlier.

Example 2.2 Suppose that X is a random variable and its probability function, the poisson, is defined by

$$f(x) = \frac{m^x e^{-m}}{x!} \quad , \quad x=0,1,2,\ldots.$$

The first question which might come to mind is to prove that this is a probability function, i.e., $f(x) \geq 0$ for every x, and that $\sum_x f(x) = 1$. To prove the second part, consider

$$\sum_x f(x) = \sum_x \frac{m^x e^{-m}}{x!} = e^{-m} \sum_x \frac{m^x}{x!} = e^{-m}(e^m) = e^0 = 1.$$

Thus we see $\sum_x f(x) = 1$. The first part is obvious since $x \geq 0$.

What is the moment generating function for X?

$$m(t) = E(e^{tX}) = \sum_x e^{tx} \frac{m^x e^{-m}}{x!} = e^{-m} \sum_x \frac{(me^t)^x}{x!} = e^{-m} e^{me^t} =$$

$e^{-m+me^t} = e^{m(e^t-1)}$. So the moment generating function for this random variable X is $m(t) = e^{m(e^t-1)}$. Now we have the form of the moment generating function for a

poisson, so if we were given a random variable Y and its moment generating function is of this form, say

$m_y(t) = e^{N(e^t-1)}$, then by the previous theorem we know that Y is distributed as a poisson and its probability function is given by

$$g(y) = \frac{N^y e^{-N}}{y!} , \quad y = 0,1,2,\ldots .$$

Let us find the moment generating function of a variate X, where X is distributed as a binomial, that is, $f(x) = \binom{n}{x} p^x q^{n-x}$, $x = 0,1,2,\ldots,n$ and $q = 1-p$. Now the moment generating function is given by $m(t) = E(e^{tX}) =$

$$\sum_{i=1}^{n} \frac{e^{tx} n!}{x!(n-x)!} \; p^x q^{n-x} = \sum_{i=1}^{n} \frac{n!}{x!(n-x)!} \; (pe^t)^x \; q^{n-x}$$

but this sum can be written as a binomial raised to the nth power because the expansion is purely algebraic and need not be interpreted in terms of probabilities. Hence

$$m(t) = (q+pe^t)^n .$$

The desired moments may be obtained by differentiation. If we differentiate m(t) with respect to t, and combine terms we get

$$m'(t) = npe^t(q+pe^t)^{n-1} \text{ and } m''(t) = npe^t(q+pe^t)^{n-2} \left[q+npe^t\right]$$

The values of these derivatives at t=0 are np and np(q+np) respectively; hence, these are the values of μ and μ_2', respectively. If q is replaced by 1-p, it will be observed that μ_2' here agrees with our previous results. For this problem, the moments are easier to obtain indirectly by means of the moment generating functions than directly from definition.

Before terminating our consideration of moment generating functions, let us prove some theorems which will be useful in the following chapters.

Theorem 2.5 The moment generating function of a linear combination of n independent variables is equal to the product of moment generating functions of the individual variables, evaluated at $a_i t$, that is,

$$M_{(a_1x_1+a_2x_2+\dots+a_nx_n)}(t) = M_{x_1}(a_1t)M_{x_2}(a_2t)\dots M_{x_n}(a_nt_n)$$

Proof:

$$\text{Consider } M_{(a_1x_1+\dots+a_nx_n)}(t) = E\left[e^{(a_1x_1+a_2x_2+\dots+a_nx_n)t}\right]$$

$$= \sum_{x_1}\sum_{x_2}\dots\sum_{x_n} e^{(a_1x_1+\dots+a_nx_n)t} f(x_1)f(x_2)\dots f(x_n)$$

$$= \sum_{x_1} e^{a_1x_1t} f(x_1) \sum_{x_2} e^{a_2x_2t} f(x_2)\dots \sum_{x_n} (e^{x_n})^{a_nt} f(x_n)$$

$$= (M_{x_1}(a_1t))(M_{x_2}(a_2t))\dots(M_{x_n}(a_nt)).$$

Hence we have the desired conclusion.

Although we used a discrete random variable to prove this theorem, it may be proved just as easily for a continuous random variable by using elementary properties of integration. So we shall assume this theorem is true for a continuous random variable.

We are often confronted, when studying estimation and testing hypotheses, with the problem of determining the density of some statistic based on a random sample of size n. A statistic which is encountered very frequently is the sample mean, $\overline{X}$. Let us now employ theorems 2.3 and 2.4 to determine the density of the random variable $\overline{X}$, where $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal with mean μ and variance σ^2.

Theorem 2.6 If X is normally distributed with mean μ and variance σ^2 and a random sample of size n is drawn, the sample mean, $\overline{X}$, will be normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

Proof:

$$\text{Consider } M_{\overline{X}}(t) = M_{\left(\frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n}\right)}(t) .$$

But by the previous theorem we have

$$M_{\overline{X}}(t) = M_{X_1}(t/n) \ldots, M_{X_n}(t/n).$$

However, since each X_i is normally distributed, we know that the moment generating function for each X_i is given by

$$M_{X_i}(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

Thus the moment generating function for each X_i is given by

$$M_{\overline{X}(t)} = \left(e^{\mu\frac{t}{n} + \frac{\sigma^2}{2}\frac{t^2}{n}}\right)\left(e^{\mu\frac{t}{n} + \frac{\sigma^2 t^2}{2n^2}}\right)\cdots\left(e^{\mu\frac{t}{n}+\frac{\sigma^2 t^2}{2n^2}}\right)$$

$$= e^{n(\mu\frac{t}{n} + \frac{\sigma^2 t^2}{2n^2})}$$

$$= e^{\mu t + \frac{\sigma^2 t^2}{2n}}$$

$$= e^{\mu t + (\frac{\sigma}{n})^2 \frac{t^2}{2}}$$

So by examining the moment generating function of $\overline{X}$, we see that it has a moment generating function which is in the form of a normal variate with mean μ and variance $\frac{\sigma^2}{n}$.

Another useful result is that the statistic $z=\frac{X-\mu}{\sigma}$ is distributed as a normal with $\mu=0$ and $\sigma^2=1$ where X is a normal variable with mean μ and variance σ^2. This result is easily seen by using moment generating functions. The moment generating function of z is given by

$$M_z(t)=E(e^{zt})=E\left[e^{t\frac{X-\mu}{\sigma}}\right]= E\left[e^{t\frac{X}{\sigma}}e^{\frac{-t\mu}{\sigma}}\right] = e^{\frac{-t\mu}{\sigma}} E\left[e^{t\frac{X}{\sigma}}\right]$$

$$= e^{\frac{-t\mu}{\sigma}} e^{\frac{t\mu}{\sigma} + \frac{1}{2}\sigma^2 \frac{t^2}{\sigma^2}} = e^{\frac{1}{2}t^2}$$

Hence z is a normal variable with mean 0 and variance 1. These results will be very useful in developing certain concepts in the following chapters.

CHAPTER III

ESTIMATION

3.1 Introduction

In the study of statistics we are often confronted with the problem of estimating the true value of a parameter in a density or probability function. For example, if we assumed the heights of students enrolled at Oklahoma State University are normally distributed, a statistical problem which could evolve from this is estimating the mean and variance of the random variable representing these heights. In this chapter we shall discuss methods which will give us "good" estimates of these parameters.

Estimation of population parameters is practically always based upon samples from the population involved. For example, we could estimate the mean of a population by taking a random sample of size n from the population and computing $\overline{X}$, the sample mean. It seems obvious that if we use $\overline{X}$ as an estimate of population mean we are using all the information we have available, even though we could use as an estimate for the mean X_1, the first observation. It seems in the latter case we have not utilized all the information at our disposal, i.e., $\overline{X}$ would seem to be a better

estimate of the mean than X_1. We shall, later in this chapter, give mathematical meaning to the word "better" or "good" estimates used in this context. Since the choice of the statistic which is to be used in a given problem must evidently be based on practical considerations, let us investigate first what approach we should use so that our estimates will impart a maximum amount of information with a minimum risk of ambiguity or misinterpretation. We shall be led to one of the most fundamental problems of statistics while trying to answer the question of how to state the results of a problem of estimation. The difficulties encountered in discussing the accuracy of an estimate are paralleled by the difficulties of explaining the relationship between an estimate and the parameter which it is supposed to estimate. These ideas are brought out more clearly in the following illustration. Let us consider the problem of estimation of the height of students at Oklahoma State University, mentioned at the outset of this chapter. Suppose we take a sample of size 20 of heights in inches of students selected in a random fashion and record the results as follows:

73, 66, 68, 66, 69, 66, 73, 70, 70, 73
66, 61, 66, 70, 68, 58, 73, 66, 65, 65

We compute the sample mean and standard deviation of this sample to be 67.6 and 3.78 inches, respectively.

Now if we want to estimate the true mean m of the

heights of students on the basis of this sample, there is practically no limit to the variety of methods we could use to state our results. Let us consider a few alternatives.

Alternative 1. The mean of the population is estimated to be equal to 67.6 inches.

Alternative 2. On the basis of a random sample of size 20, the mean of the population is estimated to be 67.6 inches. This estimate is the mean of the sample.

Alternative 3. The mean of the population is estimated to be 67.6 inches, which is based on a sample of 20 measurements which have a standard deviation of 3.78 inches.

Alternative 4. We are 95 per cent confident that the interval from 65.74 to 69.46 inches contains the actual mean of the population.

While considering the alternatives, it should be apparent that whereas the first three alternatives are, in principle, much alike, the fourth is of an entirely different nature. The first alternative seems to have several shortcomings since it gives no indication how we arrived at the estimate. Thus, it would be very difficult to draw any conclusions whatsoever about the accuracy of our results. We might have arrived at these results by just taking the average height of our best friend and ourself.

We see that the second alternative is an improvement over the first since it tells us the method by which the estimate was computed and how large a sample this estimate was based upon. However, it says nothing about the variability of the measurements which made up the sample, and it still leaves us in no position to get a real appraisal of the accuracy of the estimate.

The third alternative furnishes us with all the information a trained statistician would need in order to discuss the reliability of our estimate. However, usually we will be supplying these estimates to non-statisticians so we must state the conclusion in terms that will be meaningful to them. Actually, alternative three is used to make the statement in alternative four.

So it seems that if we are reporting to non-statisticians, alternative four would be the most meaningful to address to them. From our earlier consideration of probability it is immediately obvious that if we assign an estimate of probability .95, i.e., if we say that we are 95 per cent confident, this means that we have used a method of estimation which is successful about 95 per cent of the time. Alternative four is actually implying that it would be a fair bet to give 19 to 1 odds that the interval from 65.74 to 69.46 inches contains the mean of the population.

Had we wanted to be more certain of alternative four being true, we could have made the statement that we are 99 per cent confident that the mean of the population lies

in the interval from 65.06 to 70.14. Alternative four presents us, therefore, with a method of stating our results in a form which is understood easily by laymen and requires no further calculations to inform us directly of the reliability of our method of estimation.

In the discussion on the preceding page, we mentioned two types of estimation; alternative three gave a number as an estimate of the population mean while alternative four gave a different type of estimation, an interval estimation. The former type of estimation, called a point estimate, is very useful in developing the theory of statistics. Generally speaking, a point estimate is the familar kind of estimate, that is, it is a number obtained from computations on the observed values of the random variable which serves as an approximation to the parameter. For example, the observed proportion of defective parts in 50 consecutive parts turned out by a machine is a point estimate of the true proportion p for the machine. An interval estimate is an interval determined by two numbers obtained from computation on the observed values of the random variables that is expected to contain the true value of the parameter in its interior. Since point estimates play an important part in developing the theory of statistics, we shall devote some time to this concept on the following pages.

3.2 Maximum Likelihood Function

In order to know how to use several observations of a random variable in an intelligent manner for constructing a point estimate of a parameter of a density function of the random variable, it is desirable to have some general principles to follow. The principle, or method, should be such that the estimates obtained by using the method will possess desirable properties. For example, if two different methods are tried on the same sets of observations and if one method produces estimates that are consistently closer to the value of the parameter being estimated than those of the other method, then the first method would obviously be preferred. Properties of good point estimates will be considered later; here it suffices to describe a method of obtaining point estimates that is usually preferred by statisticians. This method of estimation, known as the maximum likelihood method, is used in the following chapter whenever problems arise of finding a point estimate of a parameter of a frequency function. We shall formalize the definition after some necessary notation has been introduced.

Let $f(x;\theta)$ be a density function of the random variable X, where θ is the parameter to be estimated. Suppose that n observations are to be made of the variable X. Let $X_1, X_2, \ldots, X_n$ denote the n random variables corresponding to these n observations. The function given by

$L = f(x_1;\Theta)\ f(x_2;\Theta)\ldots,f(x_n;\Theta)$ defines a function of the variables $x_1,\ldots,x_n$ and the parameter Θ which is known as the likelihood function.

For our purpose, we are supposing that the observational values are obtained from n independent trials of an experiment for which $f(x;\Theta)$ is the frequency function of a discrete random variable X. Then for any particular set of observational values, the likelihood function gives the probability of obtaining that set of values, since $f(x_i;\Theta)$ is the frequency function of X_i. If, however, X is a continuous variable, the likelihood function gives the probability density of a sample $(x_1,x_2,\ldots,x_n)$, i.e., the joint density of n independent random variables, where the sample point is thought of as being n dimensional.

Now, for a given set of observational values, an estimate of Θ is merely a number obtained from calculations made on the observational values, i.e., an estimate is simply a function of the observational values. For example, a useful estimator mentioned earlier is $g(X_1,X_2,\ldots,X_n) = \frac{(X_1+X_2+\ldots,+X_n)}{n}$, which is a function of the observed random variables. We shall usually refer to the function of observed random variables as an estimator, while the actual value computed will be called an estimate of the parameter. For example, in the illustration earlier in this chapter, we would call $g(X_1,X_2,\ldots,X_n) = \overline{X}$ an

estimator of the true mean of the population while the value $\bar{X} = 67.6$ is called an estimate of the mean.

Using the notation and terminology of the preceding paragraphs, the method of maximum likelihood estimation may be defined in the following manner.

Definition 4.1 A maximum likelihood estimator $\hat{\theta}$ of the parameter θ in a density or probability function $f(x;\theta)$ is an estimator that maximizes the likelihood function $L(x_1,x_2,\ldots,x_n;\theta)$, where L is considered as a function of θ.

If the x_i's are treated as fixed, the likelihood function becomes a function of θ only, say $L(\theta)$, consequently, the problem of finding a maximum likelihood estimator is the problem of finding the value of θ that maximizes $L(\theta)$. This type of problem can be handled in many instances by differentiating the likelihood function, $L(\theta)$, with respect to θ and setting the derivative equal to zero. However, any method of finding an estimator for θ which maximizes $L(\theta)$ is acceptable. The functions which we shall need to differentiate will be very simple polynomials or functions of e to some power.

Let us consider an example where we shall obtain the maximum likelihood estimator for a density function $f(x;\theta)$.

Example 4.1 Suppose X is a random variable which has a density $f(x;\theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$. Now , the likelihood function

for a random sample of size n is $L(\theta) = \frac{1}{\theta^n}$. To maximize $L(\theta)$ we must make θ^n as small as possible. If our sample is $x_1, x_2, \ldots, x_n$ we must use as an estimate of θ the largest observed value in the sample of size n, since θ cannot be smaller than the largest observed value. Hence, the maximum likelihood estimate of θ is the maximum observed value.

Many times when we are confronted with the problem of determining the maximum likelihood estimator, the likelihood function involves e to some power of θ. Hence, in our search for the estimators of θ which will maximize $L(\theta)$, we might just as well search for estimators which will maximize $\ln(L(\theta))$ since, if a $\hat{\theta}$ will maximize $L(\theta)$ it will make $\ln(L(\theta))$ a maximum also. Sometimes it will simplify the algebra to a great extent to use $\ln(L(\theta))$ to determine $\hat{\theta}$.

Let us consider an example where X is a random variable with probability function $f(x;p) = p^x(1-p)^{1-x}$, $x=0,1$. Now, the likelihood function is

$$L(x_1,x_2,\ldots,x_n;p) = p^{x_1}(1-p)^{1-x_1}\, p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n}$$

$$= p^{\sum_{i=1}^{n} x_i}\,(1-p)^{n-\sum_{i=1}^{n} x_i}$$

So to find the maximum likelihood estimator of p, we shall differentiate $\ln(L(p))$ with respect to p and then set the derivative equal to zero. We get

$$\ln(L(p)) = \Sigma x_i \ln(p) + (n - \Sigma x_i)\ln(1-p)$$

and differentiating $\ln(L(p))$ yields

$$\frac{d \ln(L(p))}{dp} = \frac{\Sigma x_i}{p} - \frac{(n-\Sigma x_i)}{1-p} = \frac{(1-p)\Sigma x_i - np + p\Sigma x_i}{p(1-p)}$$

$$= \frac{\Sigma x_i - np}{p(1-p)} .$$

Setting the derivative equal to 0, we get

$$\frac{\Sigma x_i - n\hat{p}}{\hat{p}(1-\hat{p})} = 0$$

which implies

$$\hat{p} = \frac{\Sigma x_i}{n} = \overline{X} .$$

Hence the maximum likelihood estimator, $\hat{p}$ of p is the sample mean $\overline{X}$. Suppose a sample of size 10 yields 6 ones and 4 zeros, then an estimate of the parameter p would be .6 . In the preceding remarks we mentioned that maximum likelihood estimation is the favorite method of many statisticians for obtaining point estimates of a parameter. Let us now turn our attention to some desirable properties

which we want estimators to possess.

3.3 Unbiased Estimates

One of the properties which we hope an estimator possesses is that the value of estimates would consistently be close to the parameter that we are trying to estimate, i.e., if we repeated sampling several times and compute $\hat{\theta}$ for each sample, then most of these estimates would be concentrated near the true parameter θ. We are actually saying that we would want the mean of the random variable $\hat{\theta}$ to be θ. But the mean of $\hat{\theta}$ is the $E(\hat{\theta})$ which we want to be equal to θ, i.e., we shall insist that $E(\hat{\theta}) = \theta$. This property is called unbiasedness. Let us now give a formal definition of an unbiased estimator of a parameter θ.

Definition 3.2 The statistic $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is called an unbiased estimate (estimator) of the parameter θ if the expected value of $\hat{\theta}$ is θ, i.e., $E(\hat{\theta}) = \theta$.

This property merely states that the random variable $\hat{\theta}$ possesses a distribution whose mean is the parameter θ being estimated. For example, we might expect $\bar{X}$ to be an unbiased estimator of the mean θ for a random variable X whose density or probability function is $f(x;\theta)$. Now the expected value of $\bar{X}$ is given by

$$E(\bar{X}) = \frac{E(\Sigma X_i)}{n} .$$

But since the X_i's are independent, we have

$$\frac{E(\Sigma X_i)}{n} = \frac{1}{n}\ \Sigma E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\theta = \frac{n\theta}{n} = \theta.$$

Hence $E(\overline{X}) = \theta$, i.e., $\overline{X}$ is an unbiased estimator of the mean of the random variable X.

Let us consider an illustration which shows the bias of a statistic determined by employing the expected value. Consider the expected value of the sample variance based on a random sample of size n. From the properties of E and the definition of S^2, it follows that

$$E(S^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i^{\,2}-2X_i\overline{X}+\overline{X}^2)\right]$$

$$= E\left[\frac{1}{n}\ (\Sigma X_i^{\,2}-2n\overline{X}^2 + n\overline{X}^2)\right]$$

$$= E\left[\frac{1}{n}\Sigma X_i^{\,2}-\overline{X}^2\right]$$

$$= \frac{1}{n}\ \Sigma E(X_i^{\,2})-E(\overline{X}^2)$$

$$= \frac{1}{n}\ \Sigma(\ \sigma^2+\mu^2)-(\ \sigma_{\overline{X}}^2+\mu^2)$$

$$= \sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2$$

$$= \frac{n\ \sigma^2-\ \sigma^2}{n}$$

which yields

$$E(S^2) = \frac{(n-1)\,\sigma^2}{n}$$

This shows that S^2 is not an unbiased estimator of σ^2, which means that if repeated samples of size n are taken and the resulting sample variance is averaged, the average will not approach the true variance in value but will be consistently too small by the factor $\frac{n-1}{n}$. For small samples this factor becomes important; consequently, one must be careful how he combines samples in making an estimate of the true variance when an unbiased estimate is desired. In order to overcome the bias in S^2, it is merely necessary to multiply S^2 by $\frac{n}{n-1}$, i.e.,

$$\frac{E(\Sigma(X_i-\overline{X})^2)}{n-1} = \sigma^2,$$

which says that $\frac{(X_i-\overline{X})^2}{n-1}$ is an unbiased estimate of σ^2.

Let us now consider an example in which we will actually compute the maximum likelihood estimator for a parameter in a density function. Suppose that (1,1,0,1) is a random sample taken from a point binomial distribution with parameter p. Let us obtain the maximum likelihood estimator of p. The likelihood function L(p) is given by

$$L(p) = p^{\sum_1^4 x_i}\,(1-p)^{4-\sum_1^4 x_i}$$

also

$$\frac{dL(p)}{dp} = -(4-\Sigma x_i)\, p^{\Sigma x_i} (1-p)^{3-\Sigma x_i} + x_i (1-p)^{4-\Sigma x_i} p^{\Sigma x_i - 1}$$

Setting $\frac{dL(p)}{dp} = 0$ to obtain the value of p which maximizes L(p) we get

$$-(4-\Sigma x_i) + \Sigma x_i \frac{(1-\hat{p})}{\hat{p}} = 0$$

$$-\hat{p}(4-\Sigma x_i) + \Sigma x_i - \Sigma x_i \hat{p} = 0$$

$$-4\hat{p} + \Sigma x_i = 0$$

and

$$\hat{p} = \frac{\Sigma x_i}{4}$$

Therefore the maximum likelihood estimate of p is

$\frac{1+1+0+1}{4} = 3/4$, i.e., based upon this sample we would use 3/4 as an estimate of p.

To help us become more proficient in determining the maximum likelihood estimates of a parameter in a probability function, let us consider another illustration. Suppose that (0,3,1,0,2,1,0,2) is a random sample taken from a poisson distribution. Let us obtain two unbiased estimates of m where

$$f(x) = \frac{e^{-m} m^x}{x!} .$$

Let us first determine the maximum likelihood estimator of m. The likelihood function is given by

$$L(m) = \frac{e^{-nm}\, m^{\Sigma x_i}}{\prod_{i=1}^{n} (x_i)!}$$

and the log of L(m) is given by

$$\ln L(m) = -nm + \Sigma x_i \ln m - \Sigma \ln(x_i!)$$

Taking the derivative of the ln(L(m)) with respect to m and setting it equal to zero yields

$$-n + \frac{\Sigma x_i}{\hat{m}} = 0$$

solving for $\hat{m}$ we get

$$\hat{m} = \frac{\Sigma x_i}{n} \quad .$$

Hence, in this example the actual maximum likelihood estimate of m is given by

$$\hat{m} = \frac{3+1+2+1+2+0+0+0}{8} = \frac{9}{8} \quad .$$

Now that we have an estimate of m, we need to determine if this estimate is unbiased. Recalling from the last chapter we proved if X is a poisson variate then the expected value of X, E(X), is equal to m. Hence we have

$$E(\hat{m}) = \frac{E(\Sigma X_i)}{n} = \frac{1}{n} \Sigma E(X_i) = \frac{1}{n} \Sigma m = \frac{nm}{n} = m.$$

Thus $\hat{m} = \overline{X}$ is an unbiased estimate of m. However, X_1, the first number of the sample, is also an unbiased estimate of m, since $E(X_1) = m$. Thus any element in the sample would serve as an unbiased estimate of m, however the maximum likelihood estimator, $\overline{X}$, has other desirable properties such as being a sufficient statistic for m. Thus $\overline{X}$ seems to be a better unbiased estimate of m than just any element of the sample. The concept of sufficiency will be discussed in the latter part of the chapter.

Although the property of being unbiased is a desirable one to seek in an estimator, it is not nearly so important as the property of an estimate being close in some sense to the parameter being estimated. Thus, if an estimator t gave estimates which were consistently closer to θ than another estimate t' in repeated samples of the same size, then t would certainly be preferred to t', even if t were biased and t' were unbiased. Let us now consider some desired properties which we shall want estimators to possess.

Suppose you are a manager of a big league baseball club and you are trying to find some way of determining how valuable a certain player is to your club. So you send for his record over the past two years. You have a mass of information showing how many hits he got in each game, but you realize how hard it is to draw any conclusion after considering this enormous amount of data. So you want to reduce this data to one number which will contain all the

information about the player. We could say we would just look at the first 10 games in which he participated and find the per cent of hits and base our decision upon this number. However, it seems we have lost some information which we had available. A more appropriate measure of the player's worth would be his per cent of hits in times at bat. It seems that we have lost no information contained in the sample in computing an estimate of the player's worth.

The example mentioned in the preceding paragraph leads us to an important property which we shall want estimators of a parameter to possess. In statistics we are usually furnished with a sample from a population whose density is $f(x;\theta)$, and using it we want to reduce these n random variables to a single random variable. Our objective is to try to find an estimator based upon the sample which gives as much "information" as possible about the parameter θ. If this is the case, we prefer to work with $\hat{\theta}$ rather than the n random variables $X_1, X_2, \ldots, X_n$ for the simple reason that one random variable is usually easier to use than n random variables. If we find a statistic $\hat{\hat{\theta}}$, where $\hat{\theta}$ is an estimator of the parameter θ, which contains all the information about the parameter contained in a random sample of size n from a probability function $f(x;\theta)$, then we say that $\hat{\theta}$ is a sufficient statistic for θ. A formal definition of a sufficient statistic is given in almost any mathematical statistics text, but the definition is nearly impossible to

apply. In our consideration of the concept we shall investigate some examples to help illustrate this idea and state a theorem which is much more workable than the definition. Let us consider another example to help illustrate this concept of sufficiency.

Suppose you are a senior at a high school which has an enrollment of about 500. The administration decides to try to determine how its senior students compare with other seniors who attend high schools of similar size in a particular state. Suppose further that the administration has access to certain standardized tests which measure these desired properties, but these tests are costly so they decide to pick 20 seniors at random and administer the test to this group. Now, our problem is to try to obtain an estimate of how your school compares to those other schools in your area. Suppose we know that the mean of the scores obtained on this test in the past has been 50. So the administration gives the test and obtains twenty scores. To help the administration's ego they might suggest that the highest score obtained on the test be used as an estimate of the school's worth. However, it seems that we have lost some information which was contained in the sample by using the highest score as an estimate. Hence, it would seem that this estimate is not sufficient. However, if we would compute the sample mean of the test scores it seems we would have an estimate of the school's worth which con-

tains all the information in the sample of size 20. Thus we see that $\bar{X}$ is a sufficient statistic for the unknown parameter in the population.

Let us now turn to a somewhat more mathematical method for determining whether an estimator $\hat{\theta}$ is a sufficient statistic for a parameter θ. A relatively easy criterion has been developed by J. Neyman which can be used, in many cases, for examining a statistic $\hat{\theta}$ for sufficiency. The following theorem gives us a relatively easy method for judging whether a certain statistic is sufficient. We shall state the theorem without proof.

Theorem 3.1 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the probability density $f(x;\theta)$ and let the joint density of these n random variables be $g(x_1, x_2, \ldots, x_n;\theta) = f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta)$. If this density factors as follows: $g(x_1, \ldots, x_n;\theta) = h(\hat{\theta};\theta)\, k(x_n, x_2, \ldots, x_n)$, where $k(x_1, x_2, \ldots, x_n)$ does not involve the parameter θ, then $\hat{\theta}$ is a sufficient statistic for θ.

To help recognize the usefulness of this theorem, let us consider an example. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from the probability functiom $f(x;p)^n = p^x(1-p)^{1-x}$, $x = 0,1$. Now the joint probability function is the product of the probability function of each X_i, since the sample

is random, therefore the probability function is given by

$$g(x_1,x_2,\ldots,x_n;p) = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} .$$

We shall show that $\overline{X}$ is a sufficient statistic for p. To apply the theorem just stated, we must show that g factors into two functions, one of which contains only the parameter p and the estimator $\overline{X}$ and one which contains only x_i's. Consider

$$g(x_1,x_2,\ldots,x_n;p) = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$$

$$= p^{\frac{n\Sigma x_i}{n}}(1-p)^{\frac{n(1-\Sigma x_i)}{n}}$$

$$= p^{n\overline{x}}(1-p)^{n(1-\overline{x})} \cdot 1$$

If we let $h(\overline{x};p) = p^{n\overline{x}}(1-p)^{n(1-\overline{x})}$ and $k(x_1,x_2,\ldots,x_n) = 1$, we see that these functions satisfy the conditions of the theorems, hence $\overline{X}$ is a sufficient statistic for the parameter p in f(x;p). We are actually saying that $\overline{X}$ contains all the "information" about p which is contained in the sample of size n from f(x;p).

Although we see that an estimator being a sufficient statistic is a desired property, this property alone is not always enough to give "good" estimates for a parameter. For example, in the illustration just given, we could use ΣX_i as an estimator of p. We can see from the given theorem

that ΣX_i satisfies all the hypotheses of the theorem, hence it is a sufficient statistic for p. Therefore, we must search for estimators which possess properties which will consistently give "good" estimators of the parameter, i.e., close to the parameter. Because of the difficulty or impossibility of determining whether one of two estimates is closer than the other to θ for any reasonable definition of closeness, it is customary to substitute a measure of the variability t (the estimate) about θ in place of closeness. Since the variance, or standard deviation, has been used to measure variability throughout the preceding chapters, one might naturally think of selecting one or the other of these measures. However, unless θ happens to be the mean of the distribution of t, the variance will not measure the variability about θ. The difficulty is easily overcome by using the second moment about θ as the desired measure. If θ is the mean of t, then this measure reduces to the variance of t. In view of the preceding discussion, the following definition will be introduced as a basis for choosing good estimators.

Definition 3.3 A statistic t will be called a best unbiased estimate (or estimator) of a parameter θ if t is such that $E(t-\theta)^2 \leq E(t'-\theta)^2$, where t' is any other unbiased estimate of θ.

Although there are other definitions of a best es-

timate in use, the preceding definition is one that is frequently used. It should be realized that the variance was selected (above) because it was considered to measure the concentration of the distribution of t about θ. Let us now consider an example in which we prove that a statistic is the best unbiased estimator in a class of estimators.

Let us consider the problem of determining whether some weighted average of a random sample from a population can yield a better unbiased estimate of the population mean than the sample mean which we have already seen is an unbiased estimator of the population mean. Suppose the two competing estimates are $t_1=a_1X_1 +\dots + a_nX_n$ and $t_2= \overline{X}$. The unknown a's in t_1 are to be selected to make t_1 unbiased and to minimize $E(t_1-\theta)^2$ so the estimator will satisfy the condition of the definition. In order that t_1 be unbiased, calculate

$$E(t_1)=a_1E(X_1)+\dots + a_nE(X_n) = a_1\mu +\dots + a_n\mu=(a_1+\dots +a_n)\mu.$$

So t_1 will be unbiased if we have $a_1+a_2+\dots +a_n = 1$, i.e., the sum of the coefficients in t_1 must be 1. Thus this restriction may be ignored if t_1 is written in the form

$$t_1 = \frac{c_1X_1+\dots +c_nX_n}{c_1+c_2+\dots +c_n}$$

Now that we have t_1 unbiased, its second moment about μ_{t_1} is simply its variance. Since the variables $X_1,X_2,\dots,X_n$

are independent and have the same variance, we have the variance of t_1 given by

$$\sigma_{t_1}^2 = \sigma^2 \frac{\Sigma c_i^2}{(\Sigma c_i)^2}$$

Now we must choose the c's to minimize this expression. To find the value of c which will minimize $\sigma_{t_1}^2$, we shall find the partial derivative of $\sigma_{t_1}^2$ with respect to c_k which gives

$$\frac{\partial (\sigma_{t_1}^2)}{\partial c_k} = 2(\Sigma c_i)^2 c_k - 2(\Sigma c_i^2)\Sigma c_i = 0 \quad (k = 1,2,\ldots,n)$$

Solving for c_k we get

$$c_k = \frac{\Sigma c_i^2}{\Sigma c_i} \qquad (k = 1,2,\ldots,n)$$

which implies that

$$\Sigma c_k c_i = \Sigma c_i^2 \quad \text{and} \quad c_k = c_i \quad (k = 1,2,\ldots,n)$$

This result shows that the best linear combination to use is the one in which the coefficients are all equal, since the c_k does not depend on k, in which case t_1 reduces to $\bar{X}$. Thus we have proved that no linear combination of the sample can yield a better unbiased estimate than the sample mean $\bar{X}$. Therefore, the best unbiased estimator of the sample mean of a population is $\bar{X}$.

CHAPTER IV

THE TESTING OF HYPOTHESES

4.1 Introduction

In the course of everyday living we are engaged in the process of decision making. We must decide in which courses we will concentrate our studies while in high school, what college we shall attend, what clothes we shall wear, where we shall live, what we shall eat for lunch, etc. Many decisions depend almost entirely upon a person's likes and dislikes. Some decisions are so oriented that we must face a certain amount of risk of taking the wrong alternative. It is evident that decisions we make are usually highly influenced by our past experience, our individual tastes, scientific evidence, and the like. For example, when a farmer is planting cotton, he must make a decision upon the depth he will plant the seeds. This decision is partly based upon his experience from past years when he planted cotton under similar conditions. He should also rely upon scientific evidence, such as; certain types of seeds might germinate more rapidly than other types, hence, he would probably not plant these as deep as those of slower germination. As mentioned earlier, some decisions

are based almost completely upon personal tastes. We shall exclude this type of decision making throughout this chapter.

Let us consider an example, which, though it does not apply in the scientific realm, is typical of the types of situations which we can expect to meet in problems in which we want to test a hypothesis.

Suppose you are a member of a high school baseball team and you are participating in a ball game against a rival school. Suppose further that you have just singled into left field and are now standing on first base trying to decide if you should try to steal second. The coach has signaled that you are on your own, i.e., you may try to steal if you like. You might reason as follows. You recall that in an earlier inning another one of the boys on the team stole second base off this pitcher and you are almost as fast as that boy. You also reason that if you tried to steal, the batter might hit a line drive to one of the infielders and thus not only get the batter out, but also you. If you choose to steal, you are gambling that you can beat the catcher's throw and that the batter does not hit a line drive to one of the fielders. Similarly, if you choose not to steal, you are gambling that the batter does not hit a grounder which could result in a double play. Our problem is to try, by employing some useful statistical procedure, to pick

the alternative which will have the smallest risk associated with it. The problem of evaluating these risks becomes very complicated if we permit situations in which we must choose between more than two different courses of action. In our consideration of decisions we shall limit our study to those of only two alternatives or those which can be altered to have only two alternatives.

In the example given, we see that it is not easy to find a suitable, or most suitable, method of decision for any given question. You might choose to try to steal second on the basis of your recalling that one of your teammates stole a base earlier or you might decide to not attempt to steal since the last time you tried for a stolen base you were unsuccessful. Or another alternative might be to flip a coin, "Tails you stay on 1st, or heads you try to steal 2nd".

Our concern here is to try to pick from these three methods of decision (and conceivably more) the one which is more likely to be successful. We can easily see that the only method which we can attach a probability to is the third where we trusted our fate to the flip of a coin. If we trust our fate to the coin, we can expect to make the correct decision 50 per cent of the time regardless of whether we should or should not try to steal the base. It should be observed that when we attach a probability to a method of decision we are referring to the success ratio

of the given method of decisions if it were employed a great number of times. We note that we could use the method of a flip of a coin for every decision with which we are confronted. But the probability associated with the above type of decision making method is .50. We shall strive to develop methods which will give us correct decisions with a much higher probability associated with them.

4.2 Statistical Hypotheses

Very often in practice we are called upon to make decisions about a population on the basis of sample information. Such decisions are referred to as statistical decisions. For example, we may wish to decide on the basis of sample data, whether a new serum is really effective in curing a particular disease, whether one educational procedure is better than another, whether a given coin is loaded, etc.

In attempting to reach decisions, it is useful to make assumptions or guesses about the populations involved. These assumptions, which may or may not be true, are called statistical hypotheses and in general are statements about the probability distributions of the populations.

In many instances, we formulate a statistical hypothesis for the sole purpose of rejecting or nullifying it. For example, if we want to decide whether a given coin is loaded, we formulate the hypothesis that the coin is fair,

i.e., $p = .5$, where p is the probability of heads. Similarly, if we want to decide whether one procedure is better than another, we formulate the hypothesis that there is no difference between the procedures (i.e., any observed differences are merely due to fluctuations in sampling from the same population). Such hypotheses are often called null hypotheses and we shall denote them by H_0.

Any hypothesis which differs from a given hypothesis is called an alternative hypothesis. For example, if one hypothesis is $p = .5$, alternative hypotheses are $p = .7$, $p \neq 5$ or $p \geq .5$. We shall choose the notation H_1 to represent the alternative to the null hypothesis.

If, on the basis of a particular hypothesis, we find that results observed in a random sample differ markedly from those expected under the hypothesis on the basis of pure chance using sampling theory, we would say that the observed differences are significant and we would be inclined to reject the hypothesis (or at least not accept it on the basis of the evidence obtained). For example, if 20 tosses of a coin yield 16 heads, we would be inclined to reject the hypothesis that the coin is fair, although it is conceivable that we might be wrong. These types of problems will be our concern in this chapter. Let us now consider an example which will help illustrate these new concepts.

The example presented at the outset of this chapter

is not one in which we would be concerned with the testing of a hypothesis. The methods used for testing hypotheses are usually a great deal more refined and at times also much more complicated. However, in principle they are all more or less the same. To help us visualize some of the difficulties which we shall encounter when we are asked to accept or reject a scientific hypothesis, let us illustrate this concept with an example.

Suppose you are the manager of a certified seed distributing company. Suppose further that your main income is from the selling of cotton seeds, which your company guarantees to have a germination of 70 per cent, i.e., 70 per cent of these seeds will sprout when planted during favorable conditions. For some reason or another the USDA decides to investigate this claim and it assigns one of its agents to test the hypothesis that 70 per cent of these seeds will actually germinate. The agent has instructions from his superior to take a sample of seeds of size 100 and base his final decision on the following criterion:

> He should accept the hypothesis H_0 if the sample of 100 seeds contain 61 or more seeds which will germinate when planted under favorable conditions.
>
> He should reject the hypothesis H_0 if the number of seeds that germinate are less than 61.

The investigator is actually going to base his decision on the number of seeds he finds which germinate in a sample of size 100. If, after planting these seeds under ideal conditions, he observes 61 or more seeds which germinate, then we will conclude that the germination of the seed is probably .70. On the other hand, if less than 61 of these seeds germinate, he will reject the hypothesis and charge the seed company with misleading claims about its product.

Even though you are not trained in the field of statistics, you agree reluctantly that the criterion seems fair, but you are still worried that bad luck might play tricks on you, i.e., for example, the sample might come from a sack which just by chance got wet during the process of shipment, which ruined most of the seed. You are concerned, even though the seeds usually test to have about 70 per cent germination, about the possibility that the hypothesis H_0 might be rejected.

After our consideration in chapter one on probability we must agree that it is certainly possible that the investigation might produce less than 61 good seeds despite the fact that the germination is actually .70. On the other hand, we know enough statistics to believe that such an occurrence would be extremely unlikely. Let us consider this example further and try to determine the actual probability that the hypothesis will prove unfavorable to you.

If we repeated the experiment many times we see that the proportion of successes would be concentrated close to .70 and few results would be outside of .65 to .75. It seems reasonable to assume the number of good seeds from a sample of size 100 would approach the normal density, which is actually the case since the number of successes in n trials is a binomial distribution and its limiting distribution is the normal. To simplify this example, we shall assume that X, which is equal to the number of good seeds in a sample of size 100, has a normal distribution. We need only make a small adjustment in our hypothesis to assume normality, since the normal is continuous and we must adjust the hypothesis as follows:

Accept the hypothesis H_0 if the number of seeds that germinate is greater or equal to 60.5.

Reject the hypothesis H_0 if the number of good seeds is less than 60.5.

If we investigate the change in criterion, we find this is the customary procedure of spreading a discrete variable over a continuous scale, and since we obviously cannot get 60.5 good seeds the criterion is, for all practical purposes, exactly the same as before.

To help us determine the actual probability of this "bad luck", we ask the following question, "If the true proportion of good seeds is .70, what is the probability of

getting a sample which contains less than 61 good seeds?". In other words, what is the probability of the rejection of hypothesis H_0 when it is actually true. This probability is given by the shaded area in Figure 6, and can be evaluated by using a normal table given in almost any statistics book. Since we have assumed that p=.70 and n=100, the mean and standard deviation of the sampling distribution of the number of good seeds in a sample of size 100 are, as mentioned early in Chapter I

$$m = 100\ (7/10) = 70$$

and

$$\sigma = \sqrt{(100)(7/10)(7/10)} = 7\ .$$

The z value which corresponds to the dividing line of our criterion, i.e., to 60.5 is $z = (\frac{60.5-70}{7}) = 1.36$. The normal curve area which corresponds to a z value of .4131 is represented by the area bounded by the line z = 70 and z = 60.5 . Now since 70 is assumed to be the mean, the area to the left of z = 70 is .5000. Hence the shaded area is .5000-4131 = .0869.

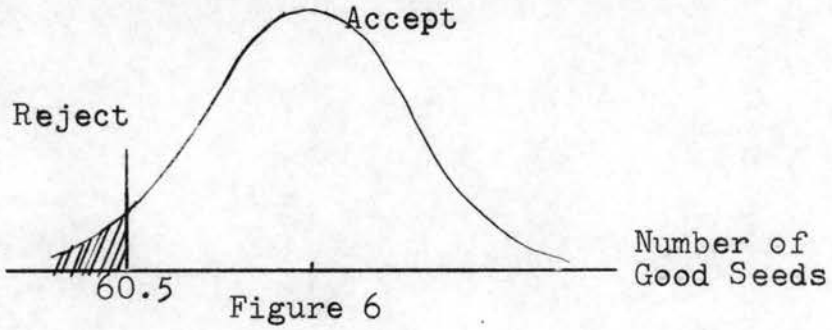

Figure 6

Thus the probability of getting a sample for which the observed number of seeds falls into the left tail of the distribution, i.e., into the rejection region, is .0869. Hence, the probability of the investigator rejecting the hypothesis H_0 on the basis of this criterion is approximately .087 if the true proportion is actually .70, i.e., if the hypothesis H_0 is actually true. So if by chance, H_0 were to be rejected when it should have been accepted, we have committed an error usually referred to as a type I error, which we shall consider in more detail later. We see that it is type I error which concerns us.

It is understandable why we are concerned, now that we know that in approximately one out of eleven experiments the results would be negative even though they should be affirmative. You feel that the risk is too high so you suggest to the investigator to use a criterion which has a smaller type of error. Actually the type I error can be made as small as we want, for instance we could always accept the hypothesis and in so doing never make type I error. Surely, this would be ideal for you, but in eliminating type I error we have left ourselves wide open for another type of error, namely, the error of accepting the hypothesis H_0 when it should have been rejected. This type of error is committed whenever we accept a hypothesis when actually it should be rejected and is called type II error. We can see, if we were in the investigator's position, that we would strive to make type II error small.

You decide that the type I error is too high, so we search for a method which will reduce this error. An obvious method will be to enlarge the acceptance region and in so doing, reduce the rejection region. We could, for example, change the criterion to read

> Accept H_0 if there are 50 or more good seeds in the sample of 100 seeds.
>
> Reject H_0 if the sample has less than 50 good seeds.

Now $z = (\frac{50-70}{5}) = -4$, which corresponds to the normal area of .4990. Hence, the area to the left of $z=-4$ is .5000-.4990 = .001. We see that this new criterion is much more favorable to you and your company since the probability of type I error is .001, but at the same time, it puts the USDA at a great disadvantage. It makes it very difficult to prove your company wrong even if the true per cent of good seeds is less than 70.

It should be evident after considering this example, that when testing a hypothesis we must concern ourselves with both type I and type II errors. For if we were in the investigators position we would try to reduce type II error, so in practice we must strive to reduce both types of errors. Let us now turn to this problem from a more mathematical point of view. Some of the techniques developed will be useful in Chapter 6.

4.3 Type I and Type II Errors

In our study of testing hypotheses we shall be concerned mainly with two types of errors which are associated with statistical decisions, type I and type II errors. If we reject a hypothesis when it should be accepted, that is, when the hypothesis is actually true, we say that a type I error has been made. If, on the other hand, we accept a hypothesis (many statisticians never say they accept a hypothesis, they say that they do not reject the null hypothesis) when it should be rejected, we say that a type II error has been made. In either case, we see a wrong decision or error in judgment has occurred.

Closely associated with the concept of type I error is the idea of level of significance. In testing a given hypothesis, the maximum probability with which we would be willing to risk a type I error is called the level of significance of the test. This probability, often denoted by α, is generally specified before any samples are drawn, so the results obtained will not influence our choice.

In practice, a level of significance of .05 or .01 is customary, although other values are used. If for example a .05 or 5% level of significance is chosen in designing a test of hypothesis, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted, i.e., we are about 95 per cent confident we have made the right decision. In such cases we say

that the hypothesis has been rejected at a .05 level of significance, which means we could be wrong with probability .05.

In order for any tests of hypotheses or rules of decision to be good, they must be designed so as to minimize errors of decisions. This is not a simple matter since, for a given sample size, an attempt to decrease one type of error is accompanied in general by an increase in the other type of error. In practice, we must remember what item we are dealing with. One type of error may be more serious than the other, and so a compromise should be reached in favor of a limitation of the more serious error. One of the best ways to reduce both types of error is to increase the sample size, which may or may not be possible.

Let us consider an example involving a normal statistic. Suppose that under a given hypothesis the sampling distribution of a statistic S is a normal distribution with the mean μ and variance σ^2. Then the distribution of the standardized variable z, as given by $z = \frac{S-\mu}{\sigma}$, is the standardized normal distribution (mean 0, and variance 1) and is shown in Figure 7.

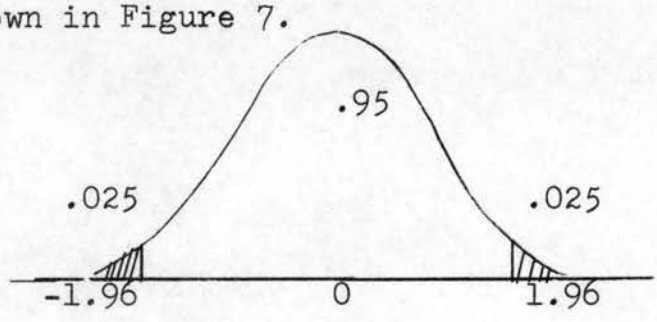

Figure 7

As indicated in Figure 7, we can be 95 per cent confident that, if the hypothesis is true, i.e., z is actually a sample statistic, z will be between -1.96 and 1.96 (since the area under the normal curve between these values is .95).

However, if on choosing a single sample at random, we find that the z score lies outside the range -1.96 to 1.96, we would conclude that such an event could happen with probability of only .05 (total shaded area in the Figure 7) if the given hypothesis were true. We would then say that this z score differed significantly from what would be expected under the hypothesis and we would be inclined to reject the hypothesis.

The total shaded area .05 is the level of significance of the test. It represents the probability of our being wrong in rejecting the hypothesis, i.e., the probability of making a type I error. Thus we say the hypothesis is rejected at a .05 level of significance or the z score of the given sample statistic is significant at a .05 level of significance.

The set of z scores outside the range -1.96 to 1.96 constitutes what is called the critical region or region of rejection of the hypothesis, or the region of significance. The set of z scores inside -1.96 to 1.96 could then be called the region of acceptance of the hypothesis, or the region of non-significance.

On the basis of the above remarks, we can formulate

the following rule of decision or test of hypothesis or significance.

> Reject the hypothesis at .05 level of significance if the z score of the statistic S lies outside the range -1.96 to 1.96 (i.e., either $z>1.96$ or $z<-1.96$). This is equivalent to saying that the observed sample statistic is significant at .05 level.
>
> Accept the hypothesis (or do not reject) otherwise.

4.4 Test of a Simple Hypothesis Against a Simple Alternative

When we are concerned with testing a hypothesis that a parameter, $\theta = \theta_0$, against the alternative that $\theta=\theta_1\neq\theta_0$, then this type of test of hypotheses is called a test of a simple hypothesis against a simple alternative.

In the preceding section we mentioned briefly the idea of acceptance and critical region. The following definition is very useful in assisting us in determining this critical region when testing a simple hypothesis against a simple alternative.

Definition 4.1 A test based on a random sample $X_1,\ldots,X_n$ from a density $f(x;\theta)$ for testing a simple hypothesis $H_0;\theta = \theta_0$ against a simple alternative, $H_1;\theta=\theta_1$ is a likelihood ratio test, if there exists a number k such that the test calls for accepting H_0 if $\lambda > k$, and rejecting H_0

if $\lambda \leq k$ and either if $\lambda = k$ where λ is the likelihood-ratio given by

$$\lambda = t(t_1,\ldots,x_n) = \frac{f(x_1;\theta_0)f(x_2;\theta_0)\ldots f(x_n;\theta_0)}{f(x_1;\theta_1)f(x_2;\theta_1)\ldots f(x_n;\theta_1)}$$

Let us consider this definition briefly. Since in an actual test θ_0 and θ_1 are fixed numbers, the inequality $\lambda \geq k$ for a fixed k defines a set of x's, i.e., for a fixed value of k there is a set of x's that satisfies the inequality $\lambda \geq k$. This set of x's is the acceptance region S_1, and the set of x's defined by $\lambda < k$ is the critical region (rejection region) S_2 for a particular value of k.

Let us consider an example where we know $k = e^{\frac{1}{2}}$. The value of k is usually determined in a given problem or we are able to determine it in a specific example.

Example 4.1 Suppose we have a distribution for a random variable X, such that X is distributed as a normal with a 0 mean and a variance equal to 1. To help illustrate the preceding definition, suppose we take a random sample of size one, say X, from the density. Our object is to try to determine the critical region for testing the null hypothesis, $H_0;\mu=-1$, against the alternative hypothesis, $H_1;\mu=0$. The likelihood ratio test gives

$$= \frac{f(x_i;-1)}{f(x_i;0)} = \frac{\frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}(x+1)^2}}{\frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}x^2}}$$

and simplifying we get

$$= \frac{e^{-\frac{1}{2}(x-1)^2}}{e^{-\frac{1}{2}(2x+1)}} = e^{-\frac{1}{2}(x^2+2x+1-x^2)}$$

$$= e^{-\frac{1}{2}(2x+1)}$$

As mentioned on the preceding page, we shall choose $k=e^{\frac{1}{2}}$ to illustrate the given definition; then $\lambda \geq k$ becomes

$$e^{-\frac{1}{2}(2x+1)} \geq e^{\frac{1}{2}} \quad \text{or} \quad e^{-x} \geq e.$$

Our problem is to determine the set of x such that $e^{-x} \geq e$. If we take the logarithm (base e) of both sides, we get the set $S_1 = \{ x \mid x \leq -1 \}$. So from this we see the acceptance region is all x which are less than -1, i.e., if the sample we took had a value less than -1, then we accept $H_0;\mu=-1$. We would not reject $H_0:\mu=-1$. If the sample value of x were greater than -1, we would reject $H_0:\mu=-1$ and accept $H_1:\mu=0$.

Let us consider another example where we use the likelihood ratio test.

Example 4.2 Suppose a random sample of size n is taken from a normal population with mean μ and variance 1. Suppose we wish to test the null hypothesis, $H_0:\mu=2$ against the alternative hypothesis, $H_1:\mu=0$.

The density of each X_i under the null hypothesis is

given by

$$f(x_i;2) = \frac{1}{(2\pi)^{\frac{1}{2}}} \quad e^{-\frac{1}{2}(x_i-2)^2} \quad , \; i=1,2,\ldots,n$$

and under the alternative the density of each X_i is given by

$$f(x_i;0) = \frac{1}{(2\pi)^{\frac{1}{2}}} \quad e^{-\frac{1}{2}x_i^2} \quad i=1,2,\ldots,n$$

thus

$$\lambda = \frac{f(x_1;2)f(x_2;2)\ldots f(x_n;2)}{f(x_1;0)f(x_2;0)\ldots f(x_n;0)} = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-2)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{n}x_i^2}}$$

$$= e^{-\frac{1}{2}\sum_{i=1}^{n}x_i^2-4x_i+4-x_i^2}$$

$$= e^{-\frac{1}{2}\sum_{i=1}^{n}(-4x_i+4)} = e^{-\sum_{i=1}^{n}(2-2x_i)} = e^{-2n+2\Sigma x_i}$$

$$= e^{2n\bar{x}-2n}$$

The likelihood ratio test consists of accepting H_0 if $\lambda = e^{2n\bar{x}-2n} > k$, which is equivalent to the statement that $2n\bar{x}-2n > \ln k$ or $\bar{x} > \frac{1}{2n}\ln k + 1$. Hence, employing the likelihood ratio test calls for accepting (not rejecting) $H_0:\mu=2$ if the value of $\bar{x}$, which we would compute from the sample, is greater than $\frac{1}{2n}\ln k + 1$ and rejecting $H_0:\mu=2$ and not rejecting $H_1:\mu=0$ if $\bar{x}$ is less than $\frac{1}{2n}\ln k + 1$, for a given value of k.

4.5 Composite Hypotheses

Thus far in the development of the theory of testing hypotheses we have concerned ourselves with the test of a simple hypothesis against a simple alternative. These are hypotheses of the form $H_0: \theta \in S_1$ with alternative $H_1: \theta \in S_2$, where S_1 and S_2 can be sets which contain one or more elements.

In this section a method will be given for constructing very useful tests for a simple hypothesis which can be extended to include some composite hypothesis. The method of construction depends upon the use of a theorem that was first proven and used by the two statisticians after whom it is named. The theorem, called the Neyman-Pearson lemma, will be stated without proof for a probability function, $f(x;\theta)$, of a single continuous variable and a single parameter. It should be noted that the following theorem applies only to a simple hypothesis against a simple alternative.

4.2 Neyman-Pearson Lemma If there exists a critical region A of size α and a constant k such that (where $X_1,\ldots,X_n$ is a random sample from $f(x;\theta)$),

$$\lambda = \frac{f(x_1;\theta_0)\ldots f(x_n;\theta_0)}{f(x_1;\theta_1)\ldots f(x_n;\theta_1)} \quad < K \text{ inside A}$$

and

$$\lambda = \frac{f(x_1;\theta_0)\ldots f(x_n;\theta_0)}{f(x_1;\theta_1)\ldots f(x_n;\theta_1)} \geq K \text{ outside } A$$

then A is the best critical region of size α.

Even though the Neyman-Pearson fundamental lemma applies specifically to problems involving a simple hypothesis against a simple alternative, we shall show in the following illustrations it can sometimes be used to advantage in composite hypotheses.

In the preceding lemma, we spoke of A as being the best region of size α. We mean it is the region in which $p(I) \leq \alpha$ and $1-p(II) = B(\theta)$ is a maximum for all θ in A. We shall refer to $1-p(II) = B(\theta)$ as the power of the test.

The usefulness and meaning of this lemma is best explained by means of illustrations, hence, consider the random variable X whose density function is given by $f(x;\theta) = \theta e^{-\theta x}$, $x \geq 0$. In order to discuss a problem somewhat more general than just testing a simple hypothesis, let us consider the hypothesis $H_0:\theta = \theta_0$, and the alternative $H_1:\theta < \theta_0$. We can change this hypothesis to one which is a simple hypothesis $H_0^*:\theta=\theta_0$, against the simple alternative $H_1^*:\theta = \theta_1 < \theta_0$. The corresponding likelihood functions are

$$L_0 = \prod_{i=1}^{n} f(x_i;\theta_0) = \theta_0^n \, e^{-\theta_0 \sum_{i=1}^{n} x_i}$$

and

$$L_1 = \prod_{i=1}^{n} f(x_i;\theta_1) = \theta_1^n e^{-\theta_1 \sum_{i=1}^{n} x_i} .$$

According to the Neyman-Pearson lemma, the region A is the region where

$$\frac{\theta_0^n e^{-\theta_0 \Sigma x_i}}{\theta_1^n e^{-\theta_1 x_i}} \geq k .$$

This inequality may be written in the form

$$e^{(\theta_1-\theta_0)\Sigma x_i} \leq \frac{1}{k} \left(\frac{\theta_1}{\theta_0}\right)^n$$

Taking logarithms, the inequality becomes

$$(\theta_1-\theta_0)\Sigma x_i \leq \ln \frac{1}{k} \left(\frac{\theta_1}{\theta_0}\right)^n$$

Since H_1 specifies that $\theta_1 < \theta_0$ dividing both sides by $\theta_1-\theta_0$ will reverse the inequality and yield

$$\Sigma x_i \geq \frac{\ln \frac{1}{k}\left(\frac{\theta_1}{\theta_0}\right)^n}{\theta_1-\theta_0} .$$

Now suppose for example we let n = 1 and $\theta_0 = 2$, and $\theta_1 = 1$, hence for this problem the value of the best critical region would be that part of the x axis to the

right of the point

$$x_0 = \frac{\ln\frac{1}{k}\left(\frac{\theta_1}{\theta_0}\right)^1}{\theta_1-\theta_0} = \ln(2k)$$

where k is chosen to make any desired probability.

Thus, the same critical region is used whatever the value of θ_1, so long as $\theta_1 < \theta_0$. The value of k necessary to produce the same $x_0 = \ln(2k)$, of course, depends upon the value of θ. This shows that $x_0 = \ln\frac{k}{2}$ gives the best critical region for testing the hypothesis $H_0: \theta=\theta_0$ against the alternative $H_1: \theta<\theta_0$. Thus, the Neyman-Pearson lemma, although designed to test a simple hypothesis against a simple alternative, can sometimes be used to solve a problem in which the alternative hypothesis is composite.

4.6 Likelihood Ratio Tests

When the Neyman-Pearson lemma fails to yield a best test, or when the hypothesis is composite rather than simple, it is sometimes necessary to place further restrictions on the class of tests and then attempt to find a best test from among this restricted class, or else it is necessary to introduce some other principle for obtaining good tests. In this section a second principle for constructing good tests will be introduced and discussed. Since any method for testing composite hypotheses will include the testing of simple hypotheses as a special

case, this principle will be introduced from the point of view of composite hypotheses.

We shall start one discussion with the consideration of a probability or density function which has more than one parameter. Suppose that a random variable X has a density function $f(x;\theta_1,\ldots,\theta_k)$ that depends upon k parameters. Let the composite hypothesis to be tested be denoted by $H_0:\theta_i=\theta_i (i=1,2,\ldots,k)$ where θ_i may or may not denote a numerical value. Thus, if there are two parameters, H_0 might be the hypothesis that $\theta_1 = 10$ with θ_2 unspecified, then $\theta_1' = 10$ and $\theta_2' = \theta_2$. With the aid of this notation, $f(x;\theta_1',\ldots,\theta_n')$ will denote the density of X when H_0 is true.

Let $\hat{\theta}_i$ denote the maximum likelihood estimator of θ_i for the likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta_1,\ldots,\theta_k)$, where the likelihood function is treated as a function of the parameters and the x_i are fixed. Similarly, let $\hat{\theta}_0$ denote the maximum likelihood estimator of θ_i when H_0 is true, that is, for the likelihood function $L(\theta') = \prod_{i=1}^{n} f(x_i;\theta_1',\ldots,\theta_n')$. Now let us consider the ratio $\lambda = \frac{L(\theta')}{L(\theta)}$. This is the ratio of two likelihood functions $L(\theta')$ and $L(\theta)$, where their parameters have been replaced by their maximum likelihood estimators. Since the maximum likelihood estimators are functions of the random variables $X_1,X_2,\ldots,X_n$, the ratio λ is a function of $X_1,X_2,\ldots,X_n$

only, and therefore is an observable random variable.

The denominator of λ is the maximum of the likelihood function with respect to all the parameters, whereas the numerator is the maximum only after some or all of the parameters have been restricted by H_0; consequently, it is clear that the numerator cannot exceed the denominator in value and therefore λ can assume values only between 0 and 1, inclusive. Now the likelihood function gives the probability density (or probability in case x is a discrete variable) at the sample point $X_1, X_2, \ldots, X_n$. Therefore, if λ is close to 1, it follows that the probability density (or probability) of the sample point could not be increased much by allowing the parameters to assume values other than those possible under H_0; consequently, a value of λ near 1 corresponds intuitively to considerable belief in the reasonableness of the hypothesis H_0. If, however, the value of λ is close to 0, it implies that the probability density (or probability) of the sample point is very low under H_0 as contrasted to its value under certain other possible values of the parameters not permitted under H_0, and therefore a value of λ near 0 corresponds to considerable belief in the unreasonableness of the hypothesis. If increasing values of λ are treated as corresponding to increasing degrees of belief in the truth of the hypothesis, then λ may serve as a statistic for testing H_0, with small values of λ leading to the rejection of H_0.

Since we have agreed to use λ as a test statistic, our next endeavor is to obtain a density function for the random variable λ. In many applied problems we can find the distribution of the statistic λ and thus make valid probability statements concerning λ.

Now suppose that H_0 is true and the density function of the random variable λ, say $g(\lambda)$, has been found. This is theoretically possible if the explicit form of $f(x;\theta_1',\ldots,\theta_k')$ is known. Suppose, further, that $g(\lambda)$ does not depend upon any unknown parameters. Then one can find a value of λ, say λ_0, such that

$$p(0 \leq \lambda \leq \lambda_0) = \alpha \tag{2}$$

The critical region of size α for testing H_0 by means of the statistic λ then is chosen to be the interval $0 \leq \lambda \leq \lambda_0$.

The preceding explanation of how likelihood ratio tests are constructed may be summarized in the following form.

Definition 4.2 To test a hypothesis H_0, simple or composite, use the statistic $\lambda = \frac{L(\theta')}{L(\theta)}$ and reject H_0 if, and only if, the sample value of λ satisfies the inequality $\lambda \leq \lambda_0$ where λ_0 is given by $p(0 \leq \lambda \leq \lambda_0) = \alpha$.

Although the use of λ as a statistic for testing hypotheses has been justified largely on intuitive grounds,

it can be shown that such tests possess several very desirable properties.

Our purpose is to acquaint the reader with some useful techniques in testing hypotheses. A more complete discussion of these concepts can be found in several mathematical statistics text, for instance, Mood and Graybill. However, the treatment given there is beyond the scope of our consideration.

CHAPTER V

REGRESSION

5.1 Introduction

Very often in practice we are concerned with the problem of determining whether there exists a relationship between two variables and, if a relationship does exist, what type of relationship it is. For example, weights of adult males depend to some degree upon their heights and areas of circles depend on their radii. We see from these examples that the relationships between variables are different. In the first example, the relation is probably linear, while in the second, the variables are related in such a way that one is proportional to the square of the other. In this chapter we shall be concerned with assumed relationships between variables and from these assumptions we will try to predict certain values of one variable when given a specific value for the other variable.

Knowledge, which is based upon experimental or observed information, has the distinguishing feature of being predictive knowledge. This means that the main value of scientific knowledge lies in the fact that, due to its very nature, it enables us to make predictions concerning

the behavior of observable phenomena. However, we must realize that when a scientist predicts the occurrence of a certain event, his prediction is quite different in nature from predictions made, for example, by prophets of ancient oracles. By this we mean a scientist does not claim to be able to predict with absolute certainty that a certain event will take place at some time in the distant future. As a matter of fact, he does not claim to be able to predict anything whatsoever with absolute certainty. Instead, he asserts his predictions in terms of probabilities, implying that he is satisfied if his predictions come true a certain percentage of the time or, better, he aims in his predictions for a success ratio which is as high as possible.

Very often in practice we are faced with the problem of determining whether certain variables are linearly related. For example, if x_i represents the score a high school student achieves on a mathematics test, and y_i represents the score achieved on a science examination, we might expect these variables to be related linearly, i.e., if a student scores well on a mathematics test, he would probably be capable of high achievement on a science test. Also, for another example, we might let x_i be a student's score on the college entrance examination and y_i represent his grade point average. Similarly, we might expect these variables to be related in a linear manner, i.e., if a student scores high on the examination, we would expect him to excel in his college work, while if a student made a low

score, we would expect him to encounter difficulties. Thus we would say that x and y are directly related. One of the problems we shall encounter is trying to determine this relation, i.e., determining a function $y = f(x)$. If we can find this function $y = f(x)$, then for a given x we can predict a y value. In the preceding example, we would be given a score on the entrance examination and we could predict the student's success in his college work.

When we have at our disposal information on two related variables, it seems natural to seek a way of expressing the form of the functional relationship. It is also desirable that we know the accuracy of this relationship. That is, we not only seek a mathematical function which tells us how the variables are interrelated, but also we wish to know how closely the values of one variable can be predicted if we are given the values of the associated variables. The techniques which shall be used to accomplish these two objectives are known as regression methods and correlation methods. Regression methods will be used to determine the "best" functional relationship between the variables, while correlation methods are used to measure the degree to which the different variables are associated.

In any analysis, it is hoped the assumed function represents some basic, or causal, mechanism associated with the factor under investigation. Because of the frequent uncertainty about basic variables and basic mechanisms, a word of warning must be sounded relative to the interpre-

tation of analyses involving these variables. The warning is: just because a particular functional relationship has been assumed and a specific computational procedure followed, do not assume that a causal relationship exists among the variables. We are actually implying that just because a particular function has been found that is a good fit to a set of observed data, we should not necessarily infer that a change in one variable causes a change in another variable. A classical example which illustrates this is: It can be shown that over a period of years there exists a linear relationship between teacher's salaries and the consumption of liquor. However, it seems reasonable we would agree that an increase in teacher's salaries had little, if any, effect upon the liquor consumption. During this period of time there was a steady rise in the wages and salaries of all types and a general upward trend of good times. Under such conditions, teacher's salaries and liquor consumption would also increase, even though no causal relationship exists.

Let us suppose that we know the average grades of six high school seniors who graduated two years ago, and also the grade point average attained the first year in college. Such an illustration may be seen in the table given on the following page.

	High School Average	Grade Point Average
Student A	90	2.8
Student B	65	2.1
Student C	95	3.7
Student D	69	2.9
Student E	76	2.6
Student F	80	2.2

Our problem now is to fit a curve to this data which will give us the best possible predictions. Since there is, logically speaking, no limit to the number of lines which can be drawn on a piece of paper, it is evident that we will need a criterion on the basis of which we can point to a single line as the one which presents us with the best fit to our data. This choice is not usually self-evident except in the special case where all points actually do fall in a straight line. Since we can hardly expect this to happen often when dealing with experimental data, we must be satisfied with a straight line which, although it cannot possibly go through all points, will have some less perfect, yet still desirable properties. As mentioned before, we are interested in determining a curve which will give us a way of predicting a value of one of the variables when we are given a specific value for the other one.

The first thing we must do when we want to fit a straight line to the data given above is to check whether it is at all reasonable to suppose that a straight line

will give a good fit. A very convenient way is to plot the points representing our data as we have done in Figure 8 below. It is almost impossible to decide if it is reasonable to treat the two variables as if they were linearly related when we have only 6 data points, however, if we examine the data points, there appears to exist a linear relation between a student's average grade in high school and his grade point in college.

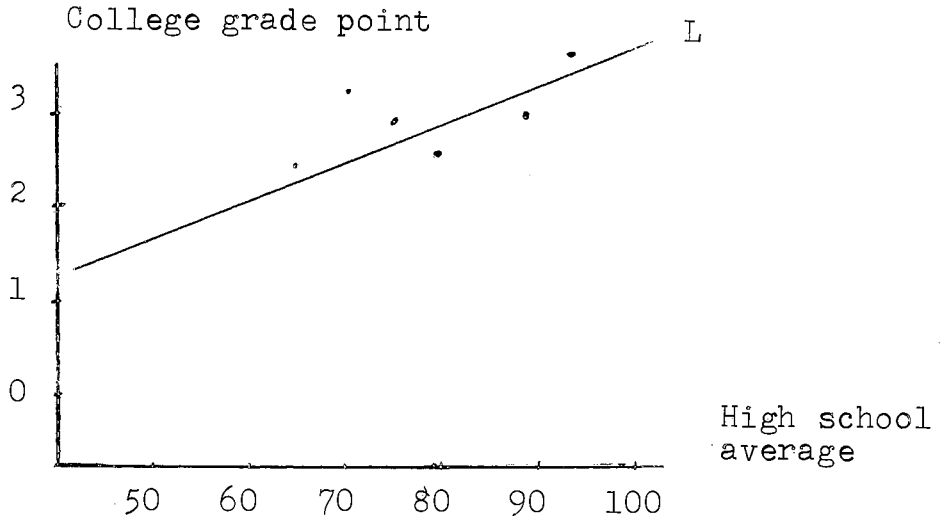

Figure 8

First, let us draw a more or less freehand line L which indicates the approximate linear relation between the variables as shown in Figure 8. We might ask ourselves how good our predictions would have been if we had actually used line L for the prediction of the college grade points for the given six students when we knew their high school grade averages. The predicted grade point for a given student could be found by considering the student's average and determining the functional value of that

particular average.

If we had known this line represented the relationship between the student's high school grade averages and their grade point average before our six students entered college, we could have used it to predict their expected success in college. For example, the student who had a high school average of 80 would, according to Figure 8, have had a predicted grade point average of 2.5. But from our data we see the student who had an average of 80 achieved a grade point average of 2.2. Consequently, the error of this prediction would have been 2.5 - 2.2 = .3.

Geometrically, the error of the prediction is measured by the vertical deviation (distance) from the point representing the actual data to the line L which we used for our prediction. In Figure 9, below, this deviation is given by the distance from A to B.

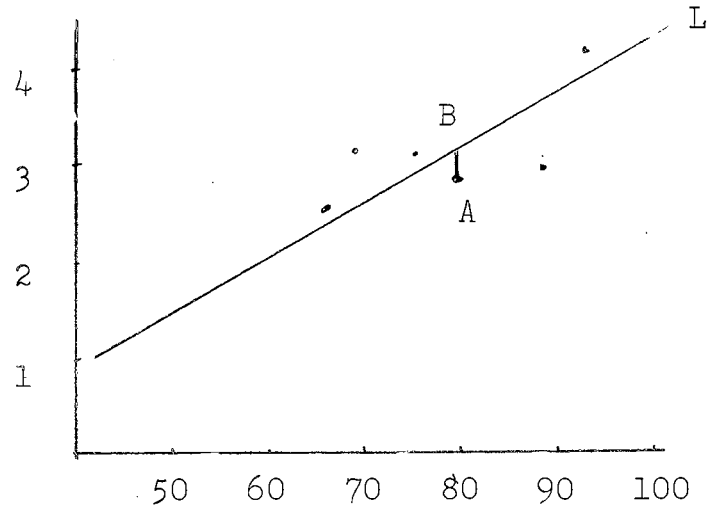

Figure 9

If we apply this method of prediction to each of the six students used in this example, the six corresponding errors of our predictions are given by the six vertical deviations from the line L. (See Figure 9). If we denote the observed grade point of the ith student by the symbol y_i and the corresponding predicted grade point by y_i', the error variance of the six predictions is given by the expression

$$\frac{\sum_{1}^{6} (y_i - y_i')^2}{6}$$

which is just the average of the squared vertical deviation from the line L.

It should be evident that if we assume a functional relationship between two variables then we should strive to find the particular function in which the error variance is as small as possible. We now have at our disposal a criterion for the goodness of the fit of a straight line. It seems quite reasonable to require the resulting errors to be small if we expect the line to be a good fit for the given data.

The line having the distinguishing property that the sum of squares of vertical deviations, the error variance of the y's, is minimum is called the regression line of y on x. If it had been desirous, instead, to predict x in terms of y we could in a similar manner have asked for a line which minimizes the sum of squares of deviations of the x's, and we would have obtained the regression of x

on y. From the discussion above, it seems reasonable that it is completely arbitrary which variable is called x and which is called y, hence, we shall simplify our work by limiting our discussion to those lines which minimize the sum of squares of the vertical deviations, i.e., we shall consider only regression lines of y on x.

Let us now suppose that we have a sample of size n of pairs of measurements, say $(x_1,y_1)\ldots,(x_n,y_n)$, which might represent the weights and corresponding heights of n individuals of about the same age, and suppose further that we are convinced that there exists a linear relationship between the x's and y's. Our problem now is to try to determine the parameters m and b in the linear function $y=mx+b$, so that the line has the property that the sum of squares of the vertical deviations will be a minimum. This means that we must find numerical values for the two constants m and b which appear in the equation $y=mx+b$ so that the line which is thus obtained has the stated properties.

A certain function has been postulated as being the "best" expression of the true state of affairs in the population, and it is now necessary to estimate the parameters of the function. The determination of these estimates and thus the specification of a particular function is commonly referred to as curve fitting. How do we go about fitting a curve to a set of data? That is, how are the estimators of the parameters obtained? Again we are faced with the problem of choosing among several methods of estimation.

The approach which we shall take should, of course, provide us with the "best" estimates. Let y_i' represent the predicted value corresponding to y_i. This value must be obtained from the equation

$$y_i' = mx_i + b$$

and if we substitute this predicted value into the expression for the error variance we can rewrite the expression as

$$\sum_1^n \frac{(y_i - mx_i - b)^2}{n} .$$

This expression is called the error variance about the regression line, and it shall be denoted by s_e^2 provided, of course, that the two constants m and b are such that we do have a regression line. We note that in the expression s_e^2 the only things which are unknown are the m and the b since we were given n pairs (x_i, y_i) of measurements which we assumed to be known from the start.

As mentioned earlier in this chapter, we are in search of parameters m and b which will minimize the expression s_e^2. To minimize s_e^2 we shall find its partial derivatives with respect to m and then with respect to b yielding two equations which can be solved for m and b.

$$\frac{\partial s_e^2}{\partial m} = \frac{-2\Sigma x_i (y_i - mx_i - b)}{n}$$

and

$$\frac{\partial s_e^2}{\partial b} = \frac{-2\Sigma (y_i - mx_i - b)}{n} .$$

If we set these expressions equal to zero the resulting equations are

$$\Sigma x_i(y_i - \hat{m}x_i - \hat{b}) = 0 \qquad \text{and} \qquad \Sigma(y_i - \hat{m}x_i - \hat{b}) = 0 .$$

From the second equation we get

$$(1) \qquad \Sigma y_i - \hat{m}\Sigma x_i = n\hat{b}$$

which gives

$$\hat{b} = \frac{\Sigma y_i}{n} + \frac{-\hat{m}\Sigma x_i}{n} .$$

From the first equation, summing over n, we get

$$(2) \qquad \Sigma x_i y_i - \hat{b}\Sigma x_i = \hat{m}\Sigma x_i^2 .$$

Now if we substitute into this equation for $\hat{b}$ from (1) we get

$$\Sigma x_i y_i - \left(\frac{\Sigma y_i}{n} + \frac{-\hat{m}\Sigma x}{n}\right) \Sigma x_i = \hat{m}\Sigma x_i^2$$

and

$$\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n} = \hat{m}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] .$$

Solving for $\hat{m}$ we find

$$\hat{m} = \frac{\Sigma x_i y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}} .$$

Therefore

$$\hat{m} = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2} .$$

Now considering equation (2) again, if we substitute the expression for $\hat{m}$ into equation (2) we get

$$\hat{b} = \frac{\Sigma y_i}{n} - \frac{(\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n})}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} \quad \frac{\Sigma x_i}{n}$$

$$\hat{b} = \frac{\Sigma y_i \left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] - \Sigma x_i (\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n})}{n(\Sigma x_i^2) - (\Sigma x_i)^2}$$

$$\hat{b} = \frac{\Sigma y_i \Sigma x_i^2 - \frac{(\Sigma y_i)(\Sigma x_i)^2}{n} - (\Sigma x_i)(\Sigma x_i y_i) + \frac{(\Sigma x_i)^2 \Sigma y_i}{n}}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$\hat{b} = \frac{\Sigma y_i \Sigma x_i^2 - (\Sigma x_i)\ \Sigma x_i y_i}{n(\Sigma x_i^2) - (\Sigma x_i)^2} \quad .$$

So if we are given n pairs of measurements and if we assume we have a linear relationship existing between the variables then we can estimate the parameters of $y = mx+b$ and obtain an estimate of that linear function by computing $\hat{m}$ and $\hat{b}$ when we are given a specific example.

Returning now to the illustration mentioned earlier concerning the relationship between high school grade

averages and the grade point average obtained in college, we can use the expressions developed on the preceding page to find the actual equation of the regression line of y on x. The necessary calculations are usually performed by means of a table similar to the following:

x	y	x^2	xy
90	2.8	8100	252.0
65	2.1	4225	136.5
95	3.7	9025	351.5
69	2.9	4761	200.1
76	2.6	5184	197.6
80	2.2	6400	176.0
471	16.3	37695	1313.7

Thus we have

$$n = 6$$

$$\Sigma x_i = 471$$

$$\Sigma y_i = 16.3$$

$$\Sigma x_i^2 = 37695$$

$$\Sigma x_i y_i = 1313.7 \ .$$

If we substitute these values into the expressions for $\hat{m}$ and $\hat{b}$ we get

$$\hat{m} = \frac{6(1313.7)-(471)(16.3)}{6(37695) - (471)^2} = \frac{204.9}{4329} = .047$$

and

$$\hat{b} = \frac{(16.3)(37695)-(471)(1313.7)}{6(37695) - (471)^2} = -\frac{4324.2}{4329} = -.999$$

and we can write the regression line as $y = .047x - .999$.

Now that we have determined the regression line of y on x we can predict a student's success (his grade point) in college. So if we are given a student's high school average we can substitute this x value into the equation $y = .047x - .999$ and get an estimate or prediction of his college grade point average. For example, if a student had a high school average grade of 70, his predicted grade point average, y', would be found by calculating

$$y' = (.047)(70) - .999 = 2.291 .$$

Since the expressions given for $\hat{m}$ and $\hat{b}$ are somewhat tedious to calculate, it is often preferable to calculate $\hat{m}$ and $\hat{b}$ by using the equation (1) and (2). After dividing (1) and (2) by n we get

$$\bar{y} - m\bar{x} = \hat{b}$$

and

$$\Sigma x_i y_i - b\Sigma x_i = \hat{m}\Sigma x_i^2 .$$

We now have two equations in two unknowns, $\hat{m}$ and $\hat{b}$, which can be solved for the unknown after making substitutions for the known values determined from the given sample. In our example, which was given on the preceding page, the resulting equations would be $2.71 - 78.5\,\hat{m} = \hat{b}$ and

$1313.7 - 471\hat{b} = 37695\hat{m}$. Solving we get

$$\hat{m} = .047$$

and

$$\hat{b} = -.999$$

which agrees with our previous results.

To help understand the concept of linear regression and curve fitting, let us consider another example. Suppose you are a farmer whose chief income is from alfalfa hay. Suppose further that your farm is in a suitable area which is accessible to irrigation, if you feel it is worthwhile to finance the expenses of an irrigation system. You decide to consult someone at the state university whom you feel might have access to some data concerning the relationship between hay yields and irrigation. It turns out that the college has recently conducted an experiment on an experimental farm and the following data is obtained concerning the hay production in tons relative to the number of inches of water which was applied. The data obtained is given in the following table:

Water (x) (treatments)	12	18	24	30	36	42	48
Yields (y)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

If we plot this data, there appears to exist roughly a linear relationship between the yields of alfalfa hay and

the number of inches of water applied, as shown below in Figure 10.

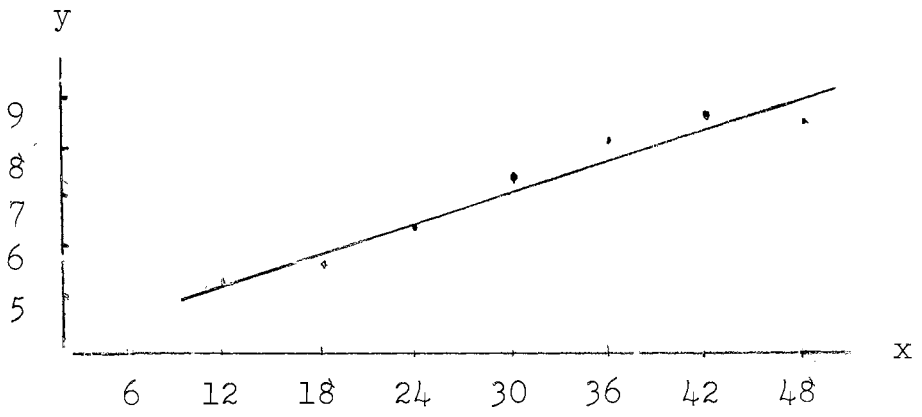

Figure 10

Since an approximate linear relation appears to exist, it should suffice to use a linear function of x. Thus the problem of prediction first requires the solution of the problem of fitting a straight line to the set of points, i.e., we must determine the values of the constants m and b in the equation $y = mx + b$. Using the methods already developed, we find that $\hat{m} = .10$ and $\hat{b} = 4.0$. Hence, the equation of the regression line is given by

$$y = .10x + 4.0.$$

Thus there seems to exist a linear relationship between the hay yield and the amount of water applied.

In fitting a straight line to a set of points, as in the preceding illustration, it is intuitively assumed that the resulting line is an estimate of a theoretical line of regression. This regression line, being an estimate of the actual or theoretical regression line, leads us to ask the

question, "How good an estimate of the true regression line is our estimate?" . A thorough investigation of this question would lead us to some advanced techniques in mathematics and statistics which is beyond the scope of our treatment here. However, we shall discuss some of these concepts from an intuitive point of view without employing a great deal of mathematical rigor.

To consider this idea of "goodness of estimate", let us return to the first example concerning the relationship between the high school student's grade average and his college grade point. Before we consider adopting the equation

"Predicted college grade point=.047(high school average)-.999" even hypothetically as a method of predicting a student's future success in college, we must first check how accurate we can expect the resulting predictions to be. To help illustrate this idea, let us assume the above formula was known at the time the six students, who were the subjects of our investigation, entered college. This assumption might seem somewhat ridiculous since we actually calculated the value of $\hat{m}$ and $\hat{b}$ on the basis of the records which these same six students established in college, and it would be impossible to know this formula in advance. What we are actually saying is that we are assuming $y = .047x - .999$ is the theoretical regression line. However, let us assume in spite of this obvious objection that as we said, to help illustrate this idea, we did have this formula when

the six students entered college. We could then have used it to predict the grade points which our students could have been expected to attain. Substituting their high school averages into the equation $y = .047x-.999$ we could have calculated the predicted value of y which, together with the indexes which the students actually obtained, are shown in the following table.

	Actual grade point y	Predicted grade point y'
Student A	2.8	3.23
Student B	2.1	2.06
Student C	3.7	3.47
Student D	2.9	2.24
Student E	2.6	2.57
Student F	2.2	2.77

If the theoretical regression line is actually a linear function of the form $y = mx + b$ the values of $\hat{m}$ and $\hat{b}$ which we calculate from a set of experimental observed data and since the calculation will change when we use a different sample, we must consider $\hat{m}$ and $\hat{b}$ as random variables. Since $\hat{m}$ and $\hat{b}$ are random variables it is possible to determine the distribution of these statistics and thus make probability statements concerning them. However, we will not endeavor to enter into a discussion concerning the distribution of these random variables.

When discussing the goodness of the estimates $\hat{m}$ and $\hat{b}$, we must realize that $\hat{m}$ and $\hat{b}$, just as most estimates, increase in accuracy for increasing values of n, i.e., good estimates of m and b are obtained only if we have a large number of pairs $(x_i y_i)$ of measurements of the two variables x and y. We should realize in actual practice, if we want to obtain good estimates, we would seldom use samples as small as the ones employed here to illustrate the techniques used in the computation of these estimates.

As mentioned earlier, a formal study of the accuracy of estimates of the regression coefficients m and b is considerably beyond the scope of our treatment here. However, as long as we base the equations which we intend to use for our predictions on reasonably large samples, our estimates of m and b will usually be sufficiently close to the true values of the regression coefficients. Therefore, if we are dealing with large samples, there would seem to be no serious objection to evaluating the goodness of the predictions by applying the equation to the identical data from which it was originally obtained. We can use as a test statistic

$$s^2 = \frac{\sum_{1}^{n} (y_i - y_i')^2}{n}$$

where y_i and y_i' are the actual and the predicted responses, respectively. If s^2 is large we would conclude that our estimates for m and b are bad, while if s^2 is small we would conclude that our estimates are good.

Our method, which we have employed in the determination of $\hat{m}$ and $\hat{b}$, has consisted of minimizing the sum of squares of deviations from the straight lines, and hence, this method is referred to as the method of least squares. This method enables us to select one line as the line which provides us with the best fit to a given set of points. It is really in this sense that we define what we mean by a good fit. Although we have used the method of least squares only for the determination of a best-fitting straight line, it can also be used to give us best-fitting curves in general even though their equations may be of a much more complicated nature.

We now have at our disposal estimates of the parameters m and b in the linear relationship $y = mx + b$ which we hope are good estimates of these parameters. As mentioned earlier, it is desirable that good estimates have the property of being unbiased. Let us see if these estimates possess this property. If we let $\hat{m}$ and $\hat{b}$ represent the estimates of m and b, respectively, we want to know if $E(\hat{m}) = m$ and $E(\hat{b}) = b$.

Consider

$$E(\hat{m}) = E\left[\frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma x_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2}\right] .$$

Thus we have

$$E(\hat{m}) = E\left[\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}\right]$$

$$= \frac{1}{\Sigma(x_i-\bar{x})^2} E\left[\Sigma(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})\right]$$

$$= \frac{1}{\Sigma(x_i-\bar{x})^2}\left[\Sigma\, x_i E(y_i) - x_i E(\bar{y}) - \bar{x}E(y_i) + \bar{x}E(\bar{y})\right].$$

But we assumed the linear relationship such that $E(y_i) = b+mx_i$. Thus we have

$$E(\bar{y}) = E\left[\frac{1}{n}\Sigma y_i\right] = \frac{1}{n}\Sigma E(y_i)$$

$$= \frac{1}{n}\Sigma(b+mx_i) = \frac{1}{n}(nb+m\Sigma x_i)$$

$$= b + m\bar{x}.$$

Substituting this back into the expression above we have

$$E(m) = \frac{1}{\Sigma(x_i-\bar{x})^2}\left[\Sigma\left(x_i(b+mx_i) - x_i(b+m\bar{x}) - x(b+mx_i) + \bar{x}(b+m\bar{x})\right)\right].$$

Thus

$$E(m) = \frac{1}{\Sigma(x_i-\bar{x})^2}\left[\Sigma\left(mx_i^2 - 2m\bar{x}x_i + m\bar{x}^2\right)\right].$$

Hence we have

$$E(\hat{m}) = \frac{m}{\Sigma(x_i-\bar{x})^2}\ \Sigma(x_i^2 - 2\bar{x}\,x_i + \bar{x}^2)$$

$$= \frac{m\Sigma(x_i-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}$$

$$= m \quad .$$

Therefore $E(\hat{m}) = m$, i.e., $\hat{m}$ is an unbiased estimate of the parameter m. Also one expression for $\hat{b}$ is $\hat{b}=\bar{y}-\hat{m}\bar{x}$. Thus

$$E(\hat{b}) = E(\bar{y})-E(\bar{x}\hat{m})$$
$$= b+m\bar{x} - \bar{x}\,E(\hat{m}) \quad .$$

But we have shown above that $E(\hat{m}) = m$, hence we have

$$E(\hat{b}) = b+m\bar{x} - \bar{x}m = b \quad .$$

Thus $\hat{b}$ is an unbiased estimate of b. So we see that these estimators possess the desirable property of being unbiased.

5.2 Correlation

In the last section we devoted a great amount of time to the problem of finding the regression line and the error variance which we employed as a measuring device to determine the goodness of the resulting predictions, i.e., the degree to which the regression line fits a given set of

measurements. However, we observed the error variance was often difficult to compute and the goodness of fit was dependent upon the units which were used, i.e., if the error variance came out to be 10 feet this would seem sufficiently large to worry about, however, if the units under consideration were miles, then it would probably not be sufficient to cause much concern. Hence, we search for a measuring device which is independent of the particular units used in the data, i.e., we want a method which will give us a number so we can decide immediately whether it is sufficiently large or not. We shall now define another measure of the goodness of the fit of the regression line, which is called the coefficient of correlation defined by

$$r=\pm\sqrt{1-s_e^2/s_y^2}\ .$$

A considerable amount of time will now be devoted to explaining the quantities r, s_e^2 and s_y^2 .

We shall denote the error variance about the regression line by s_e^2 and define it by the expression

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i-y_i')^2}{n}$$

where y_i' is the predicted value and y_i is the observed value. We can easily see that s_e^2 depends on the scale of measurement of y and it can therefore happen that the regression line will provide us with a very poor fit even

though s_e^2 is small, simply because the quantities y are small. Similarly, it can also happen if the y's are large. The error variance may be large in spite of the fact that we have an excellent fit. This obvious shortcoming of the error variance about the regression line as a measure of the goodness of fit suggests a modification which leads us to the so-called "coefficient of correlation". This new measure can be understood readily if we define it as a measure which is a combination of the following two methods of prediction.

Method 1. We shall predict each y by means of the regression line $y_i' = mx_i + b$ which was determined from the identical set of data which will also be used to evaluate the goodness of the resulting prediction.

Method 2. We shall predict for each y that it is equal to the mean of the y_i, i.e., our predictions are now based on the formula $y_i' = \bar{y}$, where $\bar{y}$ is the mean of the same set of data which is used in method 1.

The appropriateness of method 1, discussed in the last section, is simply the error variance about the regression line given by

$$s_e^2 = \frac{\Sigma(y_i - mx_i - b)^2}{n}$$

while the appropriateness of method 2 is expressed by the error variance

$$s_y^2 = \frac{\Sigma(y_i - y_i')^2}{n} = \frac{\Sigma(y_i - \bar{y})^2}{n}$$

which is simply the sample variance of the y's.

The errors which are made by these two methods of predictions are reflected by the two quantities s_e^2 and s_y^2. We shall now show how these two quantities may be employed to define a new measure of the goodness of fit of the regression line.

Let us now consider an example to help us understand the merit of these two methods. Suppose we are given the following data which shows the personal savings of people of the United States and the number of strikes in eight different years.

Savings in billions of dollars	Number of strikes
2.9	2862
4.9	2509
10.9	4288
16.1	2968
17.5	4956
19.0	4750
11.9	4985
3.0	3693

We might suspect that if people have a large amount of personal savings that there would probably be a larger number of strikes, i.e., there exists a linear relationship between personal savings and the number of strikes.

Suppose you are called upon to predict the number of strikes for any one year on the basis of the total savings recorded for that year. If we use both methods 1 and 2, we must first calculate the regression line and the mean of the y's, where y represents the number of strikes. After some necessary calculations, we find the equation of the regression to be $y' = 95x + 2852$, while the mean of the y's is $\overline{y} = 3876$, hence using method 2 we see that $y' = 3876$.

Now to see the relative merit of the two methods of prediction, let us compare the vertical deviations from the two lines.

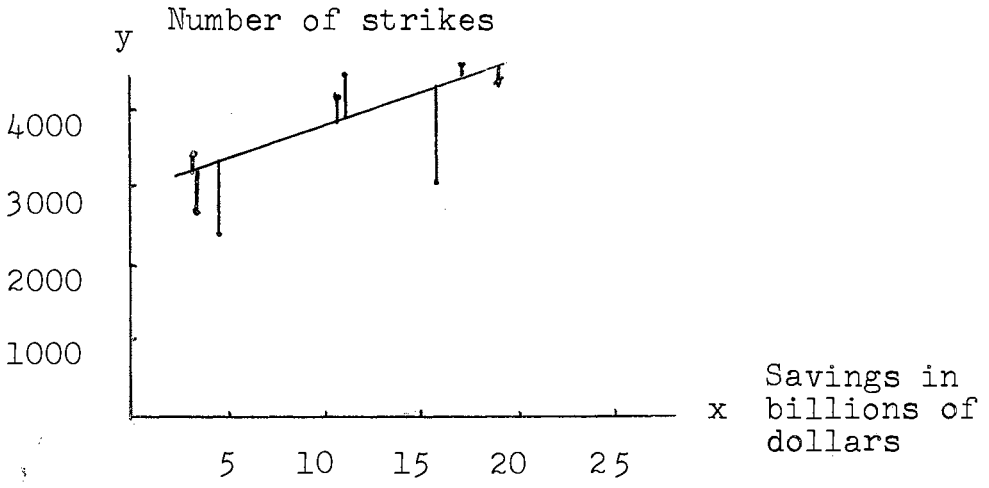

Figure 11

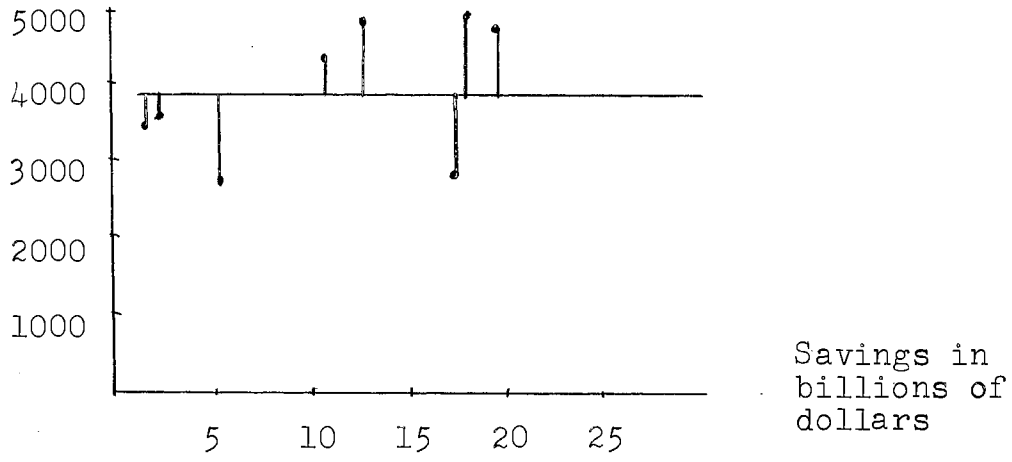

Figure 12

Using the second method to predict the number of strikes for a given year, we just compute the average of the given number of strikes in the preceding years, i.e., $y' = \bar{y}$. Now to decide which method seems to have more merit, we must remember our criterion for deciding when a method is good. We recall that, for a method to be a good one, it must minimize the expression

$$\frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n} .$$

If we consider the two Figures, 11 and 12, we might get the impression that the deviations in the Figure 11 are slightly smaller than in Figure 12, implying that method 1 is slightly more accurate than method 2. To convince

ourselves that this actually is the case, let us calculate the two error variances, s_e^2 and s_y^2. Using the expression given earlier for s_e^2 and s_y^2 we find the values to be

$$s_e^2 = 549{,}012$$

and

$$s_y^2 = 887{,}200 .$$

This parallels our previous rough judgment that the first method of prediction was slightly better than the second. Before we try to decide how much better the first method is, let us consider another illustration.

Suppose seven students, whose I.Q.'s are known, are given a test and the test scores and the I.Q.'s are as follows:

Test Scores	I.Q.
22	113
27	116
32	119
37	122
42	126
47	129
52	131

Using the same methods as before, we must determine the regression line, of y or x where y represents the I.Q. and x represents the score on the test. After some calculations

we find these to be as follows:

$$\bar{y} = 122.3$$

and

$$y' = .62x + 99.34 .$$

The errors which are made by methods 1 and 2 are reflected by the vertical deviations from the line in the figures below.

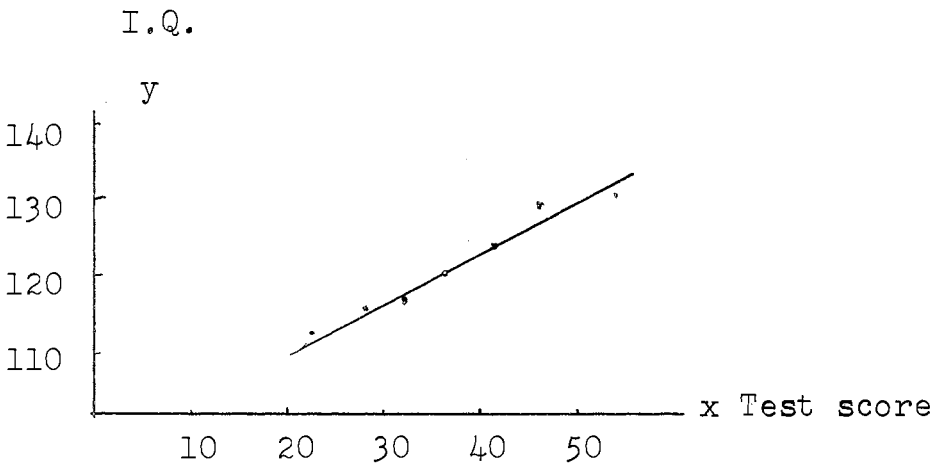

Figure 13

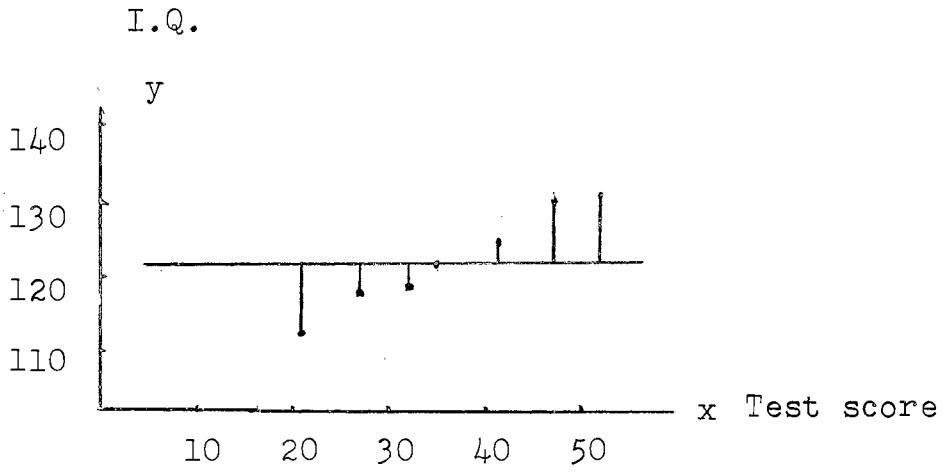

Figure 14

In Figure 13 we observe a small amount of difference between the errors of methods 1 and 2, however in this illustration we see there exists a marked difference between the deviations using the two different methods. The two actual error variances can be computed by methods mentioned earlier and are found to be

$$s_e^2 = .1584$$

and

$$s_y^2 = 38.0 .$$

This indicates strongly that the predictions which were based on the regression line were far superior to those which were based on the mean $\bar{y}$.

If we recall how we found the regression line, we found the coefficients to minimize the expression s_e^2, hence, if we use any other estimate for a predicted value other than $y = mx + b$, we see that s_e^2 never exceeds s_y^2. Also, we note that if the regression line fits a set of data very closely, the error variance of method 1 should be much smaller than method 2. If, on the other hand, the fit of the regression line is poor, method 1 provides us with only a slight improvement over method 2. This type of reasoning and the two preceding illustrations suggest a comparison of the given two methods of prediction might provide us with a new measure of the goodness of the fit for the regression line, which does not actually depend on the scale of the y's.

Now that we have decided to employ s_e^2 and s_y^2 to define a new measure of the goodness of the fit of the regression line, we must decide how to define this new measure. It has been the custom to define the following measure of goodness of fit of the regression line as the coefficient of correlation as defined earlier by the expression

$$r = \pm\sqrt{1-s_e^2/s_y^2} \ .$$

As noted before, if the fit is poor, s_e^2 will be almost as large as s_y^2, so the ratio s_e^2/s_y^2 will be close to 1 and the coefficient of correlation r will be close to 0. However, if the fit is good, s_e^2 will be much smaller than s_y^2 and the ratio s_e^2/s_y^2 will be close to 0. Thus the coefficient of correlation will be close to either plus or minus 1.

The coefficient of correlation can be computed by using the expression

$$r = \pm\sqrt{1-s_e^2/s_y^2} \ .$$

However, to compute r by this method we note it is necessary to first compute an estimate using the expression given earlier for $\hat{m}$ and $\hat{b}$ in the regression line. Since the computation of these regression coefficients, $\hat{m}$ and $\hat{b}$, involves a considerable amount of work, we search for an expression for r which is easier to compute. To avoid a good part of this work, we shall now give an alternative

expression for r which can be shown to be equivalent to the expression for r given on the preceding page. This expression is as follows:

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}} .$$

Many times r is also written in the form

$$r = \frac{\Sigma(x-\bar{x})(y_i-\bar{y})}{n\, s_x s_y}$$

where s_x^2 and s_y^2 are the sample variances of x and y, respectively.

The coefficient of correlation, with which we have been concerned in the preceding remarks, is by far the most widely used measure of the strength of the linear relationship between two variables. It not only expresses the goodness of the fit of the regression line, but it also tells us whether or not it is reasonable to say that there exists a linear relationship (correlation) between the two variables x and y. The magnitude of r determines the strength of the relationship, whereas the sign of r tells one whether y tends to increase or decrease, with x, i.e., if y increases as x increases, or decreases as x decreases, then r will be positive, while if y decreases as x increases, or increases as x decreases, r will be negative. If the numerical value of r, which has been computed from

a certain set of data, is close to 0, we say that the relationship is weak or nonexistent. If r is close to either + 1 or - 1, however, we say that the relationship is strong, with the tacit understanding that we are referring to a linear relationship and nothing else.

Let us now calculate r for the two examples presented earlier in this chapter. We must first determine whether r is positive or negative. As mentioned on the preceding page, r is positive if x increases as y increases, hence from the data and regression line we would surely agree r should be positive. In the first example, employing the expression

$$r = \sqrt{1 - s_e^2/s_y^2}$$

we get

$$r = \sqrt{1 - \frac{549,012}{887,200}} = .62$$

whereas in the second illustration the value of r is given by

$$r = \sqrt{1 - \frac{.158}{38.0}} = .998 .$$

The value of r for the two expressions shows what we had suspected, namely, that the relationship between the test score and I.Q. is very strong (at least for this particular group of students who were included in this study) while the relationship in the first illustration does not show a strong linear relationship. We should note that we used

in both illustrations samples which were too small to permit us to make far-reaching generalizations, thus confining ourselves to descriptive statistics, we can safely say only that the second set of data fits a regression line much better than the first. This does not reveal to us a great deal of information about the variables. However, it is about as far as we can go without assuming the risk of inductive generalizations.

It is always advisable to be extremely careful in the analysis and interpretation of the value of r which has been calculated from a given sample. The interpretation of a correlation coefficient as a measure of the strength of the linear relationship between two variables is a purely mathematical interpretation and is completely devoid of any cause or effect implications. The fact that two variables tend to increase or decrease together does not imply that one has direct or indirect effect on the other. Both may be influenced by other variables in such a manner as to give rise to a strong mathematical relationship. A classical example, mentioned earlier, illustrates this. It can be shown that over a period of years the correlation coefficient between teacher's salaries and liquor consumption is .90. However, during this period of time, there was a steady rise in the wages and salaries of all types and a general upward trend of good times. Under such conditions, teacher's salaries and liquor sales would also increase. Moreover, the general upward trend in wages and buying

power would be reflected in increased purchases of liquor. Thus, this high correlation merely reflects the common effect of the upward trend of the two variables. Hence, correlation coefficients must be handled with care if they are to give sensible information concerning relationships between pairs of variables. Success with correlation coefficients requires familiarity with the field of applications as well as with their mathematical properties.

Let us consider an example in which we will calculate the correlation coefficient of two variables. Suppose we are given the following data where x represents the father's height in inches, while y represents the son's height.

x	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70

We would probably expect that there would exist a strong linear relation between these two variables. Let us use the expression for r given by

$$r = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2)\,(n\Sigma y_i^2 - (\Sigma y_i)^2)}} .$$

To compute these different values, let us construct another table which will be useful in computing r. Such a table is illustrated below.

x	y	x^2	xy	y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\Sigma x = 800$	$\Sigma y = 811$	$\Sigma x^2 = 53,418$	$\Sigma xy = 54,107$	$\Sigma y^2 = 54,849$

Using the calculations we see that r is given by

$$r = \frac{(12)(54,107) - (800)(811)}{\sqrt{((12)(53,418)-(800)^2)\ ((12)(54,849)-(811)^2)}}$$

$$= .7027.$$

We can also determine the regression equation very easily by using these calculations. We see that $\hat{m}$ is given by

$$\hat{m} = \frac{n\Sigma x_i y_i - (\Sigma y_i)(\Sigma x_i)}{n\Sigma y_i - (\Sigma y_i)^2}$$

$$= 1.036$$

and

$$\hat{b} = \frac{\Sigma y_i \Sigma x_i^2 - (\Sigma x_i)(\Sigma y_i)}{n(\Sigma x_i^2) - (\Sigma x_i)^2}$$

$$= -3.38 .$$

Hence the regression line is $y' = 1.036x - 3.38$. So if we are given that a father's height is 70 inches, we would predict the son's height to be $y = (1.036)(70) - 3.38$, which yields $y = 69.14$.

There is little difficulty in explaining the meaning of the coefficient of correlation when it is either 0 or ± 1. Since if $r = 0$, we can see that the fit of the regression line is so poor that we would be just as well off not using it at all in predicting values of y. A correlation of +1 or -1, on the other hand, tells us that all points fall precisely on a straight line, and we can make extremely accurate predictions of y by employing the regression line. However, values of r which fall between

0 and 1 and 0 and -1, are somewhat more difficult to explain. Each time we take a sample of size n we get an r value, which will probably be different in each case. Thus we see that r could take on any value between -1 and 1, hence r is a random variable. If we view r as a random variable it would lead to the question, "What is the probability function of r?" . If we could determine the probability functions for r, we would then be in the position to make probability statements relating to r. We could formulate and test hypotheses and draw other useful conclusions concerning r. However, a thorough discussion of these topics, which can be found in several more advanced statistic books, such as one by Mood and Graybill, is beyond the scope of this treatise.

CHAPTER VI

SEQUENTIAL ANALYSIS

Thus far in our consideration of testing hypotheses, we started with a fixed sample size and from this sample we constructed estimates and formulated procedures for testing hypotheses based upon this fixed sample size. However, in general practice, it might be feasible to make certain statistical inferences based upon a sample smaller than the original intended sample. For example, suppose we are investigating a certain manufacturing process in which we have a criterion for determining whether a produced item will be accepted or rejected. Suppose further that at the outset we had decided to take a sample of size 100 and if we found 70 acceptable items we would continue the process, but if we found more than 30 defective items, we would stop the manufacturing process and investigate it. If the sampling process is extremely costly, we might try to minimize the sample size required to test the original hypothesis. If, after we sampled 50 items, we found almost all of these items were acceptable we might feel we had sufficient evidence to accept the original hypothesis and thus reduce the cost of the sampling procedure. Similarly, if after sampling 50 items, we observed a large part of these

to be defective, it seems feasible that we might be able to save on both sampling cost and loss of material by discontinuing the manufacturing process for investigation. This type of formulation of test of hypotheses is called a sequential test of statistical hypotheses. We shall devote a small amount of time to explaining the basic concepts concerning sequential analysis and illustrate by an example some theory which will be useful in testing hypotheses of this type. An essential feature of the sequential test, as distinguished from the current test procedure, is that the number of observations required by the sequential test depends on the outcome of the observations and is, therefore, not predetermined, but a random variable since for each experiment n may be different.

Formally, the sequential method of testing a hypothesis H may be described as follows: A rule is given for making one of the following three decisions at any state of the experiment (at the nth trial for each integer n):

(1) to accept the hypothesis H

(2) to reject hypothesis H

(3) to continue the experiment by making additional observations

Thus, such a test procedure is carried out sequentially. On the basis of the first observation, one of the aforementioned three decisions is made. If the first or second decision is made, the process is terminated. If the third decision is made, a third trial is performed, and so on.

The process is continued until either the first or the second decision is made. The number n of observations required by such a test procedure is a random variable, since the value of n depends on the outcome of the observations.

From the discussion above we see that sequential analysis is a method of statistical inferences whose characteristic feature is that the number of observations required by the procedure is not determined in advance of the experiment. The decision to terminate the experiment depends, at each stage, on the results of the observations previously made. A merit of the sequential method, as applied to testing statistical hypotheses, is that the test procedure can be constructed which requires, on the average, a substantially smaller number of observations than equally reliable test procedures based on a predetermined number of observations.

We shall now employ a method developed by Wald which will provide us with a procedure for testing a simple hypothesis. This procedure will employ techniques which are very similar to the likelihood ratio test discussed in an earlier chapter. After giving a formal definition for this test procedure, we will illustrate its usefulness by considering a simple example.

If we are given that, for a positive integer n, the probability that sample $X_1, X_2, .., X_n$ is obtained is given by p_{1n} when H_1 is true, i.e., p_{1n} is the likelihood

function when the alternative hypothesis is true, and by p_{0n} when H_0 is true, i.e., the likelihood function when H_0 is true, then the sequential probability ratio test for testing H_0 against H_1 is defined as follows:

Definition 6.1 The positive constants A and B($B<A$) are chosen. At each stage of the experiment (at the end of the nth trial for any integer n), the probability ratio $p_{1n}/p_{0n} = R_n$ is computed. Then one of these three decisions is made:

(1) If $B<R_n<A$, the experiment is continued by taking an additional observation (or set of observations)

(2) If $R_n \geq A$, the process if terminated with rejection of H_0 (acceptance of H_1)

(3) If $R_n \leq B$, the process is terminated with the acceptance of H_0

If, for a particular sample, $p_{1n} = p_{0n} = 0$, then R_n is defined as 1.

If, for some sample, $P_{1n} > 0$ but $P_{0n} = 0$, the inequality $R_n \geq A$ is considered fulfilled and H_0 is rejected.

One of the first questions which comes to mind is, "How do we determine these positive constants A and B?" . A complete discussion of the derivation of these constants can be found in [6]. This derivation is beyond the scope of our treatment and hence, we shall only state and use the results. The constants A and B are determined according to the desired values of a and b, where a is the

probability of making a type I error, i.e., a is the probability of rejecting H_0 when H_0 is actually true, and b is the probability of making a type II error, i.e., b is the probability of accepting H_0 when it should be rejected. Usually when we are engaged in an experiment, we have already determined in advance the values of a and b which will be used, hence, if we know how A and B depend upon a and b, we could determine A and B in advance. It can be shown, [6], that a and b are known functions of A and B, and a very simple but accurate approximation is given by the following:

$$A = \frac{1-b}{a}$$

$$B = \frac{b}{1-a} \quad .$$

This definition is very long and involved so let us now consider an example which will help us to recognize its usefulness. Suppose you are in the manufacturing business and you have a machine producing certain items. Suppose further, that you have decided upon a criterion by which you determine if a product is accepted or rejected. Let us define a random variable to help us determine a probability function which describes this process by letting x = 1 if the item is good, and x = 0 if the item is rejected. The probability function which gives the desired probabilities is given by $f(x:p) = p^x (1-p)^{1-x}$, $x = 0,1$

where p represents the true proportion of good items of any given number of items.

To apply the procedure mentioned at the outset of this chapter, we must determine A and B mentioned in definition 6.1. Suppose we agree that we can tolerate a type I error and type II error of a = .05 and b = .05. Thus we can calculate A and B which are given by

$$A = \frac{1-.05}{.05} \quad \text{or} \quad A = 19$$

and

$$B = \frac{.05}{1-.05} \quad \text{or} \quad B = .053 \ .$$

Suppose that the desired test of hypothesis is given as follows:

$$H_0: p = .7$$

$$H_1: p = .3 \ .$$

Suppose you take a sample of size one and observe that it is a 1, i.e., an acceptable item. We might be inclined to think that on the basis of this observation that our sequential probability ratio test would yield a value which would lead to the acceptance of H_0. Let us compute R_1 and see if this is actually the case. Now R_1 is given by

$$R_1 = \frac{(.3)^1 (1-3)^0}{(.7)^1 (1-.7)^0}$$

Hence

$$R_1 = 3/7 \ .$$

But $B < R_1 < A$, hence we must make additional observations. Suppose you take nine more observations which turn out to be given as follows: (including the first observation taken before)

$$(1,0,1,1,0,0,1,1,1,1)$$

Now R_{10} is given by the following expression:

$$R_{10} = \frac{(.3)^7 (1-.3)^3}{(.7)^7 (1-.7)^3} \ .$$

Simplifying, we get

$$R_{10} = \frac{(.3)^7 (.7)^3}{(.7)^7 (.3)^3}$$

and

$$R_{10} = (3/7)^4 = .012 \ .$$

Thus we see that $R_{10} < B$, hence we would accept $H_0: p=.7$.

From the above discussion, we see how useful this type of test of hypotheses can be in applied problems. Our aim in the treatment of the procedure here is to acquaint the reader with some of the basic techniques which can be implemented in testing hypotheses of this nature. A more advanced and complete development of the theory concerning this concept can be found in [6].

BIBLIOGRAPHY

(1) Cox, D.R. "Sequential Tests for Composite Hypotheses" Proceeding Cambridge Philosophical Society, XLVIII (1952), pp. 290-299.

(2) Freund, John E. Modern Elementary Statistics. Prentice-Hall, Inc., 1952, pp. 123-135, 159-171.

(3) Hoel, Paul G. Introduction to Mathematical Statistics. 3rd ed. John Wiley and Sons, Inc. 1962, pp. 160-168.

(4) Johnson, N.L. "Some Notes on the Application of Sequential Methods in the Analysis of Variance." Annals of Mathematical Statistics, XXIV (1953) pp. 614-623.

(5) Mood, Alexander M., and Graybill, Franklin A. Introduction to the Theory of Statistics. 2nd ed. McGraw-Hill Book Company, Inc., 1963.

(6) Wald, Abraham. Sequential Analysis, John Wiley & Sons, Inc., 1947, pp. 23-48.

(7) Commission on Mathematics College Entrance Examination Board. Introductory Probability and Statistical Inferences for Secondary Schools. 1957, pp. 71-103.

VITA

WAYNE F. HAYES

Candidate for the Degree of

Doctor of Education

Thesis: SPECIAL TOPICS IN ELEMENTARY MATHEMATICAL STATISTICS

Major Field: Higher Education Minor: Mathematics

Biographical:

Personal Data: Born near Vinson, Oklahoma, July 13, 1937, the son of Bert F. and Grace F. Hayes.

Education: Attended elementary and secondary school in Vinson, Oklahoma; graduated from Vinson High School in 1955; received the Bachelor of Science degree from Southwestern State College, Weatherford, Oklahoma, with a major in mathematics and a minor in physics, in January, 1960; received the Master of Science degree from Oklahoma State University, Stillwater, Oklahoma, with a major in mathematics, in May, 1965; completed requirements for the Doctor of Education degree at Oklahoma State University, in May, 1967.

Professional Experience: Two years of teaching at Eakly, Oklahoma; one and one-half years teaching at Amarillo, Texas. Two years as a graduate assistant at Oklahoma State University.

Professional Organizations: Mathematics Association of America.