AD-A016 496

RIA-80-U344

AD-A016 496

THE FOUNDATIONS OF PROBABILITY AND MATHEMATICAL STATISTICS

Gus W. Haggstrom

RAND Corporation
Santa Monica, California

March 1975

TECHNICAL LIBRARY

USADACS Technical Library
5 0712 01001016 2

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

KEEP UP TO DATE

Between the time you ordered this report—which is only one of the hundreds of thousands in the NTIS information collection available to you—and the time you are reading this message, several *new* reports relevant to your interests probably have entered the collection.

Subscribe to the **Weekly Government Abstracts** series that will bring you summaries of new reports *as soon as they are received by NTIS* from the originators of the research. The WGA's are an NTIS weekly newsletter service covering the most recent research findings in 25 areas of industrial, technological, and sociological interest—invaluable information for executives and professionals who must keep up to date.

The executive and professional information service provided by NTIS in the **Weekly Government Abstracts** newsletters will give you thorough and comprehensive coverage of government-conducted or sponsored research activities. And you'll get this important information within two weeks of the time it's released by originating agencies.

WGA newsletters are computer produced and electronically photocomposed to slash the time gap between the release of a report and its availability. You can learn about technical innovations immediately—and use them in the most meaningful and productive ways possible for your organization. Please request NTIS-PR-205/PCW for more information.

The weekly newsletter series will keep you current. But *learn what you have missed in the past* by ordering a computer **NTISearch** of all the research reports in your area of interest, dating as far back as 1964, if you wish. Please request NTIS-PR-186/PCN for more information.

WRITE: Managing Editor
5285 Port Royal Road
Springfield, VA 22161

Keep Up To Date With SRIM

SRIM (Selected Research in Microfiche) provides you with regular, automatic distribution of the complete texts of NTIS research reports *only* in the subject areas you select. SRIM covers almost all Government research reports by subject area and/or the originating Federal or local government agency. You may subscribe by any category or subcategory of our WGA **(Weekly Government Abstracts)** or **Government Reports Announcements and Index** categories, or to the reports issued by a particular agency such as the Department of Defense, Federal Energy Administration, or Environmental Protection Agency. Other options that will give you greater selectivity are available on request.

The cost of SRIM service is only 45¢ domestic (60¢ foreign) for each complete microfiched report. Your SRIM service begins as soon as your order is received and processed and you will receive biweekly shipments thereafter. If you wish, your service will be backdated to furnish you microfiche of reports issued earlier.

Because of contractual arrangements with several Special Technology Groups, not all NTIS reports are distributed in the SRIM program. You will receive a notice in your microfiche shipments identifying the exceptionally priced reports not available through SRIM.

A deposit account with NTIS is required before this service can be initiated. If you have specific questions concerning this service, please call (703) 451-1558, or write NTIS, attention SRIM Product Manager.

This information product distributed by

NTIS U.S. DEPARTMENT OF COMMERCE
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

AD-A011 049

THE FOUNDATIONS OF PROBABILITY AND MATHEMATICAL STATISTICS

By Gus W. Haggstrom

March 1975

PRICES SUBJECT TO CHANGE

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
US Department of Commerce
Springfield, VA. 22151

P-5394

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

PREFACE

These notes were written for an introductory course in probability and statistics at the post-calculus level that was presented during the fall term of 1974 to students in the Rand Graduate Institute. Most of the material is devoted to the basic concepts of probability theory that are prerequisite to learning mathematical statistics: probability models, random variables, expectation and variance, joint distributions, conditioning, correlation, and sampling theory. Among the distributions treated are the binomial, hypergeometric, Poisson, negative binomial, normal, gamma, lognormal, chi-square, and bivariate normal. The last section of the notes provides an introduction to some of the basic notions of parameter estimation: bias, efficiency, sufficiency, completeness, consistency, maximum likelihood, and least-squares estimation. Proofs of the Rao-Blackwell, Lehmann-Scheffé, and Gauss-Markov Theorems are included.

The author wishes to thank the following RGI students for their constructive comments on an earlier version of these notes, their assistance in eliminating many (but surely not all) of the errors, and their patience and goodwill: Joe Bolten, Tom Carhart, Chris Conover, Wendy Cooper, Roger DeBard, Steve Glaseman, Masaaki Komai, Ragnhild Mowill, Captain Michael A. Parmentier, and Hadi Soesastro.

Preceding page blank

Preceding page blank

TABLE OF CONTENTS

SECTION I. - INTRODUCTION

Statistics is the branch of applied mathematics that is concerned with techniques for (1) collecting, describing, and interpreting data; and (2) making decisions and drawing inferences based upon experimental evidence. The term "statistics" is also used to refer to the data themselves or numbers calculated from the data, as in the expression "lies, damned lies, and statistics." Sometimes it is not clear which usage is intended, as in the old saw, "You can prove anything with statistics." At any rate, statistical terminology, measures, and analytical techniques have become commonplace in the scientific community for describing and interpreting experimental results, and a knowledge of statistics has become a prerequisite for scientific research in many fields.

As a branch of applied mathematics, statistics relies heavily on mathematical models. The solution to a statistics problem typically involves four steps:

(1) Statement of the real problem.

(2) Specification of a mathematical model to fit the problem.

(3) Solution of the mathematical problem.

(4) Application to the real problem.

Even if the real problem is completely specified in a particular application, the choice of the mathematical model and therefore the solution may still be practically unlimited. Obviously, the mathematical model should contain the essential features of the physical situation, but in most cases this will not lead to a unique specification of the model, and it will be meaningless to refer to a "correct" choice. The final choice of the model will be affected by the intuition and subject matter knowledge of the model builder and perhaps by his ability to carry out the mathematical solution. For now, let us assume the choice has been made.

The next step, solving the mathematical problem, will often be straightforward, since the model will probably be chosen using ease of solution as a criterion. The final step, identifying the solution of the mathematical model with the answer to the real world problem, would appear immediate, but this is often the step where the experimenter discovers that his presumably well-conceived mathematical model yields a solution that cannot possibly satisfy the real problem.

Since the mathematical models for statistical applications are primarily probability models and since statistical theory depends heavily on probability theory, we shall begin our study of statistics with a consideration of those probability concepts that will be needed in the sequel. But, before we proceed along that path, it may be helpful to provide a single example of a statistical problem to introduce some terminology and to indicate the applicability of the models that will be treated.

Consider the problem of estimating the proportion of some population who share a common attribute based upon a sample of a certain size from that population. For example, the population might consist of the voters in a certain state, and the problem might be to estimate the proportion of the voters who favor a given candidate based upon the stated preferences of a relatively small number of voters. As a second example, consider estimating the proportion of defective transistors produced by a given machine based upon a sample of transistors chosen from that machine's output. Here, the population of interest is not a group of people, but the set of transistors produced by the machine.

As these examples illustrate, the problem under consideration is a common one. So as not to confuse the issues involved, let us pretend that the population of interest is a big can of marbles that contains an unknown proportion p of red ones and that the sample

will consist of drawing 10 marbles one by one "with replacement" from the can. A sample is said to be drawn with (or without) replacement if, after each draw, the marble is (or is not) returned to the can. In either case, the sample is said to be a random sample if on each draw every marble in the can has the same chance of being selected. Your problem: estimate (guess) the value of p based upon a random sample of size 10 taken with replacement.

As a first step toward specifying a mathematical model to fit this situation, note that the data of the experiment is conveniently represented by a vector $x = (x_1, x_2, \ldots, x_{10})$ where x_i is 1 or 0 according as the i^{th} marble drawn is red or not. Thus, if the first two marbles drawn are red and the others are all white, then $x = (1,1,0,0,0,0,0,0,0,0)$. This is an example of a sample point, i.e., a point that summarizes the data for a particular realization of an experiment. The set of all possible sample points x is called the sample space for the experiment. Your estimate $\hat{p}$ can be taken as any value computed from the vector x. Three possibilities that you might consider are $\hat{p}_1 = \bar{x} = \Sigma_{i=1}^{10} x_i/10$ or perhaps $\hat{p}_2 = (1 + 8\bar{x})/10$ or even $\hat{p}_3 = 1/2$, which ignores the data and guesses that p is 1/2 no matter what the data indicates.

Note that the values of $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$ are prescribed by the formulas above for all sample points x. These are examples of statistics, i.e., numbers calculated from the data points. These particular statistics are also called estimators of the parameter p to differentiate them from other statistics in this example, such as Σx_i, $x_1 - x_2$, $\max(x_1, x_2)$, and $7x_{10} + 52$. The values of the estimators at a particular sample point are called estimates. Thus, for the sample point $(1,1,0,0,0,0,0,0,0,0)$, the three estimates of p are $\hat{p}_1 = 1/5$, $\hat{p}_2 = 0.26$, and $\hat{p}_3 = 1/2$. Of course, if the actual proportion of red marbles in the can is $p = 1/2$, then $\hat{p}_3$ provides the best estimate of p. However our intuition tells us that for values of p near 0 or 1 the estimators $\hat{p}_1$ and $\hat{p}_2$ will usually provide more reliable estimates.

As this example indicates, estimates themselves have little intrinsic interest, because one can always specify an estimator that will yield any value whatsoever. In some applications of this model, measures of goodness can be prescribed for comparing estimators, in which case the problem of choosing an estimator reduces to solving the mathematical problem of determining the one that is best in the sense of these criteria. However, such instances are rare. In most applications, clear-cut goodness criteria for estimators do not exist, and one is content to report the value of the "usual" estimator of p, namely, $\hat{p}_1 = \bar{x}$. As will be seen later, this estimator has many desirable properties and contains all the information about p that is provided by the sample.

A further discussion of this problem is deferred until the elementary probability concepts required for this and other statistical problems are treated. For a nontechnical discussion of the nature of statistics, its uses and misuses, see W. Allen Wallis and Harry V. Roberts, Statistics, A New Approach, Free Press, Glencoe, Illinois, 1956, Chapters 1-3. For a pleasant diversion that is somewhat related to the subject, see Darrell Huff, How to Lie with Statistics, W. W. Norton and Co., New York, 1954.

SECTION II. - PROBABILITY MODELS

References:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1970, Chapters 1 and 2.

Seymour Lipschutz, Theory and Problems of Probability, Schaum's Outline Series, McGraw Hill, New York, 1968, Chapters 1 and 3.

Emanuel Parzen, Modern Probability Theory and Its Applications, Wiley, 1960, Chapter 1.

William Feller, An Introduction to Probability Theory and its Applications, Vol. I, 3rd Edition, Wiley, 1968, Chapter 1.

Paul E. Pfeiffer, Concepts of Probability Theory, McGraw-Hill, New York, 1965, pp. 1-40.

Certain physical experiments have the property that their outcomes are somewhat unpredictable and appear to "depend on chance." As examples, consider flipping a coin, throwing dice, picking three students by lot from a class, spinning a roulette wheel, finding the lifetime of a light bulb, and determining the time between successive telephone calls coming into an exchange. If we rule out the uninteresting cases for the moment (e.g., two-headed coins or dice controlled electronically so that "7" must appear), each of these experiments has the property that the outcome of the experiment cannot be predicted with certainty. Yet, when the experiment is repeated many times, a certain regularity may appear. For example, if a slightly bent coin is tossed many times, the relative frequency of heads, computed after each toss and based upon all the outcomes up to that toss, may seem to fluctuate less and less around a particular number, say 2/3. Similarly, the successive averages of the times between incoming telephone calls during a certain part of the day may appear to "tend" to a certain number.

These experiments also suggest questions about the "chance" or "likelihood" or "probability" of certain outcomes or collections of outcomes occurring: "If two dice are tossed, what is the probability of getting a total of seven or more?" "If telephone calls come into an exchange at an average rate of 4 per minute, what is the probability of getting more than 10 in any one minute during the next hour?" "What is the probability of drawing a straight flush in poker?"

Before tackling a formal definition of probability, we shall first put the idea of a random experiment into a mathematical framework. To the outcomes of interest of the experiment, we make correspond the elements of a set S called a sample space. That is, a sample space of an experiment is a set S such that each element of S corresponds to one of the outcomes of the experiment. For example, if the experiment consists of tossing a coin, we might take as our sample space the set $S = \{H, T, E\}$, i.e., S is the set consisting of the three letters H, T, and E, where "H" stands for "heads," "T" for "tails," and "E" for "edge." As this example shows, the choice of S is somewhat arbitrary.

As a second example consider the experiment of throwing two dice. For convenience let us assume that the dice are painted red and green to distinguish them. Then we can designate the outcome that 3 turns up on the red die and 4 turns up on the green die by the pair (3,4). Using similar designations for the other possible outcomes, we see that an appropriate sample space S for this experiment is the set of pairs (x,y) where x and y are integers from 1 to 6. This sample space can be visualized by plotting the pairs as indicated in the figure below. We can write S in set notation by listing all the elements of S as follows:

$$S = \{(1,1), (1,2), (1,3), \ldots, (6,6)\}.$$

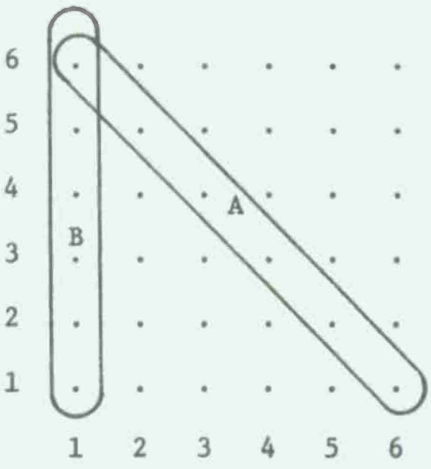
Alternatively, we can write

$$S = \{(x,y) : x \text{ and } y \text{ are integers from } 1 \text{ to } 6\},$$

which can be read as "S is the set of pairs (x,y) such that x and y are integers from 1 to 6."

The elements of a sample space S are sometimes called sample points (or just points), and an event is a collection of sample points, i.e., a subset of S. (For the moment, any subset of S will be referred to as an event; later, for technical reasons, the term "event" will be reserved

for subsets of S in a certain class.) For example, the pair (5,2) is a sample point in the sample space S above; this can be written as $(5,2) \in S$, where the symbol "$\in$" stands for "is an element of" or "belongs to." In the game of craps it is of interest to consider the event A corresponding to a total of seven on both dice. (See the figure.) In set notation this event could be written as:



$$A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$$

or $$A = \{(x,y) : x + y = 7\}.$$

The event B designated in the figure corresponds to the result that the red die turns up 1:

$$B = \{(x,y) : x = 1\}.$$

In general, an event A is said to <u>occur</u> if the outcome of the experiment corresponds to a sample point s in S such that $s \in A$. Thus, if A and B are the events defined above and if the result of tossing the dice is 5 on the red die and 2 on the green, then A occurs but B does not occur. If A and B are events such that A is a subset of B, written $A \subset B$ or $B \supset A$, then clearly whenever A occurs, B must also occur.

It will be convenient to have notation for the <u>union</u> and <u>intersection</u> of any two events A and B. As the words suggest, the union of A and B, denoted by $A \cup B$, is the set of all those points that belong to at least one of the sets A and B, whereas the intersection of A and B, denoted by $A \cap B$, consists of those points which belong to both A and B. Thus, in the example above,

$$A \cup B = \{(x,y) : x = 1 \text{ or } x + y = 7\}$$

$$A \cap B = \{(x,y) : x = 1 \text{ and } x + y = 7\} = \{(1,6)\}.$$

Note that the event $A \cup B$ occurs if either A or B occurs (or both), whereas $A \cap B$ occurs if and only if both A and B occur. Also note that the notions of union and intersection can be extended to more than two events. For example, if A, B, and C are events, then $A \cap B \cap C$ is the set of points common to all three sets. Also, if A_1, A_2, ...

is a sequence of events, then $\bigcup_{i=1}^{\infty} A_i$ (or $A_1 \cup A_2 \cup \ldots$) is the set of all points that belong to at least one of the sets A_i, and $\bigcap_{i=1}^{\infty} A_i$ is the set of points that belong to all the sets A_i.

If two events A and B have no points in common, we say that the events are disjoint (or mutually exclusive). Introducing the symbol $\emptyset$ to denote the "empty set" (i.e., the set having no elements), we can write this as $A \cap B = \emptyset$. For example, in the dice throwing sample space above, if

$A = \{(x,y) : x + y = 7\}$ and

$B = \{(1,1), (1,2), (2,1), (6,6)\}$, then $A \cap B = \emptyset$.

The complement of an event A, denoted by A^c, is the event consisting of those points in S that do not belong to A. Symbolically, $A^c = \{s : s \notin A\}$; here, "$\notin$" stands for "does not belong to." Note that $A \cap A^c = \emptyset$ and $A \cup A^c = S$.

Example. Let S be the Cartesian plane, i.e., $S = \{(x,y) : x$ and y are real numbers$\}$. Then the "curve" $y = x^2$ is the set $A = \{(x,y) : y = x^2\}$. The set $B = \{(x,y) : x^2 + y^2 < 1\}$ is the set of points inside the circle of radius 1 centered at the origin. If $C = \{(x,y) : x^2 + y^2 = -1\}$, then $C = \emptyset$. To "solve" the set of equations $x + y = 5$ and $3x - y = 3$ means to find the intersection of the sets $D = \{(x,y) : x + y = 5\}$ and $E = \{(x,y) : 3x - y = 3\}$, namely, $D \cap E = \{(2,3)\}$. The set $F = \{(x,y) : 3x - y < 3\}$ is the set of points above the line $y = 3x - 3$; F^c is the set of points on or below this line. Note that $F^c \cap B = \emptyset$.

We shall want to talk about the probability of any event A, denoted by $P(A)$. As this notation suggests, P will be defined as a function of events. To begin with, let us assume that the sample space is finite, say $S = \{s_1, s_2, \ldots, s_n\}$. Then a finite probability model is prescribed by assigning numbers p_i to the sample points s_i such that

(a) each p_i is nonnegative, and

(b) $\Sigma_{i=1}^{n} p_i = 1$.

In this case, the probability $P(A)$ of any event A is the sum of the p_i's assigned to the points that belong to A.

For example, consider the coin-tossing example where the sample space chosen was $S = \{H,T,E\}$. In this case, there are only 8 events, namely,

$\emptyset$, {H}, {T}, {E}, {H,T}, {H,E}, {T,E}, S.

If the coin is fairly fat and bent a little, an appropriate assignment of probabilities p_i to the points H, T, and E might be 1/2, 1/3, and 1/6, in which case the probabilities of the events are

$$P(\emptyset) = 0 \qquad P(\{H,T\}) = 5/6$$
$$P(\{H\}) = 1/2 \qquad P(\{H,E\}) = 2/3$$
$$P(\{T\}) = 1/3 \qquad P(\{T,E\}) = 1/2$$
$$P(\{E\}) = 1/6 \qquad P(S) = 1$$

Although we might want to choose another P to fit a particular coin, this choice of P is at least consistent with some of our intuitive notions about probability, namely:

I. $0 \le P(A) \le 1$ for all events A.

II. $P(\emptyset) = 0$, $P(S) = 1$.

IIIa. If A and B are events such that $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

Similarly, if S is countably infinite, say $S = \{s_1, s_2, \ldots\}$, one can assign probabilities to all subsets of S in a consistent way by first assigning probabilities p_i to the points s_i where $p_i \ge 0$ and $\Sigma p_i = 1$. Then, for any event A, P(A) is defined by

$$P(A) = \sum_{s_i \in A} p_i.$$

It is easily checked that P satisfies conditions I, II, and IIIa above as well as:

III. If $A_1, A_2, \ldots$ are events such that $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

In general, a set function P on the class of events of a sample space S, countable or not, is said to be a <u>probability measure</u> if P satisfies conditions I-III above. (Condition IIIa follows from III by setting $A_3 = A_4 = \ldots = \emptyset$ in III.) Its value P(A) for any event A is then called the <u>probability</u> of A. To sum up the discussion above, if the sample space S is countable (in which case it is said to be <u>discrete</u>), P can be prescribed by assigning nonnegative values p_i

that sum to unity to the individual sample points s_i, in which case the probability of any event is the sum of the probabilities assigned to the points that belong to that event.

Using the properties I-III above, one can easily show that for any probability measure P and any events A and B,

(1) $P(A^c) = 1 - P(A)$

(2) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(3) $P(A \cup B) \leq P(A) + P(B)$

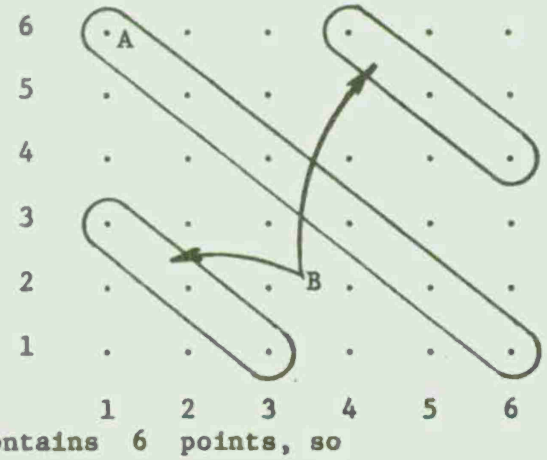
(4) If $B \subset A$, $P(B) \leq P(A)$.

For the present we shall assume that P is given or that there is a "natural" choice of P suggested by the problem. Whether the probabilities P(A) actually fit the physical situation in some sense or how they are measured in practice does not enter the picture at this stage. This is analogous to the situation in trigonometry when one is given the lengths of the sides of a triangle and is asked to determine the area.

The case where the physical experiment, when properly viewed, has N outcomes which appear to be "equally likely" can be handled immediately in this framework, at least theoretically. The key words in such problems are "chosen at random," "fair coin," "honest dice," "selected by lot," etc. For such situations, one can choose an appropriate sample space S with N points and assign probability 1/N to each point. Then, for any event A, P(A) = (number of elements in A)/N.

Example. (Dice throwing) If two dice are thrown, find the probability of getting a total of (a) seven, (b) four or ten.

Solution. The problem remains unchanged if we consider the dice distinguishable, say red and green. Let $S = \{(x,y) : x, y \text{ are integers from 1 to 6}\}$. For example, the sample point (3,4) corresponds to 3 on the red die and 4 on the green. Assuming equally likely outcomes (honest dice), we assign probability 1/36 to each point.



(a) The event A, "seven occurs," contains 6 points, so

$P(A) = 6/36 = 1/6.$

(b) The event B, "four or ten occurs," also contains 6 points, so

$P(B) = 1/6.$

Example. (Coin-tossing) If a fair coin is tossed four times (or if four fair coins are tossed), what is the probability of getting at least two heads?

Solution. An appropriate sample space for a single toss is $S = \{H,T\}$. For four repetitions of the experiment, we can let $S^4 = S \times S \times S \times S = \{(x_1,x_2,x_3,x_4) : x_i \in S\}$.[1]
The sample point (T,T,T,H), for example, corresponds to obtaining tails on the first three tosses and heads on the fourth. There are $2^4 = 16$ points in S^4, and we assign probability 1/16 to each point. The complement A^c of the event A, "at least two heads," contains five points: (T,T,T,T), (H,T,T,T), (T,H,T,T), (T,T,H,T), (T,T,T,H). Therefore,

$$P(A) = 1 - P(A^c) = 1 - (5/16) = 11/16.$$

Exercise. An absent-minded hatcheck girl has 4 hats belonging to 4 men. Since she cannot remember which hat belongs to each man, she returns them at random. Find the probability that

(a) exactly two men get their own hats back. Ans. 1/4.

(b) at least two men get their own hats back. Ans. 7/24.

(Set up an appropriate sample space and show the correspondence between the sample points and the outcomes of the experiment.)

As another example of an experiment that fits the equally likely outcomes case, consider the experiment of choosing a sample of size r at random without replacement from some population of n objects, say $\Pi = \{a_1,a_2,\ldots,a_n\}$ where $n \geq r$. For purposes of illustration, let $r = 3$, and suppose that the experiment is conducted by first choosing one of the elements a_i in Π in such a way that each element has the same chance of being chosen. Then a second element is chosen at random from those remaining. Finally, a third element is chosen at random from those remaining after the first and second have been chosen. If the elements a_5, a_7, a_1 are chosen in that order, this outcome can be represented by the 3-tuple (a_5,a_7,a_1). Similarly, the result of choosing a sample of size r can be represented by an r-tuple $(x_1,x_2,\ldots,x_r)$ where the components x_i are

---

[1] This notation uses an obvious generalization of the notation for the Cartesian product $C \times D$ of two sets C and D as defined by:

$$C \times D = \{(c,d) : c \in C, d \in D\}.$$

Thus, $C \times D$ is the set of all ordered pairs having the property that the first component belongs to C and the second component belongs to D.

different elements of the population. This r-tuple is an example of a permutation, i.e., an arrangement of r symbols from a set of size n in which repetitions are not allowed.

The number of permutations of n symbols taken r at a time, denoted by $P(n,r)$, can be determined as follows. The first component of the r-tuple can be filled by any of the n symbols, the second by any of the $n-1$ symbols not already used in filling the first component,..., the r^{th} by any of the $n-(r-1)$ symbols not already used in filling the first $r-1$ components. The total number of different ways of filling the r components is

$$P(n,r) = n(n-1)(n-2)\cdots(n-r+1) = n!/(n-r)! \quad \text{for } r = 1,2,\ldots,n$$

where $n! = n(n-1)(n-2)\cdots(3)(2)(1)$ and $0! = 1$. The reason for setting $0! = 1$ is to have the formula $P(n,r) = n!/(n-r)!$ hold for $r = n$, in which case $P(n,r) = n!$.

If the elements in the sample are drawn simultaneously so that the order in which the elements are drawn is unknown, the outcomes of the experiment can be represented using combinations (subsets) of size r instead of r-tuples. For example, if $r = 3$, the subset $\{a_1,a_5,a_7\}$ corresponds to drawing the elements a_1, a_5, and a_7 in some order. Note that for each subset of size three, say $\{a_1,a_5,a_7\}$, there are $3! = 6$ permutations, namely,

$$(a_1,a_5,a_7),\ (a_1,a_7,a_5),\ (a_5,a_1,a_7),\ (a_5,a_7,a_1),\ (a_7,a_1,a_5),\ (a_7,a_5,a_1).$$

Hence, the number of subsets of size three is the number of permutations of size three divided by $3!$. In general, if $\binom{n}{r}$ denotes the number of different subsets of size r from a set of size n, then it follows by an argument similar to that above for the case $r = 3$ that

$$\binom{n}{r} = \frac{P(n,r)}{r!} = \frac{n!}{r!(n-r)!} \quad \text{for } r = 0,1,\ldots,n.$$

Theorem 2-1. Given any set of size n, say $\Pi = \{a_1,\ldots,a_n\}$, the number of ordered r-tuples (permutations) $(x_1,\ldots,x_r)$ such that the x_i's are different elements of Π is

$$P(n,r) = n(n-1)\ldots(n-r+1) = n!/(n-r)! \quad \text{for } r = 1,2,\ldots,n.$$

The number of subsets (combinations) of size r from Π is

$$\binom{n}{r} = n!/r!(n-r)! \quad \text{for } r = 0, 1,\ldots, n.$$

The following example illustrates how the above results are used in sampling inspection.

Example. A box contains 12 items of which 9 are defective. What is the probability that a random sample of size 4 taken without replacement will contain exactly 3 defectives?

Let the set of 12 items be denoted by $\Pi = \{D_1,\ldots,D_9,G_1,G_2,G_3\}$. Two solutions will be given below, the first using subsets of Π of size 4 as sample points and the second using permutations of size 4 as sample points. Although the sample spaces are quite different, the solutions to the problem yield the same answer.

Solution A. Set $S = \{x : x \text{ is a subset of size 4 from } \Pi\}$. The number of points in S is

$$\#S = \binom{12}{4} = \frac{12!}{4!8!} = \frac{12\cdot11\cdot10\cdot9}{4\cdot3\cdot2\cdot1} = 495.$$

Assign probability $1/495$ to each point. Let $A = \{x\in S:$ x contains 3 D's and 1 G$\}$. Since $\#A$ = (no. of ways of choosing 3 of 9 D's) $\times$ (no. of ways of choosing 1 of 3 G's) $= \binom{9}{3}\binom{3}{1} = 252$,

$$P(A) = \binom{9}{3}\binom{3}{1}/\binom{12}{4} = 252/495 = 28/55.$$

Solution B. Set $S = \{(x_1,x_2,x_3,x_4): x_i\in\Pi,\ x_i\neq x_j \text{ for } i\neq j\}$. Then $\#S = P(12,4) = 12\cdot11\cdot10\cdot9$. Let $A = \{x\in S:$ exactly three x_i's are D's$\}$. Then

$\#A$ = (no. of ways of choosing 3 of 9 D's) $\times$ (no. of ways of choosing 1 of 3 G's) $\times$ (no. of ways of ordering the four chosen symbols) $= \binom{9}{3}\binom{3}{2}4!$,

so that $P(A) = \binom{9}{3}\binom{3}{1}4!/P(12,4) = \binom{9}{3}\binom{3}{1}/\binom{12}{4} = 28/55$.

The above argument is easily generalized to prove the following theorem:

Theorem 2-2. A random sample of size n is taken without replacement from a lot of N items of which the proportion p are defective. The probability $p(x)$ that the sample will contain exactly x defectives is

$$p(x) = \binom{Np}{x}\binom{Nq}{n-x}/\binom{N}{n} \quad \text{for } x = 0,1,2,\ldots,n$$

where $q = 1-p$.

Now suppose that the sample is taken with replacement. Then an appropriate sample space for the experiment is

$$S = \{(s_1, s_2, \ldots, s_n): \ s_i \in \Pi\}.$$

Since each component of the sample points can be filled in N ways and repetitions are permitted, the number of points in S is N^n. Let A be the event that exactly x of the items drawn are defective. The number of points in A is the number of ways of choosing x of the n components to be filled by D's [namely, $\binom{n}{x}$] multiplied by the number of ways of filling the x chosen components with D's [namely, $(Np)^x$] multiplied by the number of ways of filling the remaining $n-x$ components with G's [namely, $(Nq)^{n-x}$ where $q = 1-p$]. Therefore, the number of points in A is

$$\#A = \binom{n}{x}(Np)^x(Nq)^{n-x},$$

and

$$P(A) = \#A/N^n = \binom{n}{x}p^x q^{n-x} \qquad \text{for } x = 0,1,2,\ldots,n.$$

This proves the following result:

Theorem 2-3. If a random sample of size n is taken with replacement from a lot of N items of which the proportion p are defective, then the probability $p(x)$ that the sample will contain exactly x defectives is

$$p(x) = \binom{n}{x}p^x q^{n-x} \qquad \text{for } x = 0,1,2,\ldots,n$$

where $q = 1-p$.

As an example of an experiment that requires an infinite sample space, imagine a person tossing a fair coin until a head occurs. The previous example suggests using the sample space

$$S_1 = \{(H), (T,H), (T,T,H), \ldots, (T,T,\ldots)\}$$

where $(T,T,\ldots)$ corresponds to never obtaining heads. A slightly simpler sample space $S = \{1,2,3,\ldots,\infty\}$ is obtained by considering the so-called "waiting time" for heads, i.e., the number of the trial on which heads first occurs. Since the coin is assumed fair, we let $P\{1\} = 1/2$. Analogy

with the previous example, where we set $P\{(T,T,T,H)\} = 1/16 = 1/2^4$, prompts us to set $P\{4\} = 1/2^4$. Similar considerations for any n leads us to set $P\{n\} = 1/2^n$ for every n. Then, since

$$\sum_{n=1}^{\infty} P\{n\} = \sum_{n=1}^{\infty} 1/2^n = 1,$$

we must have $P\{\infty\} = 0$, which is consistent with our intuitive notion that, if the coin is really fair, it cannot come up tails infinitely many times.

Having assigned probabilities to the elementary events, we can compute the probability of any event. For example, the probability that at least 4 tosses are needed is

$$1 - P\{1,2,3\} = 1 - (1/2 + 1/4 + 1/8) = 1/8.$$

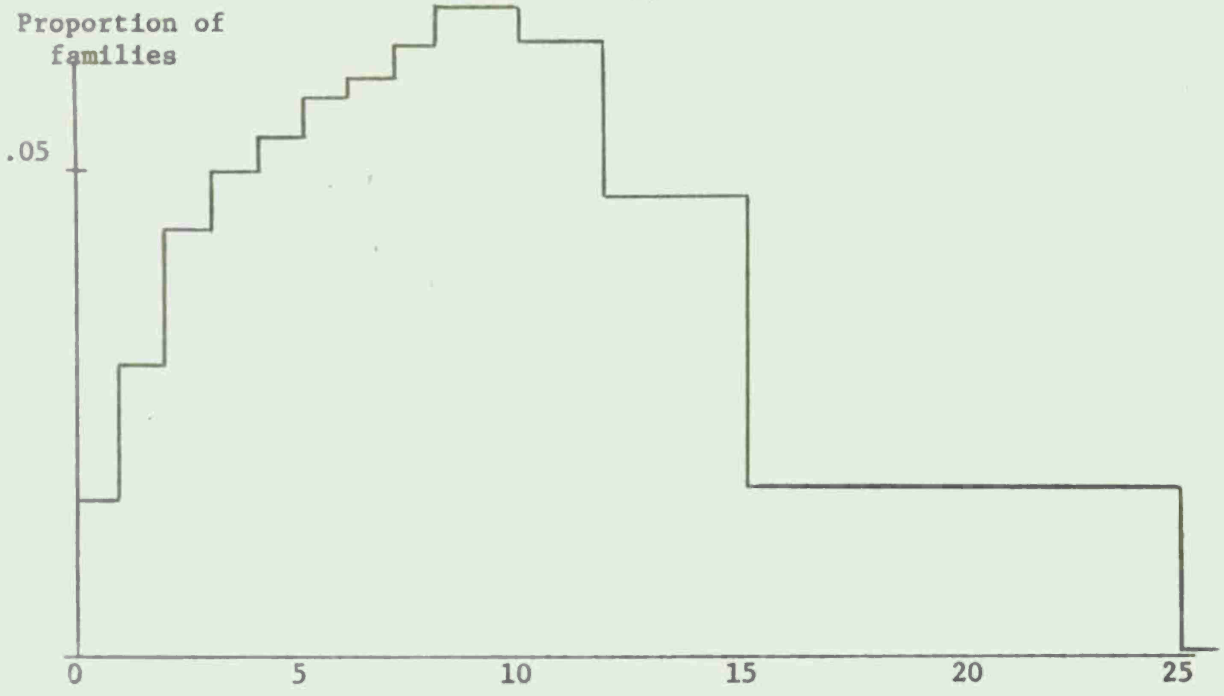
Also, the probability that the waiting time is odd is

$$P\{s : s \text{ is odd}\} = \sum_{k=1}^{\infty} 2^{-2k+1} = \frac{1/2}{1 - (1/4)} = 2/3$$

Example. According to the U.S. Bureau of the Census (Current Population Reports, Series P-60, No. 78, May 20, 1971), the "distribution" of family income in 1970 in the United States was as follows:

Family Income	Percent of Families	Family Income	Percent of Families
Under $1000	1.6	$7000-7999	6.3
$1000-1999	3.0	$8000-9999	13.6
$2000-2999	4.3	$10000-11999	12.7
$3000-3999	5.0	$12000-14999	14.1
$4000-4999	5.3	$15000-24999	17.7
$5000-5999	5.8	$25000-49999	4.1
$6000-6999	6.0	$50000 up	0.5

This distribution can be represented graphically using a histogram as indicated in the figure below. Note that the heights of the rectangles above the income intervals have been chosen in such a way that the areas of the rectangles are proportional to the percentages given in the table.



Although the reason for doing so will not be apparent at this time, one can build a probability model around the distribution above by considering the experiment of choosing a family "at random" from the population of all families and recording, as the outcome of the experiment, the family income of the family selected. As a sample space for this experiment, we can take the set of nonnegative real numbers: $S = [0,\infty)$. Guided by the table above, we can choose our class of events to be the sets $\emptyset$, [0, 1000), [1000, 2000), ..., and unions of these intervals. To be consistent with the table above, we let our probability measure P have values:

$$P([0,1000)) = .016, \quad P([1000,2000)) = .030, \quad \text{etc.}$$

If a family is chosen at random from the population, the event A corresponding to selecting one having income less than \$3000 is the event

$$A = [0,1000) \cup [1000,2000) \cup [2000,3000),$$

and the probability of this event is

$$P(A) = .016 + .030 + .043 = .089,$$

which is the proportion of families in the population having income less than \$3000 according to the Bureau of Census estimates.

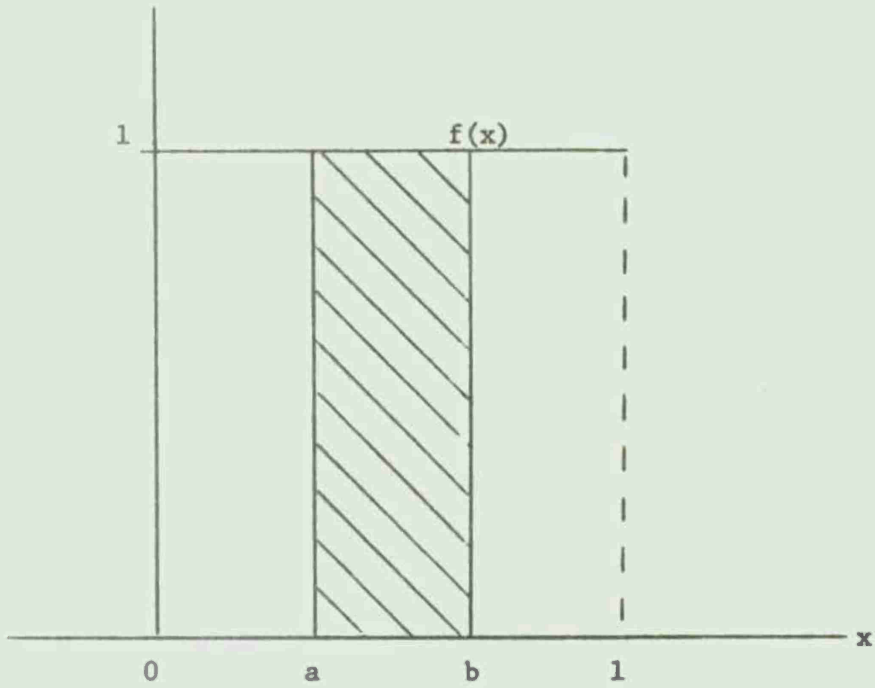
Note that our class of events did not include every subset of S in this case. Our class of events was restricted to those subsets of S whose probabilities were determined either directly from the table or by application of the axioms for a probability measure. The next example indicates another reason for considering classes of events that do not include all subsets of the sample space.

Example. (Spinning a spinner) Imagine trying to choose a real number between 0 and 1 "at random." A hypothetical physical model for this would be to spin a perfectly balanced spinner on a circle with uniform markings from 0 to 1. Here, an obvious choice for the sample space is $S = [0,1]$, which is uncountable. In order for the numbers to be "equally likely," each singleton set must have probability zero in this case, so that the scheme used to assign probabilities in the discrete case breaks down. However, we clearly want to have, for example, $P[.3,.4] = .1$ and $P(.25,.39] = .14$, which leads us to assign probability to any interval (a,b) [or $(a,b]$ or $[a,b)$ or $[a,b]$] its "length" $b - a$. Following condition III for a probability measure, probability can also be assigned to any set which is a countable union of disjoint intervals, and this value again coincides with out notion of the "length" of the set.

Is there a consistent way of defining "length" for every subset of $[0,1]$? Unfortunately, the answer is no. (Reference: H. L. Royden, Real Analysis, Macmillan, New York, 1963, p. 43.) One way out of this difficulty is to restrict the class of events, i.e., the class of subsets of $[0,1]$ for which probability is assigned.

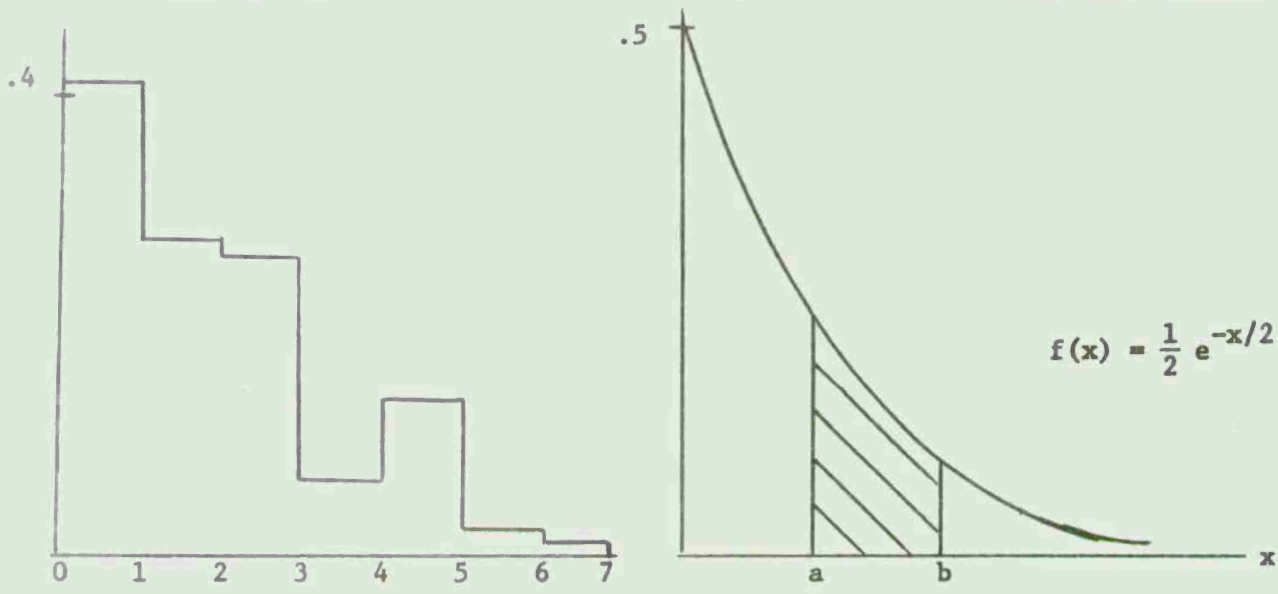
One such restriction is to the smallest class of subsets which contains the intervals and is closed under countable unions, countable intersections, and complementation. For our purposes it suffices to know that such a class exists and that there is a way of defining a probability measure on this class which corresponds to our intuitive notion of length.

Note that in this example the probability of any interval $[a,b]$ with $0 \le a < b \le 1$ can be visualized as the area under the "curve" $f(x) = 1$ for $0 \le x \le 1$ and between the ordinates $x = a$ and $x = b$, as illustrated in the figure below.



The next example shows how other curves can be used to prescribe probability measures on the line.

Example. Consider the waiting time in minutes between telephone calls coming into an exchange. A histogram based upon the observed waiting times for 100 calls coming into the exchange during a certain period of the day may look like the figure on the left below. The figure is intended to depict a case where 42 out of 100 waiting times were less than one minute.



Let $S = (0,\infty)$. Theory to be developed later in this course suggests that, if the average waiting time between calls is 2 minutes, then a reasonably well-fitting model might be obtained by assigning probabilities to intervals [a,b] using areas under the curve $f(x) = (1/2)e^{-x/2}$ as is illustrated in the figure on the right above. That is,

$$P([a,b]) = \int_a^b (1/2)e^{-x/2}\,dx = e^{-a/2} - e^{-b/2}.$$

The theory will also suggest that, under certain assumptions about the waiting times between calls, a histogram based upon thousands of waiting times (using a finer partition of the x-axis than is indicated in the figure above) should fit the curve on the right quite well. Also, the relative frequency of the observed waiting times falling in a particular interval [a,b] should be close to the preassigned probability $P([a,b])$.

As in the spinner example, the probability of any countable union of disjoint subintervals of S can be computed by adding the probabilities of the individual intervals. As before, technical difficulties preclude assigning probabilities to all subsets of S, but we can again restrict ourselves to the smallest class of events that contains the intervals and is closed under countable set operations (unions, intersections, and complements).[1] It can be shown that any probability measure on this class of sets is completely determined by its values on the intervals. Thus the function f above completely specifies the assignment of probabilities to this class of sets through the relationship

$$P([a,b]) = \int_a^b f(x)\,dx.$$

The function f is an example of a <u>density function</u>, i.e., a nonnegative function whose integral over the real line is equal to one. Clearly, any density function can be used to specify a probability measure on the line, and it is often convenient in applications of probability to use density functions in specifying probability measures (or "distributions") on the line.

---

[1] The smallest class of subsets of the line that contains the intervals and is closed under countable set operations is often referred to as the class of <u>Borel sets</u> of the line.

SECTION III. - CONDITIONAL PROBABILITY AND INDEPENDENCE

References:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1970, Chapter 3.

Seymour Lipschutz, Theory and Problems of Probability, Schaum's Outline Series, McGraw-Hill, New York, 1968, Chapter 4.

Emanuel Parzen, Modern Probability Theory and Its Applications, Wiley, 1960, Chapters 2 and 3.

William Feller, An Introduction to Probability Theory and its Applications, Vol. I, 3rd Edition, Wiley, 1968, Chapter 5.

Paul E. Pfeiffer, Concepts of Probability Theory, McGraw-Hill, New York, 1965, pp. 41-105.

Consider choosing a person at random from a population of N voters of whom N_F are female and N_C are planning to vote for Charles Charmer. Let C be the event that the person plans to vote for Charmer and F the event that the person is female. Then

$$P(C) = \frac{N_C}{N} \quad \text{and} \quad P(F) = \frac{N_F}{N} .$$

Now suppose that we are informed that the person chosen was a woman. This eliminates many sample points as possible outcomes of the experiment, and it may not be the case that the proportion of women favoring Charmer is the same as the corresponding proportion $P(C)$ for the entire population. If in fact N_{CF} women plan to vote for Charmer, then our revised assessment of the probability that the person chosen will vote for Charmer is N_{CF}/N_F. This ratio is called the conditional probability of C given F and is denoted by $P(C|F)$. If it happens that $P(C) = P(C|F)$, so that knowing

that the event F occurred does not change our assessment of the probability of C, then the events C and F are said to be independent. These concepts are defined for arbitrary sample spaces below.

Conditional Probability

For any two events A and B such that $P(B) > 0$, the conditional probability of A given B is defined by

$$P(A|B) = P(A \cap B)/P(B).$$

Note that, for fixed B, the conditional probability $P(A|B)$ is proportional to $P(A \cap B)$ with the constant of proportionality chosen to make $P(B|B) = 1$.

In a finite probability model $S = \{s_1, s_2, \ldots, s_n\}$ with equally likely points, the probability of any event C is $\#(C)/n$ where $\#(C)$ denotes the number of points in C. Therefore

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\#(A \cap B)/n}{\#(B)/n} = \frac{\#(A \cap B)}{\#(B)},$$

so that in this case $P(A|B)$ is the proportion of the points in B that also belong to A. In general, $P(A|B)$ is the proportion of the probability assigned to B that also belongs to A.

It follows immediately from the definition of $P(B|A)$ that

$$P(A \cap B) = P(A)\, P(B|A).$$

More generally, if $A_1, A_2, \ldots, A_k$ are any events for which $P(A_1 \cap A_2 \cap \ldots \cap A_{k-1}) > 0$, then

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap \ldots \cap A_{k-1}).$$

These results are sometimes useful in computing probabilities of joint occurrences of events when it is obvious what the conditional probabilities must be by reference to the reduced sample spaces.

Exercises. 1. Two fair dice are thrown, one red and one green. What is the conditional probability that the sum is ten or more given that (a) an observer reported that the red die turned up as a five? (b) a colorblind observer has reported that one of the dice turned up a five (not intending to exclude the possibility that both turned up fives)? Ans. (a) 1/3, (b) 3/11.

2. Consider drawing two balls at random without replacement from an urn containing six numbered balls where balls 1 to 4 are white and 5 and 6 are red. Let A be the event that the first ball drawn is white and B the event that the second ball drawn is white. Is it not obvious from the physical situation that $P(B|A) = 3/5$? Is it equally obvious that $P(A|B) = 3/5$? Do you believe that $P(A) = P(B)$? Set up a sample space for this experiment with equally likely outcomes and verify your answers.

3. A batch of 10 light bulbs contains three defectives. Bulbs are selected at random without replacement and tested one by one. Find the probability that the second defective occurs on the sixth draw. Ans. 1/6. Hint: Let A be the event that there is exactly one defective in the first five draws and B the event that there is a defective on the sixth draw. Evaluate $P(A \cap B)$ using conditional probabilities.

4. Let Q be the set function defined on a class of events by $Q(A) = P(A|B)$ where P is a probability measure and B is an event for which $P(B) > 0$. Show that Q is a probability measure, thus verifying that conditional probabilities "act like" probabilities.

Bayes' Theorem

A partition of a sample space is a set of disjoint events $B_1, B_2, \ldots, B_k$ such that their union is the entire sample space S. For example, any event B and its complement B^c constitute a partition. If the sample space corresponds to some population, then any stratification of that population, say by race, income level, or sex, constitutes a partition of S.

The following result, the second part of which is called Bayes' Theorem, is easily proved.

Theorem 3-1. Let $B_1, B_2, \ldots, B_k$ be a partition of S such that $P(B_i) > 0$ for each i. Then for any event A

(i) $$P(A) = \sum_j P(A \cap B_j) = \sum_j P(A|B_j)P(B_j)$$

(ii) if $P(A) > 0$,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}.$$

Example. Suppose 20% of the people in a certain group are bad drivers. Of these, 40% drive sports cars. Of the good drivers, 5% drive sports cars. If you pick a person at random and he drives a sports car, what is the probability that he is a bad driver?

Let V, B, and G denote the events corresponding to sports car drivers, bad drivers, and good drivers in a sample space S that corresponds to the population of interest. Then

$$P(V) = P(V|B)P(B) + P(V|G)P(G)$$
$$= (.4)(.2) + (.05)(.8) = .12.$$

Thus, $$P(B|V) = \frac{P(V|B)P(B)}{P(V)} = \frac{(.4)(.2)}{.12} = \frac{2}{3}.$$

Exercises. 1. Prove the theorem above.

2. A plant produces three grades of components: 20% of all components produced are of grade A, 30% of grade B, and 50% of grade C. The percentage of defective components in the three grades are 5, 4, and 2 percent respectively. (a) What proportion of all components produced in the plant are defective? (b) If a component selected at random from the plant's output is defective, what is the probability that it is of grade A? Ans. (a) 0.032, (b) 5/16.

3. A certain disease is present in about one out of 1000 persons in a certain population. A test for the disease exists which gives a "positive" reading for 95% of the victims of the disease, but it also gives positive readings for 1% of those who do not have the disease. What proportion of the persons who have positive readings actually have the disease? Ans. 0.087.

Independent Events, Independent Experiments, and Bernoulli Trials

Two events A and B are said to be independent if

$$P(A \cap B) = P(A)P(B).$$

If $P(B) > 0$, this condition is clearly equivalent to having $P(A|B) = P(A)$. Thus, A and B are independent if and only if knowing that B has occurred does not change the probability that A will occur. In the equally likely outcome case, two events A and B are independent if the proportion of the points in B that also belong to A is the same as the proportion of points in the entire sample space that belong to A.

Three or more events $A_1, A_2, \ldots, A_n$ are said to be <u>independent</u> if for any subsequence of k integers $i_1 < i_2 < \ldots < i_k$ from 1 to n

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

In particular, three events A, B, and C are independent if the following four conditions hold:

$$P(A \cap B) = P(A)P(B)$$
$$P(A \cap C) = P(A)P(C)$$
$$P(B \cap C) = P(B)P(C)$$
$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

<u>Example</u>. Referring back to the probability model for throwing two fair dice, one can readily check that any two of the three events A = "3 on the green die," B = "4 on the red die," and C = "total of seven" are (pairwise) independent. However, it is not the case that $P(A \cap B \cap C) = P(A)P(B)P(C)$, because $P(A \cap B \cap C) = 1/36$ whereas $P(A)P(B)P(C) = (1/6)^3 = 1/216$. Hence, these three events are not independent.

The probability model for tossing two fair dice is an instance of a model for two "independent experiments." Let $S_1 = \{s_1, s_2, \ldots\}$ and $S_2 = \{t_1, t_2, \ldots\}$ be discrete sample spaces for two experiments, and let P_1 and P_2 be the corresponding probability measures for the separate experiments.[1] Then a sample space for the combined experiment is

---

[1] In the dice-throwing example, both S_1 and S_2 consist of the integers from 1 to 6, and both probability measures P_i assign probability 1/6 to each point in S_i.

$$S = S_1 \times S_2 = \{(s,t) : s \in S_1,\ t \in S_2\}.$$

The two experiments are said to be <u>independent</u> if probabilities are assigned to the points of S using the formula:

$$P\{(s,t)\} = P_1\{s\}P_2\{t\}.$$

To see the connection between independent experiments and independent events, let A be any event in the combined sample space S that depends on the outcome of the first experiment only (e.g., "3 or more on the red die"). Then A is of the form $C \times S_2 = \{(s,t) : s \in C\}$ where C is an event in S_1 (e.g., $C = \{3,4,5,6\}$). Similarly, let $B = S_1 \times D$ be any event that depends on the outcome of the second experiment only (e.g., "2 on the green die"). Then it is easily verified that

$$P(A \cap B) = P(C \times D) = P_1(C)\ P_2(D) = P(A)P(B).$$

Thus, if probabilities are defined multiplicatively on S using the rule indicated above, any event that depends on the outcome of the first experiment only is independent of any event that depends on the outcome of the second experiment only.

To extend the notion of independent experiments to more general sample spaces, one is led by the discussion above for discrete sample spaces to proceed as follows. Let S_1 and S_2 be any two sample spaces with probability measures P_1 and P_2. If C is any event in S_1 and D is any event in S_2, define the probability of the "rectangle" $C \times D$ in the product space $S = S_1 \times S_2$ by

$$P(C \times D) = P_1(C)P_2(D).$$

It follows from this definition that any event $A = C \times S_2$ that depends on the outcome of the first experiment is independent of any event $B = S_1 \times D$ that depends on the result of the second experiment only, since

$$P(A \cap B) = P(C \times D) = P_1(C)P_2(D) = [P_1(C)P_2(S_2)][P_1(S_1)P_2(D)] = P(A)P(B).$$

More generally, one can combine the sample spaces $S_1, S_2, \ldots, S_n$ for n separate experiments and define probabilities multiplicatively on the product space $S = S_1 \times S_2 \times \ldots \times S_n$ to provide a model for n <u>independent</u>

experiments. It will then follow that, if $A_1, A_2, \ldots, A_n$ are events such that A_i depends on the result of the ith experiment only, these events are independent.

For example, consider n trials of exactly the same type (e.g., repeated tosses of a coin, or successive draws at random with replacement from a population) where each trial results in one of two outcomes of interest, say 1 and 0 (for success or failure, or heads and tails, or employed and unemployed), with probabilities p and $q = 1 - p$ on each trial. Such trials are called Bernoulli (or binomial) trials.

A probability model for n Bernoulli trials is prescribed by taking the sample space $S = \{(x_1, \ldots, x_n) : x_i = 1 \text{ or } 0\}$ and assigning probabilities, for example, as follows:

$$P\{(1,1,0,1,\ldots,0)\} = ppqp\cdots q.$$

To see how to compute probabilities of certain events of interest, consider the event A_3 that exactly three of the n trials result in successes. Then A_3 consists of all sample points in S that have exactly three 1's. Since the probability assigned to any such point is p^3q^{n-3}, it follows that $P(A_3) = \#(A_3)p^3q^{n-3}$ where $\#(A_3)$ is the number of points in A_3. But the number of points in A_3 is clearly the number of ways of choosing three of the n components for the 1's. That is,

$$\#(A_3) = \binom{n}{3} = \frac{n!}{3!(n-3)!}.$$

In particular, if $n = 4$, the number of points in A_3 is $4!/3!1! = 4$, namely, (1,1,1,0), (1,1,0,1), (1,0,1,1), and (0,1,1,1).

Similarly, if A_k is the event that there are exactly k successes in n Bernoulli trials, then

$$P(A_k) = \binom{n}{k} p^k q^{n-k} \quad \text{for } k = 0,1,\ldots,n.$$

For example, the probability of n successes is p^n, the probability of n failures is q^n, and the probability of at least one success is $1 - q^n$.

Exercises. 1. Find the probability that, if four fair coins are tossed, (a) all will turn up heads, (b) three will turn up heads. Ans. (a) 1/16, (b) 1/4.

2. Balls are drawn at random with replacement from an urn containing 1/3 red balls and the rest white. Find the probability that (a) five successive draws will yield two red balls, then three white balls, (b) there are exactly two red balls in the five draws, (c) there are at least two red balls in five draws. Ans. (a) 8/243, (b) 80/243, (c) 131/243.

3. If only 25% of the voters favor a certain candidate, what is the probability that a random sample of size 10 will show 8 or more favoring him? Ans. $436/4^{10} = 0.0004$.

SECTION IV. RANDOM VARIABLES AND THEIR DISTRIBUTIONS

References:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1970, Chapter 4.

Seymour Lipschutz, Theory and Problems of Probability, Schaum's Outline Series, McGraw-Hill, New York, 1968, Chapter 5.

Paul E. Pfeiffer, Concepts of Probability Theory, McGraw-Hill, New York, 1965, Chapter 3.

Consider the dice-throwing example again, where the sample space chosen was $S = \{(x,y) : x,y \in \{1,2,\ldots,6\}\}$. In the game of "craps," one is not interested in the particular outcome (x,y) that occurs, because only the sum is relevant. This leads us to consider the "random variable" Z on S defined for all points (x,y) by $Z(x,y) = x + y$. In general, a random variable is a real-valued function defined on a sample space.[1] Roughly speaking, the key idea behind the notion of a random variable is that it is a variable that depends on the result of a random experiment; its value for a particular outcome of an experiment is a number computed from the data point.

---

[1] This definition suffices for discrete sample spaces, where all subsets of S are events, and for the applications of probability models to be considered in this course. For arbitrary sample spaces, in which not all subsets are events, probabilists prefer to define a random variable X as a real-valued function on S such that the subset $\{s : X(s) \leq c\}$ is an event for every real number c. The purpose of this additional restriction is to assure that, under certain reasonable assumptions on the class of events, probabilities of the form $P(X \leq c)$, $P(X < c)$, and $P(a < X < b)$ are all defined for any random variable X, as well as any probabilities of the form $P(X \in B)$ where B is a countable union of intervals (open, half-open, or closed) on the line. For our purposes, we can consign this bit of pedantry to a footnote and refer the mathematically oriented reader to books on probability theory, e.g., the book by Pfeiffer cited above.

Some other random variables on the same sample space are:

$X(x,y) = x$

$Y(x,y) = y$

$$V(x,y) = \begin{cases} 1 & \text{if } x + y = 7 \text{ or } 11 \\ 0 & \text{otherwise.} \end{cases}$$

Note that we have used capital letters X, Y, and Z to denote random variables rather than the usual function notation of calculus (e.g., f, g, h). This usage has become traditional in probability and statistics to distinguish the random variables from their values, which in turn are often denoted in lower-case letters.

Sometimes random variables are defined implicitly as functions of other random variables. For example, Z could have been defined above using usual function notation as $Z = X + Y$.

Ordinarily random variables are defined verbally rather than explicitly using function notation. Thus, one might refer to the number of successes X in n Bernoulli trials. Relative to the sample space S at the end of the previous section, this means that for any sample point $s = (x_1, x_2, \ldots, x_n)$ consisting of 1's and 0's, X(s) = (number of 1's in s). Note that if X_i denotes the result of the ith trial (i.e., $X_i(s) = x_i$), then $X = \Sigma_{i=1}^{n} X_i$. This illustrates how a random variable can sometimes be represented as a function of other random variables of a simpler nature. Here, each X_i has only two possible values 0 and 1. The utility of such representations will be exhibited later.

The following examples of random variables refer to problems discussed in Section II.

1. Hatcheck girl problem.

Let S be the set of the 4! permutations of the integers 1,2,3,4, namely, (1,2,3,4), (2,1,3,4), etc. The point (2,4,3,1), for example corresponds to the outcome that the first man receives the second man's hat, the second man receives the fourth man's hat, the third man receives his own hat, and the fourth man receives the first man's hat. Let X be the random variable corresponding to the number of hats returned correctly, so that X(2,4,3,1) = 1, X(1,2,3,4) = 4, etc.

2. Spinner problem.

The sample space chosen to correspond to the set of possible readings of the spinner was the unit interval $[0,1]$.

(a) Let X be the number chosen at random:

$$X(s) = s \quad \text{for all} \quad s.$$

(b) $Y(s) = \sin 2\pi s$. [No one said that random variables had to be of particular interest for the experiment under consideration. This one happens to be of interest in another context, that of choosing a direction at random, specified by a point $(\cos 2\pi s, \sin 2\pi s)$ on the unit circle.]

(c) $X^2(s) = s^2$. [Note the strange, but unambiguous, notation.]

(d) $Z(s) = \begin{cases} 1 & \text{if} \quad 0 \le s \le 1/4, \\ 0 & \text{if} \quad s > 1/4. \end{cases}$

3. Telephone problem.

(a) Let X be the waiting time in minutes until a telephone call comes into the exchange, i.e., $X(s) = s$ for all $s > 0$.

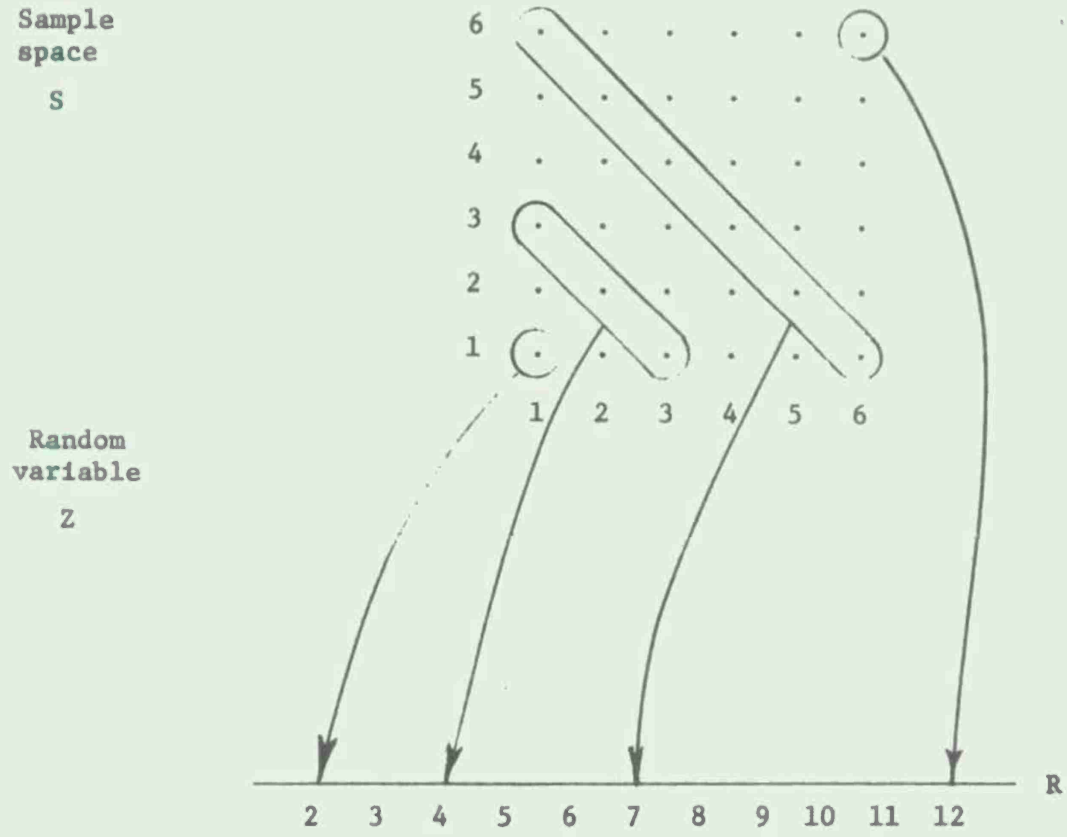
(b) $Y = X/60$, the corresponding waiting time in hours.

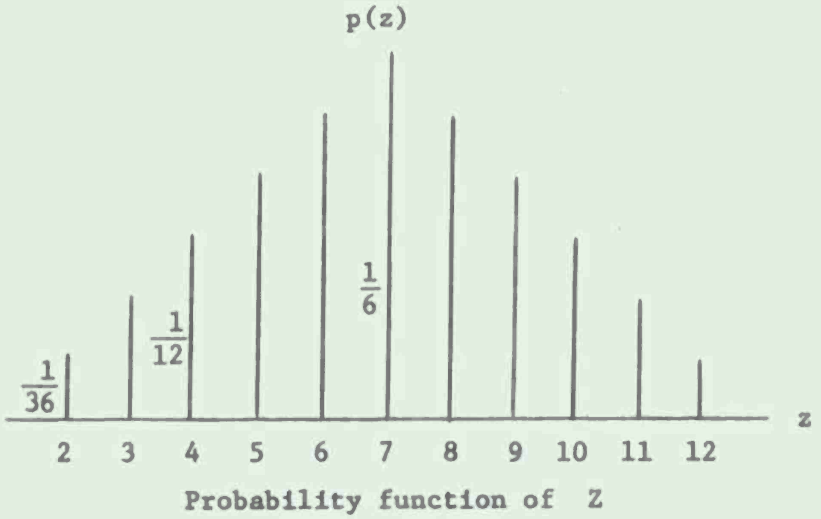
(c) Z = integral part of X. For example, if $X(s) = 6.875$ (minutes), then $Z = 6$.

Just as a random variable X "maps" (or "carries") sample points from S into the real number line, it also carries probabilities on S into the real line R, inducing a probability measure on R that is called the distribution of the random variable X. As we shall see, distributions of random variables play a central role in statistical theory.

To get a feeling for the notion of a distribution of a random variable, let us return once again to the dice-throwing example and consider the sum of the outcomes on the two dice, $Z(x,y) = x + y$. The figure on the next page attempts to depict the way that the random variable Z maps points in S into R and thereby induces a probability distribution on R. The top part of the figure indicates the correspondence between events in S and the possible values of Z: $2,3, \ldots, 12$. Since Z has value 4 on the event $\{(3,1), (2,2), (1,3)\}$, and this event has probability $P(Z = 4) = 3/36 = 1/12$, the number 4 receives probability $1/12$ under the distribution induced by Z. The function depicted in the bottom half of the figure indicates the probabilities assigned to the other

values of Z. This is a graph of the "probability function" of Z, one method of characterizing the distribution of a "discrete" random variable.





Probability function of Z

Figure IV - 1

In general, a random variable X is said to be <u>discrete</u> if there is a countable set of real numbers, say $A = \{x_1, x_2, \ldots\}$, such that $P(X \in A) = 1$. In this case, the function p on A defined by

$$p(x) = P(X = x)$$

is called the <u>probability function</u> of X. Some obvious properties of the probability function are:

(a) $p(x) \geq 0$ for all x in A,

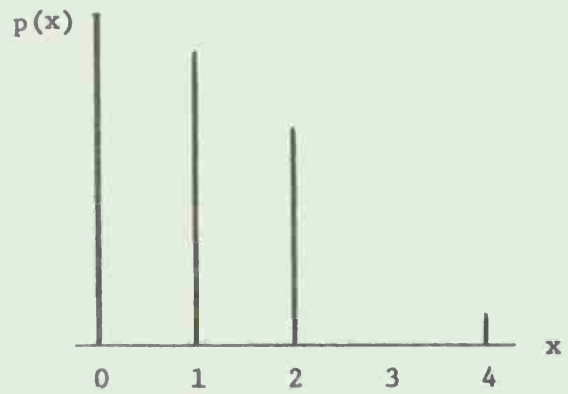
(b) $\sum_{x \in A} p(x) = 1$.

The probability function of the random variable Z in the dice-throwing example was depicted at the bottom of Figure IV-1. As a second example, let X be the number of hats returned correctly in the hatcheck girl problem. As was seen in an exercise in Section II,

$$p(2) = P(X = 2) = 1/4.$$

Other values of the probability function of X are given below.

x	$p(x)$
0	3/8
1	1/3
2	1/4
3	0
4	1/24



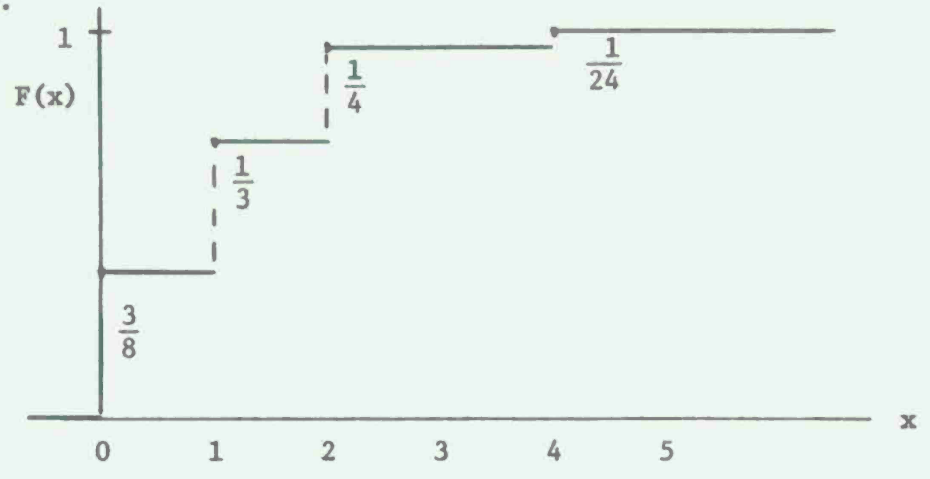
Clearly, any random variable on a sample space that has only countably many points must be discrete. As an example of a discrete random variable with infinitely many values, consider the waiting time for heads in repeated independent tosses of a fair coin. Examples of discrete random variables on uncountable sample spaces are given by Examples 2(d) and 3(c) above. The other examples of random variables for the spinner and telephone problems are not discrete, and since $P(X = x) = 0$ for all values of x for both the random variable X in the spinner problem and the waiting time in the telephone problem, we shall require characterizations other than the probability function to specify their distributions.

The distribution function (or cumulative distribution function) of a random variable X is defined for all real x by

$$F(x) = P(X \leq x).$$

The importance of the distribution function of X is that it provides a simple characterization and description of the distribution of X, whether X is discrete or not.

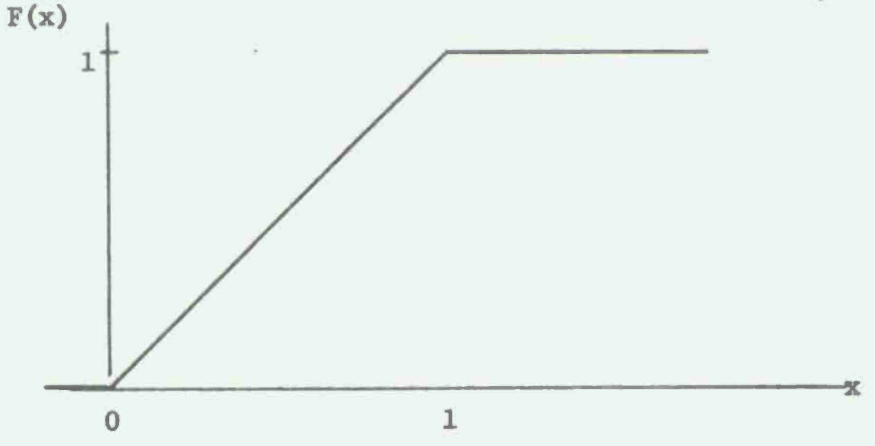
Examples. 1. Let X be the random variable in the hatcheck girl problem. The graph of the distribution function F of X is given below.



Comparing this graph with that of the probability function above, we note that the distribution function has jumps at 0, 1, 2, and 4, the values which X takes on with positive probabilities.

2. If X is the random variable in the spinner problem, then

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ x & \text{if} \quad 0 \leq x \leq 1 \\ 1 & \text{if} \quad x > 1. \end{cases}$$

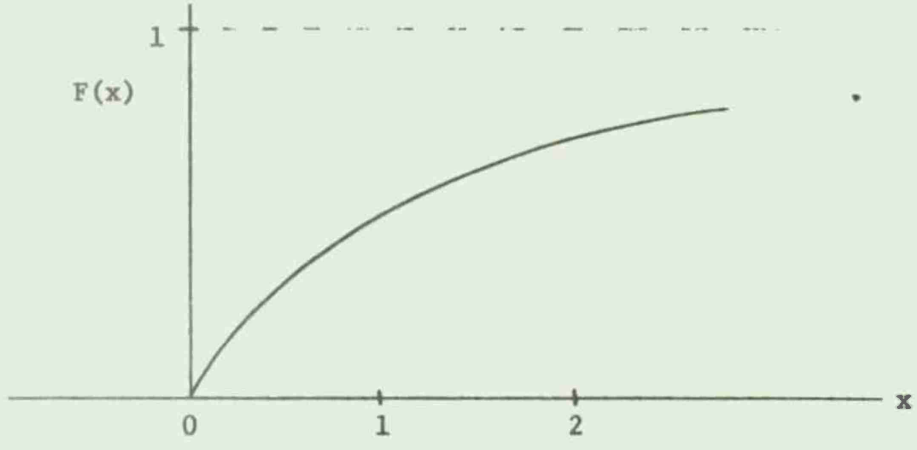


3. Let X be the random variable in the telephone problem. Then

$$F(x)=\int_{-\infty}^{x} f(t)\,dt \quad \text{where} \quad f(t)=\begin{cases} 0 & \text{if} \quad t < 0 \\ 1/2\, e^{-t/2} & \text{if} \quad t \geq 0 \end{cases}$$

so that

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 - e^{-x/2} & \text{if} \quad x \geq 0. \end{cases}$$



Although the three distribution functions above are quite different in nature, they share a number of common properties. In general, the distribution function F of a random variable X must satisfy the following properties:

(a) $0 \leq F(x) \leq 1$ for all real numbers x.

(b) F is monotonically increasing, i.e., if $a < b$, then $F(a) \leq F(b)$.

(c) $F(-\infty) = 0$, $F(\infty) = 1$.

(d) F is right continuous, i.e., $F(x + 0) = F(x)$ for all x [here, $F(x + 0)$ denotes $\lim F(y)$ as y tends to x from above].

(e) $P(a < X \leq b) = F(b) - F(a)$.

(f) $P(X = b) = F(b) - F(b-0)$ [this is the jump in F at b].

If X is discrete and has probability function p, then

$$F(x) = \sum_{x_i \leq x} p(x_i).$$

In most instances, the probability function is preferable to the distribution function in describing a particular discrete distribution. We now turn to

another class of distributions for which a characterization other than the distribution function is usually preferable.

A random variable X with distribution function F is said to have a continuous distribution if there is a nonnegative function f on R (called the density function of R) such that

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

for each real value of x.

Examples 2 and 3 above provide examples of random variables having continuous distributions. The density function of the random variable X in Example 2 is given by

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The density function in Example 3 is clearly specified. Since

$$P(a < X \le b) = F(b) - F(a) = \int_{a}^{b} f(x)\, dx,$$

these probabilities can be visualized as areas under the curve $f(x)$ and between the ordinates $x = a$ and $x = b$, as was illustrated earlier in Section II.

It follows from (1) above that, if X has a continuous distribution, then its distribution function F is continuous. However, the converse of this statement is not true since there are continuous distribution functions F for which no density function f exists. (An attempt to depict such a function F is given on page 193 in Introduction to Measure and Integration by M. E. Munroe.) Therefore some writers prefer to say that X has an absolutely continuous distribution when (1) holds.

Some observations which follow from (1) are:

(a) $\frac{dF(x)}{dx} = f(x)$ at every continuity point x of f:

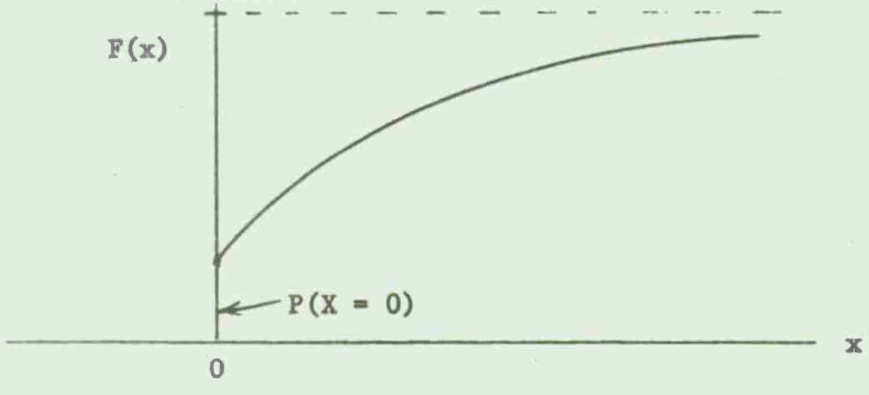
(b) $\int_{-\infty}^{\infty} f(x)\, dx = 1$;

(c) if X has a continuous distribution, then

$$P(X = x) = F(x) - F(x-0) = 0$$

for every real value of x.

To indicate an application which gives rise to a random variable which has a distribution which is neither continuous nor discrete, consider measuring the lifetime of a lightbulb, where it is reasonable to assume that there is a nonzero probability that the bulb will not burn at all. A distribution function like the one pictured below might be appropriate in this situation.



Exercises. 1. Five balls are chosen at random from an urn containing 9 balls of which 3 are white. Let X be the number of white balls in the sample. Find and sketch the probability function of X if the balls are chosen (a) with replacement, (b) without replacement.

Ans. (a) 32/243, 80/243, 80/243, 40/243, 10/243, 1/243.

(b) 1/21, 5/16, 10/21, 5/42, 0, 0.

2. Suppose Y has a density function of the form

$$f(y) = cy \quad \text{for } 0 < y < 1.$$

(a) What is the value of c?

(b) Find $P(Y < 1/2)$.

(c) Find and sketch the distribution function of Y.

(d) Find and sketch the density function of $U = 3Y$. [Note that $P(U \le u) = P(Y \le u/3)$.]

(e) Find and sketch the density function of $V = Y + 1$.

Ans. (a) 2, (b) 1/4, (c) $F(y) = 0$ for $y \le 0$, y^2 for $0 < y < 1$, 1 for $y \ge 1$.

(d) $2u/9$ for $0 < u < 3$, (e) $2(v-1)$ for $1 < v < 2$.

3. Let $Y = \sqrt{X}$ where X is the random variable in the spinner problem.

(a) Find $P(Y < 1/2)$.

(b) Find $P(Y < 1/2 | X < 3/4)$.

(c) Find and sketch the distribution function of Y.

(d) Find and sketch the density function of Y.

Ans. (a) 1/4, (b) 1/3, (c) same as 2(c), (d) $2y$ for $0 < y < 1$.

4. Let X be the random variable in the telephone problem. Show that $P(X > a+b | X > a) = P(X > b)$ for all positive values of a and b.

SECTION V. - CHARACTERISTICS OF DISTRIBUTIONS

References:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1970, Chapter 7.

Seymour Lipschutz, Theory and Problems of Probability, Schaum's Outline Series, McGraw-Hill, New York, 1968, Chapter 5.

Paul E. Pfeiffer, Concepts of Probability Theory, McGraw-Hill, New York, 1965, Chapter 5.

Consider the experiment of drawing a tag at random from a box containing N tags of which $1/2$ are marked "1," $1/3$ are marked "2," and $1/6$ are marked "3." Let X be the number on the tag that is drawn. With an appropriate sample space for this experiment consisting of N equally likely outcomes, X is a random variable having probability function

$$\begin{array}{c|ccc} x & 1 & 2 & 3 \\ \hline p(x) & 1/2 & 1/3 & 1/6 \end{array}.$$

The "expected value" of X, denoted by $E(X)$, will be defined below as a weighted average of the possible values of X using the probabilities $p(x)$ as weights. In this case,

$$E(X) = 1(1/2) + 2(1/3) + 3(1/6) = 5/3.$$

Before proceeding with a formal definition, we note two interpretations of $E(X)$ in this example. First, the arithmetic average (mean) of all the numbers on the N tags in the box is

$$\frac{1(N/2)+ 2(N/3) + 3(N/6)}{N} = 1(1/2) + 2(1/3) + 3(1/6) = 5/3.$$

Thus, in this case $E(X) = 5/3$ coincides with the ordinary average of the tag numbers in the box. Next, suppose we repeat the experiment independently

a large number of times, say n, and let n_1, n_2, n_3 be the number of times that tags numbered 1, 2, and 3 are drawn. Then the average of the numbers drawn on the n trials is

$$\frac{1n_1 + 2n_2 + 3n_3}{n} = 1(n_1/n) + 2(n_2/n) + 3(n_3/n).$$

In a large number of trials, we would anticipate that the sample proportions n_1/n, n_2/n, and n_3/n would be close to the probabilities 1/2, 1/3, and 1/6. Therefore, we would expect that the average of the numbers drawn would be close to $E(X) = 5/3$. The validity of this second interpretation of $E(X)$ will be established later.

Definition. Let X be a discrete random variable having possible values x_1, x_2, ... and probability function p. Then the expected value (expectation, mean) of X is defined by

$$E(X) = \Sigma_k \, x_k \, p(x_k)$$

provided that $\Sigma \, x_k \, p(x_k)$ converges absolutely. If $\Sigma |x_k| p(x_k)$ diverges we say that the expected value of X does not exist (or that the expectation of X is infinite).

Examples.

1. A random variable X is said to have a Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = q = 1 - p$. In this case,

$$E(X) = 1 \cdot p + 0 \cdot q = p.$$

2. Suppose X has probability function $p(x_i) = 1/n$ where $x_1, x_2, \ldots, x_n$ are n (distinct) real numbers. Then $E(X) = \Sigma \, x_i/n$.

3. Let X be the waiting time for a "1" if a fair die is tossed repeatedly until "1" occurs for the first time. Then X has probability

function $p(x) = q^{x-1}p$ where $p = 1/6$ and $q = 5/6$. Therefore, $E(X) = \sum_{x=1}^{\infty} xq^{x-1}p = 1/p = 6$. [In general, $\sum_{k=0}^{\infty} z^k = 1/(1 - z)$ for $|z| < 1$; taking derivatives on both sides in this equation yields $\sum_{k=1}^{\infty} kz^{k-1} = 1/(1 - z)^2$ for $|z| < 1$.]

4. An example of a discrete random variable that does not have an expectation is provided by letting X be a random variable such that $P(X = 2^n) = 1/2^n$ for $n = 1, 2, \ldots$ In this case, each term in the series $\Sigma x_k\, p(x_k)$ is equal to one, and hence the series does not converge.

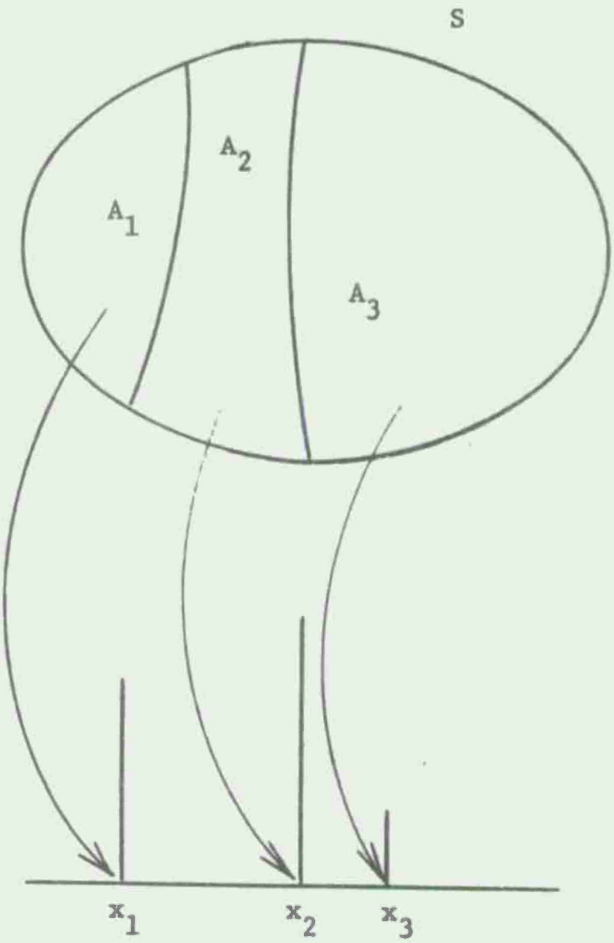
Exercises.

1. Let X be the number of heads that occur in three tosses of a fair coin. Show that $E(X) = 3/2$.

2. Five balls are chosen at random from an urn containing 9 balls of which 3 are white. Let X be the number of white balls in the sample. Show that $E(X) = 5/3$ whether the sampling is done with or without replacement. [You derived the probability function(s) of X in Exercise 1, page 36.]

3. If two fair dice are tossed and Z is the sum of the results, show that $E(Z) = 7$. (See page 31 for the probability function of Z.) Now suppose that the two dice are colored red and green. Let X be the result on the red die, and Y the result on the green die. Show that $E(X) = E(Y) = 7/2$, thus verifying the $E(X + Y) = E(X) + E(Y)$ in this case.

To derive some of the fundamental properties of expectation, let us first restrict our attention to discrete sample spaces $S = \{s_1, s_2, \ldots\}$, so that the random variables involved will necessarily be discrete.



For purposes of illustration, let X be a random variable on S having only three possible values x_1, x_2, and x_3, and consider the partition of the sample space into the sets $A_i = \{s : X(s) = x_i\}$. Denoting the elements of A_i by s_{i1}, s_{i2}, ..., we have that

$$\begin{aligned} E(X) &= \Sigma x_k \, p(x_k) \\ &= x_1 P(A_1) + x_2 P(A_2) + x_3 P(A_3) \\ &= x_1(P\{s_{11}\} + P\{s_{12}\} + \dots \quad) \\ &\quad + x_2(P\{s_{21}\} + P\{s_{22}\} + \dots \quad) \\ &\quad + x_3(P\{s_{31}\} + P\{s_{32}\} + \dots \quad) \\ &= \sum_{i,j} X(s_{ij}) \, P\{s_{ij}\}. \end{aligned}$$

This shows that, in discrete probability models, our definition of $E(X)$ is equivalent to setting

$$E(X) = \sum_{s} X(s) P\{s\}.$$

This means that $E(X)$ can also be interpreted as a weighted average of the values of X at each of the sample points where the weights are the probabilities $P\{s\}$.

One of the implications of this second representation is that, if X and Y are any two random variables on S having finite expectations, and if $Z = X + Y$, then $E(Z) = E(X) + E(Y)$, because

$$E(Z) = \sum_s Z(s)P\{s\} = \Sigma[X(s) + Y(s)]P\{s\} = \Sigma X(s)P\{s\} + \Sigma Y(s)P\{s\}$$
$$= E(X) + E(Y).$$

Also, if $W = aX + b$ where a and b are any constants, then

$$E(W) = \sum_s W(s)P\{s\} = \Sigma[aX(s) + b]P\{s\} = a\Sigma X(s)P\{s\} + b\Sigma P\{s\}$$
$$= aE(X) + b.$$

This motivates the following results, which are true for all probability models, not just discrete ones. (Ref. Pfeiffer, Chapter 5.)

Theorem 5-1. If X and Y are any two random variables that have finite expectations, then

(a) $E(X + Y) = E(X) + E(Y)$, and

(b) $E(aX + b) = aE(X) + b$ for any constants a and b.

Corollary. If $X_1, X_2, \ldots, X_n$ are n random variables having finite expectations, then $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$.

Example. A gambler at the "craps" tables in Las Vegas can place a 4-to-1 bet on the occurrence of "7" when two fair dice are tossed. If he bets a dollar and "7" occurs, he wins \$4; otherwise, he loses \$1. Let G be his gain in dollars on a single trial. Since the probability of winning on each toss is 1/6, $P(G = 4) = 1/6$ and $P(G = -1) = 5/6$, so that

$$E(G) = 4(1/6) - 1(5/6) = -1/6.$$

Alternatively, one could set X equal to 1 or 0 according as the result is "7" or not, in which case $G = 5X - 1$ and

$$E(G) = 5E(X) - 1 = 5(1/6) - 1 = -1/6.$$

Suppose that he bets a dollar on every toss of the dice for an hour where tosses occur at a rate of two a minute, and let G_i be his gain on the ith trial. Then his expected overall gain on the 120 trials is

$$E(\Sigma_{i=1}^{120} G_i) = \Sigma_{i=1}^{120} E(G_i) = 120 \cdot (-1/6) = -20.$$

Another implication of the representation $E(X) = \Sigma X(s)\, P\{s\}$ for discrete probability models is that, if $Y = g(X)$ where g is some real-valued function on R, then

$$\begin{aligned} E(Y) &= \sum_s Y(s)P\{s\} = \Sigma g(X(s))P\{s\} \\ &= g(x_1)P\{s:X(s) = x_1\} + g(x_2)P\{s:X(s) = x_2\} + \ldots \\ &= \sum_k g(x_k)p(x_k), \end{aligned}$$

where p is the probability function of X. That is, one can compute the expectation of $Y = g(X)$ without first deriving the probability function of Y.

Theorem 5-2. If X is a discrete random variable having probability function p and if the expectation of $Y = g(X)$ exists, then

$$E(Y) = \Sigma g(x_k)\, p(x_k).$$

The more general applicability of the theorems above becomes apparent when two facts are observed. First, the expectation of a discrete random variable X depends only on the probability function p of X and not on the nature of the sample space upon which X is defined. Therefore, in considering expectations of discrete random variables (or functions of discrete random variables), there is no loss of generality in assuming that the underlying sample space is discrete. Second, for any random variable X on any sample space S there is a discrete random variable X_n such

that $|X(s) - X_n(s)| \leq 1/n$ for all sample points s, namely,

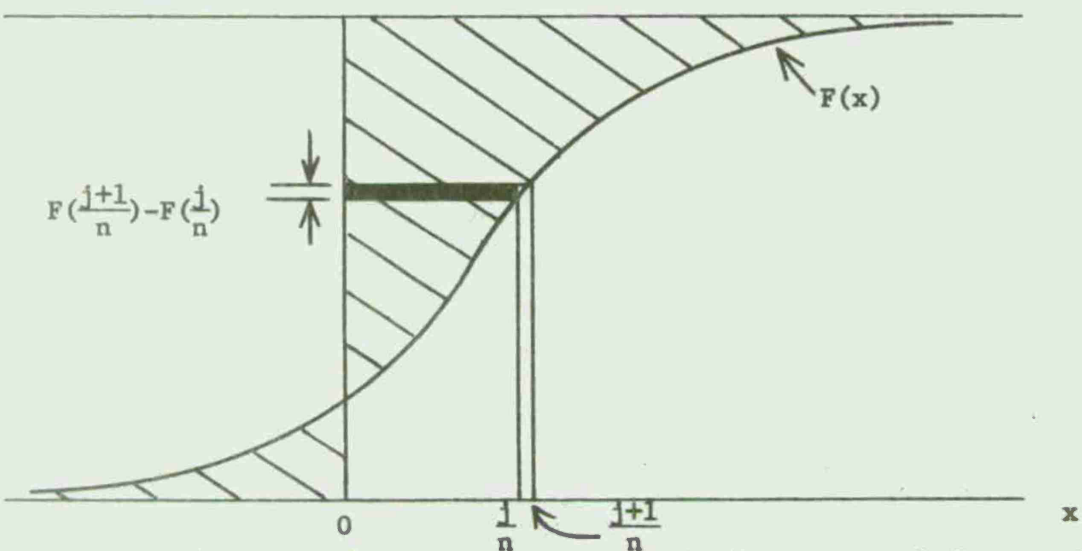
$$X_n(s) = j/n \quad \text{if} \quad j/n < X(s) \leq (j+1)/n$$

where j is restricted to integer values.

Using this second observation, one is motivated to define the expectation of any random variable X as the limit of the expectations of the discrete random variables X_n, assuming that the limit exists. If X has distribution function F, then $P(X_n = j/n) = F(\frac{j+1}{n}) - F(\frac{j}{n})$, so that

$$(1) \qquad E(X_n) = \sum_j (\tfrac{j}{n})[F(\tfrac{j+1}{n}) - F(\tfrac{j}{n})]$$

As the figure below indicates, as $n \to \infty$, the sum of the positive terms in



$E(X_n)$ tends to the area of the shaded portion to the right of the origin, and the sum of the negative terms tends to the negative of the area of the shaded portion to the left. This provides a valid geometrical interpretation of $E(X)$ as the difference between the two shaded areas depicted.[1]

[1] More precisely, $E(X) = \int_0^\infty [1 - F(x)]dx - \int_{-\infty}^0 F(x)dx$. If X is a non-negative random variable, then the second term is zero, and $E(X) = \int_0^\infty [1 - F(x)]dx = \int_0^\infty P(X > x)dx$.

Now suppose X is a continuous random variable having density function f. Then, (1) above can be written as

$$(2) \qquad E(X_n) = \sum_j (j/n) \int_{j/n}^{(j+1)/n} f(x)\,dx = \int_{-\infty}^{\infty} g_n(x)\cdot f(x)\,dx$$

where $g_n(x) = j/n$ if $j/n < x < (j+1)/n$. Since $g_n(x) \to x$ as $n \to \infty$, it follows that

$$E(X_n) \to \int_{-\infty}^{\infty} x\, f(x)\,dx$$

provided that $\int |x|\, f(x)\,dx < \infty$. This motivates the following definition:

Definition.

Let X be a continuous random variable having density function f. Then the expectation of X is defined by

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

provided that $\int_{-\infty}^{\infty} |x| f(x)\,dx < \infty$.

Examples.

1. Suppose Y has density $f(y) = 2y$ for $0 < y < 1$.

Then $E(Y) = \int_0^1 2y^2\,dy = 2/3$.

2. Let X be the waiting time in the telephone problem. (See page 34.)

Then X has density $f(y) = \lambda e^{-\lambda x}$ for $x > 0$ where $\lambda = 1/2$, and

$$E(X) = \int_0^{\infty} x\lambda e^{-\lambda x}dx = -\, xe^{-\lambda x}\Big|_{x=0}^{\infty} + \int_0^{\infty} e^{-\lambda x}dx = 1/\lambda = 2.$$

3. If Z has density $f(z) = 1/\pi(1+z^2)$, then $E(Z)$ does not exist because $\int_{-\infty}^{\infty} |z|/\pi(1+z^2)\,dz = \infty$.

Note that, in the above definition of expectation for the continuous case as well as in the corresponding definition for the discrete case, the expected value of a random variable X is analogous to the centroid (or center of gravity) of a unit mass spread out on the line according to the

probability distribution of X. In the discrete case, if one has masses $p(x_1)$, $p(x_2)$, ... at the points x_1, x_2, ... on the line, then the centroid of that distribution of masses is at $E(X) = \Sigma x_k\, p(x_k)$. Similarly, if a unit mass is distributed continuously over the real line according to the density function f, then the centroid of the distribution of mass is at $\int x\, f(x)\, dx$. The following theorem becomes apparent from this interpretation of E(X).

Theorem 5-3. If a random variable X having finite expectation has a probability or density function that is symmetric about a point c, then $E(X) = c$.

A second measure of the center of a distribution is the "median." Roughly speaking, the median of a distribution is a value such that half of the probability lies to the left of the value and half to the right with an appropriate adjustment for the discrete case.

Definition. The median of a random variable X (or of the distribution of X) is defined to be any value m such that $P(X \geq m) \geq 1/2$ and $P(X \leq m) \geq 1/2$.

For example, if X has probability function $p(x_i) = 1/n$ where $x_1, x_2, \ldots, x_n$ are n distinct real numbers such that $x_1 < x_2 \ldots < x_n$, then the median of X is $x_{(n+1)/2}$ if n is odd, and any number between $x_{n/2}$ and $x_{(n+2)/2}$ if n is even. Ordinarily, in the latter case, one defines the median to be the average of $x_{n/2}$ and $x_{(n+2)/2}$. Thus, if $n = 10$, the median is usually defined as the average of x_5 and x_6.

If X has a continuous distribution, then there is at least one value m such that $P(X \leq m) = 1/2$. Since $P(X \leq x) = F(x)$ where F is the distribution function of X, the median of X is any solution of the equation $F(m) = 1/2$.

Whether X has a continuous distribution or not, if the distribution is symmetric about some point c, then the median of the distribution is equal to c.

Exercises.

1. Let X be the number of hats returned correctly in the hat-check girl problem. (See page 29.) Show that the median of X is 1, and $E(X) = 1$. Verify that the geometric interpretation of $E(X)$ given on page 44 holds in this case.

2. Show that if Y has density function $f(y) = (2-y)/2$ for $0 < y < 2$, then $E(Y) = 2/3$, and the median of Y is $2 - \sqrt{2}$.

3. Show that if X has density function $f(x) = 1/(b - a)$ for $a < x < b$, then $E(X) = (a + b)/2$, and the median of X has the same value.

The linearity properties of expectation specified in Theorem 5-1 hold whether the random variables are discrete or not. The theorem that corresponds to Theorem 5-2 in the continuous case is:

Theorem 5-4. If X is a continuous random variable having density function f and if $Y = g(X)$ is a random variable such that $E[g(X)]$ exists, then

$$E(Y) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx.$$

Exercises.

1. Let X be the random variable in the spinner problem, and let $Y = X^2$. Apply the theorem above to show that $E(Y) = 1/3$.

2. Show that the density function of Y in the preceding problem is $f(y) = 1/2\sqrt{y}$ for $0 < y < 1$. Compute $E(Y)$ from the definition and thus verify the result in Problem 1.

Definition. The variance of a random variable X, denoted by $\mathrm{Var}(X)$ or σ_X^2, is defined by $E(X - \mu)^2$ where $\mu = E(X)$, provided that this expectation exists. The standard deviation of X, denoted by σ_X, is defined as the positive square root of the variance.

The variance and standard deviation are measures of the "spread" of the distribution of X. Another measure of spread is the mean absolute deviation, defined as $E|X - \mu|$. The reason that the variance and standard deviation are more widely used is that these measures are more tractable for reasons that will become apparent later.

Examples.

1. If the distribution of X is entirely concentrated at a single point c, so that $P(X = c) = 1$, then $E(X) = c$ and $\mathrm{Var}(X) = 0$.

2. Let X be the number of heads in five tosses of a fair coin. Then X has probability function $p(x) = \binom{5}{x}(1/2)^x(1/2)^{5-x} = \binom{5}{x}(1/2)^5$. The values of p are as follows:

x	0	1	2	3	4	5
$p(x)$	1/32	5/32	5/16	5/16	5/32	1/32

By the symmetry of p around $x = 5/2$, it follows from Theorem 5-3 that $E(X) = 5/2$. The value of $\mathrm{Var}(X)$ can be computed directly from the definition:

$$\mathrm{Var}(X) = E(X - \mu)^2 = \Sigma_{x=0}^{5} (x - 5/2)^2 p(x)$$

$$= (-5/2)^2(1/32) + (-3/2)^2(5/32) + (-1/2)^2(5/16) + (1/2)^2(5/16)$$

$$+ (3/2)^2(5/32) + (5/2)^2(1/32) = 5/4.$$

Thus, the standard deviation of X is $\sigma_X = \sqrt{5}/2 = 1.12$.

The following theorem often facilitates the calculation of variance.

Theorem 5-5. If X is a random variable for which $E(X) = \mu$ and $E(X^2) < \infty$ and if a and b are any constants, then

(a) $\mathrm{Var}(X) = E(X^2) - \mu^2$.

(b) $\mathrm{Var}(X + b) = \mathrm{Var}(X)$.

(c) $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, and $\sigma_{aX} = |a|\sigma_X$.

(d) $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

Proof: $\mathrm{Var}(X) = E(X - \mu)^2 = E(X^2 - 2\mu X + \mu^2)$. Using the linearity properties of expectation (Theorem 5-1) gives

$$\mathrm{Var}(X) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2.$$

Parts (b) and (c) follow from (d):

$$\mathrm{Var}(aX + b) = E(aX + b - a\mu - b)^2 = Ea^2(X - \mu)^2$$
$$= a^2E(X - \mu)^2 = a^2\mathrm{Var}(X).$$

Examples.

1. Applying part (a) of the above theorem, one could have computed $\mathrm{Var}(X)$ in the previous example by first computing $E(X^2)$:

$$E(X^2) = \Sigma x^2 p(x) = 0(1/32) + 1(5/32) + 4(5/16) + 9(5/16) + 16(5/32) + 25(1/32) = 15/2.$$

Hence, $\mathrm{Var}(X) = E(X^2) - \mu^2 = 15/2 - (5/2)^2 = 5/4$.

2. Let X be a discrete random variable having probability function $p(x_i) = 1/n$ where $x_1, x_2, \ldots, x_n$ are n distinct real numbers. Then since $E(X) = \bar{x} = \Sigma x_i/n$, $\mathrm{Var}(X) = \Sigma(x_i - \bar{x})^2/n$. Applying Theorem 5-5(a), one can compute $\mathrm{Var}(X)$ in this case using the formula $\mathrm{Var}(X) = (\Sigma x_i^2/n) - \bar{x}^2$.

Theorem 5-6.

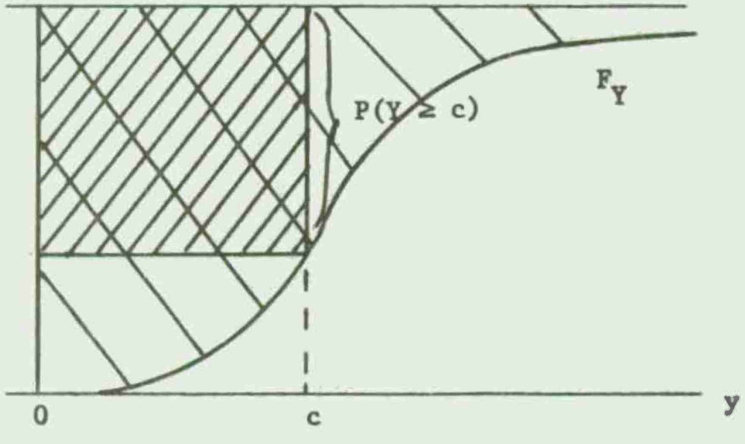
(a) If Y is a nonnegative random variable, then $P(Y \geq c) \leq E(Y)/c$ for all $c > 0$.

(b) (Chebyshev's Inequality) For any random variable X having finite variance σ^2,

$$P(|X - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2 \quad \text{for all} \quad \epsilon > 0.$$

In particular, $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ for all $k > 0$.

Proof: (a) It follows immediately from the geometric interpretation of $E(Y)$ that $E(Y) \geq c\, P(Y \geq c)$ for all $c > 0$. See the figure below.



(b) $P(|X - \mu| \geq \epsilon) = P((X - \mu)^2 \geq \epsilon^2) \leq E(X - \mu)^2/\epsilon^2 = \sigma^2/\epsilon^2$.

It follows from part (b) of the theorem that $P(|X - \mu| < k\sigma) \geq 1 - 1/k^2$ for all $k > 0$. The table on the next page compares these "Chebyshev bounds" on the probabilities $P(|X - \mu| < k\sigma)$ with the actual probabilities for two distributions:

(A) The distribution of the number of heads in five tosses of a fair coin. (See Example 2, page 48.)

(B) The continuous distribution having the "bell-shaped" density function $f(x) = (2\pi)^{-1/2}\, e^{-x^2/2}$, which has mean 0 and variance 1.

Table 1

A COMPARISON OF CHEBYSHEV BOUNDS WITH ACTUAL PROBABILITIES

	$P(\lvert X - \mu \rvert < k\sigma)$		
k	Chebyshev bound	Actual (A)	Actual (B)
1	≥ 0	5/8 = 0.625	0.683
2	$\geq 3/4$	15/16 = 0.938	0.954
3	$\geq 8/9$	1	0.997
4	$\geq 15/16$	1	1.000
5	$\geq 24/25$	1	1.000

Exercises.

1. The probability function of the random variable X in the hat-check problem was:

x	0	1	2	3	4
$p(x)$	3/8	1/3	1/4	0	1/24

Here, $E(X) = 1$. Compute $Var(X)$ directly from the definition, and check your result by computing $Var(X)$ using the formula $Var(X) = E(X^2) - E^2(X)$.

2. Five balls are chosen at random from an urn containing 9 balls of which 3 are white. Let X be the number of white balls in the sample. Find $Var(X)$ if the sampling is done (a) with replacement, (b) without replacement. (See Exercise 1, page 36 and Exercise 2, page 40.)
Ans. (a) 10/9, (b) 5/9.

3. Let X be the random variable in the spinner problem, so that X has density function $f(x) = 1$ for $0 < x < 1$. (a) Show that $Var(X) = 1/12$. (b) Show that $P(|X - E(X)| < 2\sigma) = 1$ and $P(|X - E(X)| < \sigma) = 1/\sqrt{3} = 0.577$.

4. Show that if X has mean μ and variance σ^2, then $Z = (X - \mu)/\sigma$ has mean 0 and variance 1.

5. Suppose X has density function $f(x) = (2 - x)/2$ for $0 < x < 2$.

(a) Sketch the density function of X and find $P(0 < X < 1)$.

(b) Find and sketch the distribution function of X.

(c) Find $E(X)$ and $Var(X)$.

(d) Find $P(|X - \mu| \geq 2\sigma)$.

Ans. (a) 3/4, (b) $F(x) = 0$ for $x < 0$, $x(4 - x)/4$ for $0 \leq x \leq 2$, 1 for $x > 2$, (c) 2/3, 2/9, (d) 0.04.

6. If X is the sum of two numbers chosen independently and at random between 0 and 1, then X has density $f(x) = 1 - |1 - x|$ for $0 < x < 2$. Find (a) $P(1/2 < X < 3/2)$, (b) $E(X)$, (c) $Var(X)$, (d) $P(|X - \mu| > 2\sigma)$.

Ans. (a) 3/4, (b) 1, (c) 1/6, (d) 0.03.

7. Show that, if X is a random variable such that $Var(X)$ exists, then among all real numbers c, $E(X - c)^2$ is minimized by $c = E(X)$.

SECTION VI. - SOME SPECIAL DISTRIBUTIONS

References:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1960, Chapters 8-9.

Seymour Lipschutz, Theory and Problems of Probability, Schaum's Outline Series, McGraw-Hill, New York, 1968, Chapter 6.

The table on the next page gives the probability (or density) functions, means, and variances of some frequently encountered distributions. Examples of random variables that have these distributions are given below.

Bernoulli. Any random variable that takes on only the two values 1 and 0 with probabilities p and $q = 1-p$ has a Bernoulli distribution with parameter p.

Binomial. The number of successes in n Bernoulli trials with probability p of success on each trial has a binomial distribution with parameters n and p. (See page 26.)

Hypergeometric. If X is the number of defectives in a sample of size n taken without replacement from a lot of N items of which Np are defective, then X has a hypergeometric distribution. (See Theorem 2-2.)

The values of the probability function of the hypergeometric distribution for certain values of n, p, and N are given in Table 2. In each case, the values of n and p are chosen so that the expected number of defectives is $E(X) = np = 2$. Note that for fixed values of n and p the distribution becomes more variable as the population size N increases. Since the variance of the hypergeometric distribution is $\mathrm{Var}(X) = npq(\frac{N-n}{N-1})$, as $N \to \infty$ the variance tends to npq, the variance of a binomial distribution with parameters n and p.

If the sample of size n is taken with replacement instead of without replacement, then X has a binomial distribution with parameters n and p. As intuition would suggest, if the population size is much larger than the

Table 1

A SHORT TABLE OF DISTRIBUTIONS

Distribution and range of parameters	Probability or density function	Mean E(X)	Variance Var(X)
Bernoulli (p) $0 < p < 1$	$p^x q^{1-x}$, $x = 0,1$	p	pq
Binomial (n,p) $0 < p < 1$ $n = 1,2,\ldots$	$\binom{n}{x} p^x q^{n-x}$, $x = 0,1,\ldots,n$	np	npq
Hypergeometric $N = 1,2,\ldots$ $n = 1,2,\ldots,N$ $p = 0, 1/N,\ldots,(N-1)/N,1$	$\dfrac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$, $x = 0,1,\ldots,n$	np	$npq\left(\frac{N-n}{N-1}\right)$
Poisson (λ) $\lambda > 0$	$\dfrac{e^{-\lambda}\lambda^x}{x!}$, $x = 0,1,2,\ldots$	λ	λ
Geometric $0 < p < 1$	pq^{x-1}, $x = 1,2,\ldots$	$1/p$	q/p^2
Negative Binomial $0 < p < 1$ $r = 1,2,\ldots$	$\binom{x-1}{r-1} p^r q^{x-r}$, $x = r, r+1,\ldots$	r/p	rq/p^2
Uniform (a,b) $-\infty < a < b < \infty$	$\dfrac{1}{b-a}$, $a < x < b$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Normal (μ, σ^2) $\sigma > 0$	$\dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2
Negative Exponential (λ) $\lambda > 0$	$\lambda e^{-\lambda x}$, $x > 0$	$1/\lambda$	$1/\lambda^2$
Gamma (r,λ) $r > 0$, $\lambda > 0$	$\dfrac{\lambda(\lambda x)^{r-1}}{\Gamma(r)} e^{-\lambda x}$, $x > 0$	r/λ	r/λ^2
Chi-square (n) $n = 1,2,\ldots$ See Gamma $(\frac{n}{2},\frac{1}{2})$	$\dfrac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$, $x > 0$	n	$2n$
Cauchy (μ,λ) $\lambda > 0$	$\dfrac{\lambda}{\pi\{\lambda^2 + (x-\mu)^2\}}$	∞	∞
Laplace (μ,λ) $\lambda > 0$	$\dfrac{1}{2\lambda} e^{-\lvert x-\mu\rvert/\lambda}$	μ	$2\lambda^2$
Pareto (α,c) $c > 0$, $\alpha > 0$	$\dfrac{\alpha}{c}\left(\dfrac{c}{x}\right)^{\alpha+1}$, $x > c$	$\dfrac{\alpha c}{\alpha-1}$ if $\alpha > 1$	$\dfrac{\alpha c^2}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$

Table 2

A COMPARISON OF HYPERGEOMETRIC, BINOMIAL, AND POISSON PROBABILITIES

Sample size	p	x	Hypergeometric N=5	10	20	50	100	Binomial	Poisson
5	0.4	0	-	.024	.051	.067	.073	.078	.135
		1	-	.238	.255	.259	.259	.259	.271
		2	1.0	.476	.397	.364	.354	.346	.271
		3	-	.238	.238	.234	.232	.230	.180
		4	-	.024	.054	.069	.073	.077	.090
		5	-	-	.004	.007	.009	.010	.036
10	0.2	0		-	.043	.083	.095	.107	.135
		1		-	.248	.266	.268	.268	.271
		2		1.0	.418	.337	.318	.302	.271
		3		-	.248	.218	.209	.201	.180
		4		-	.043	.078	.084	.088	.090
		5		-	-	.016	.022	.026	.036
		6		-	-	.002	.004	.006	.012
		7		-	-	.000	.000	.001	.003
		8		-	-	.000	.000	.000	.001
		9		-	-	.000	.000	.000	.000
		10		-	-	.000	.000	.000	.000
20	0.1	0			-	.067	.095	.122	.135
		1			-	.259	.268	.270	.271
		2			1.0	.364	.318	.285	.271
		3			-	.234	.209	.190	.180
		4			-	.069	.084	.090	.090
		5			-	.007	.022	.032	.036
		6			-	-	.004	.009	.012
		7			-	-	.000	.002	.003
		8			-	-	.000	.000	.001
		9-20			-	-	.000	.000	.000

sample size, then the hypergeometric probabilities $P(X = k)$ differ little from the corresponding binomial probabilities, and as $N \to \infty$ the hypergeometric probabilities tend to the binomial probabilities. Table 2 compares the two sets of probabilities for $N = 100$ and for three sample sizes $n = 5$, 10, and 20.

Poisson. Suppose that events of a certain type (such as traffic accidents, arrivals at a checkout counter, emissions of α-particles from a radioactive source, vacancies in the Supreme Court during a year, etc.) are occurring randomly over time in such a way that certain assumptions are satisfied (e.g., the events occur singly, and the numbers of occurrences in disjoint time intervals are "independent"). Then the number of occurrences X in a unit time interval can be assumed to have a Poisson distribution with parameter λ, where λ is the mean number of occurrences in an interval of length one.[1] The number of occurrences in a time interval of length t has a Poisson distribution with parameter λt.

The Poisson distribution also arises as a limit of binomial distributions as $n \to \infty$ and $p \to 0$ in such a way that $np \to \lambda$. Table 2 gives the Poisson probabilities for $\lambda = 2$. Compare these probabilities with the binomial probabilities for (a) $n = 5$, $p = 0.4$; (b) $n = 10$, $p = 0.2$; and (c) $n = 20$, $p = 0.1$. In all three cases, $np = 2$. Note that as n increases, the differences between the binomial and Poisson probabilities become smaller.

Geometric and Negative Binomial. These distributions occur in considering the number of Bernoulli trials required until a certain number of successes occur. If X is the number of trials required until r successes occur, then X has a negative binomial distribution with parameters r and p, where p is the probability of a success on each trial. If X is the waiting time for the first success (i.e., the special case where $r = 1$), then X has a geometric distribution. For example, if two fair dice are tossed again and again until a total of seven occurs for the first time, then the number of the trial on which seven occurs has a geometric distribution with parameter $p = 1/6$, and the expected number of trials is 6.

---

[1] Meyer, op. cit., pp. 166-168.

Uniform. A random variable U has a uniform distribution on an interval (a,b) if the probability that U takes on values in any subinterval (c,d) of (a,b) is proportional to the length of the subinterval, and the probability that U takes on values outside the interval (a,b) is zero. For example, the random variable in the spinner problem has a uniform distribution on $(0,1)$.

Normal. This distribution is the most frequently used of all distributions in statistical applications for two reasons: (a) many statistical calculations are greatly simplified if the random variables involved are assumed to have normal distributions, (b) the normal distribution provides a reasonable approximation for distributions of repeated measurements of many physical phenomena--cranial lengths, ballistic measurements (coordinates of deviations from the target), logarithms of incomes, heights, IQ scores, sums or averages of several test scores, etc. The normal distribution is also the limiting distribution of many distributions (binomial, hypergeometric, Poisson, negative binomial, and distributions of sums and averages of random variables that satisfy certain properties).

A random variable Z is said to have a standard normal distribution if Z has density function $\varphi(z) = (2\pi)^{-1/2} e^{-z^2/2}$ for $-\infty < z < \infty$. This bell-shaped density function is symmetric about zero. It is easily verified that $E(Z) = 0$ and $Var(Z) = 1$. The distribution function of Z, commonly denoted by Φ in the statistical literature, is tabulated in Table 3. For example, $P(Z < 2) = \Phi(2) = 0.9772$. The values of $\Phi(z)$ for negative values of z can be computed using the formula $\Phi(z) = 1 - \Phi(-z)$, which follows from the symmetry of the distribution about zero. For example, $P(Z < -2) = 1 - \Phi(2) = 0.0228$. Note that $P(-2 < Z < 2)$ is approximately 0.95.

If X is a random variable such that $Z = (X - \mu)/\sigma$ has a standard normal distribution, then X is said to have a normal distribution with parameters μ and σ^2, which is often abbreviated to $X \sim N(\mu,\sigma^2)$.

Exercise. Verify that (a) the random variable Z having a standard normal distribution has mean 0 and variance 1, (b) $P(-1 < Z < 1) = 0.68$, (c) $X = \mu + \sigma Z$ has mean μ and variance σ^2, (d) X has density function

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

Table 3

CUMULATIVE NORMAL DISTRIBUTION

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

x	.00	.01	.02	.03	.01	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6661	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9901	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

x	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$\Phi(x)$	.90	.95	.975	.99	.995	.999	.9995	.99995	.999995
$2[1-\Phi(x)]$	.20	.10	.05	.02	.01	.002	.001	.0001	.00001

If $X \sim N(\mu,\sigma^2)$, then one can use a table of the standard normal distribution function to compute any probability of the form $P(a < X < b)$:

$$P(a < X < b) = P(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}).$$

For example, if $X \sim N(28,4)$, then

$$P(25 < X < 27) = \Phi(\frac{27-28}{2}) - \Phi(\frac{25-28}{2}) = \Phi(-0.5) - \Phi(-1.5)$$

$$= 0.31 - 0.07 = 0.24.$$

The normal distribution frequently occurs as the limiting distribution of sums or averages of a large number of random variables. In the simplest case, consider a sequence of Bernoulli trials with probability p of success on each trial. Let X_i be 1 or 0 according as the ith trial is a success or not, and let $S_n = X_1 + X_2 + \ldots + X_n$. Then S_n is the number of successes in the first n trials, which has a binomial distribution with parameters n and p, so that $E(S_n) = np$ and $Var(S_n) = npq$. For large values of n, the distribution of S_n is approximately normal with mean $\mu = np$ and variance $\sigma^2 = npq$ in the sense that

$$P(a \le \frac{S_n - np}{\sqrt{npq}} \le b) \doteq \Phi(b) - \Phi(a)$$

for any real numbers $a < b$, and as $n \to \infty$ the probability on the left tends to the limit on the right. This is called the DeMoivre-Laplace Central Limit Theorem. For a proof, see W. Feller, An Introduction to Probability Theory and Its Applications, Volume I, 3rd Edition, John Wiley, 1968, pp. 182-186. (A more general result on the limiting distribution of sums of outcomes of independent trials is contained in Section VIII.)

It follows that for any integer k,

$$P(S_n \le k) = P(\frac{S_n - np}{\sqrt{npq}} \le \frac{k - np}{\sqrt{npq}}) \doteq \Phi(\frac{k - np}{\sqrt{npq}}).$$

This approximation is usually improved by first replacing $P(S_n \le k)$ by the equivalent quantity $P(S_n \le k + 1/2)$ and then proceeding as before to obtain

$$P(S_n \le k) = P(S_n \le k + 1/2) \doteq \Phi\left(\frac{k+(1/2)-np}{\sqrt{npq}}\right).$$

This so-called "continuity correction" is motivated by the fact that a step function (namely, the distribution function of S_n) is being approximated by a continuous function (the distribution function of a normal distribution) that tends to pass through the "midpoints" of the steps.

Table 4 on the next page compares the two normal approximations for the case where $n = 12$ and $p = 1/4$. The entries in the column headed "Poisson approximation" are the probabilities $P(X \le k)$ where X has a Poisson distribution with parameter $\lambda = 3$.

Example. If 55 percent of the voters of a large city are in favor of a given proposal, what is the probability that a random sample of 100 voters would not show a majority in favor?

Let X be the number in the sample favoring the proposal. If the sampling is done with replacement, then X has a binomial distribution with parameters $n = 100$ and $p = 0.55$, so that $E(X) = np = 55$ and $\sigma = \sqrt{npq} = 4.98$. Hence

$$P(X \le 50) = P(X \le 50.5) = P\left(\frac{X-55}{4.98} \le \frac{50.5-55}{4.98}\right) \doteq \Phi(-0.90) = 0.18.$$

Exercises. 1. A man claims to be able to predict whether a fair coin will result in heads before it is flipped. To test his contention you toss a fair coin 100 times and record the number of times that he predicts the result correctly. What is the approximate probability that he will predict the result correctly 60 or more times if his predictions are mere guesses? Ans. 0.03.

2. Suppose that the lifetimes of components of a certain type have a $N(\mu,\sigma^2)$ distribution with $\mu = 1000$ hours and $\sigma = 100$ hours. What is the approximate probability that, among 45 components chosen at random from components of this type, 10 or more will last less than 900 hours? [To make the arithmetic easy, assume that $\Phi(-1) = 1/6$.] Ans. 0.21.

Table 4

A COMPARISON OF THE NORMAL AND POISSON APPROXIMATIONS TO THE BINOMIAL PROBABILITIES $P(S_n \leq k)$ FOR THE CASE $n = 12$, $p = 0.25$

k	$P(S_n \leq k)$	Normal approximation		Poisson approximation
		Without continuity correction	With continuity correction	
0	.0317	.0228	.0478	.0498
1	.1584	.0913	.1587	.1991
2	.3907	.2525	.3695	.4232
3	.6488	.5000	.6305	.6472
4	.8424	.7475	.8413	.8153
5	.9456	.9087	.9522	.9161
6	.9857	.9772	.9902	.9665
7	.9972	.9962	.9987	.9881
8	.9996	.9996	.9999	.9962
9	1.0000	1.0000	1.0000	.9989
10	1.0000	1.0000	1.0000	.9997
11	1.0000	1.0000	1.0000	.9999
12	1.0000	1.0000	1.0000	1.0000

The Lognormal Distribution. A random variable X is said to have a lognormal distribution if $Y = \log X$ has a $N(\mu,\sigma^2)$ distribution. This is equivalent to saying that X has a lognormal distribution if there is a normally distributed random variable Y such that X has the same distribution as e^Y. Since X has distribution function

$$F(x) = P(X \le x) = P(e^Y \le x) = P(Y \le \log x) = \int_{-\infty}^{\log x} f_Y(y)\,dy \quad \text{for } x > 0,$$

X has density function

$$f(x) = F'(x) = f_Y(\log x)\,\frac{d(\log x)}{dx}$$

$$= (1/\sigma x\sqrt{2\pi}) \exp\{-(\log x - \mu)^2/2\sigma^2\} \quad \text{for } x > 0.$$

Using the fact that $E(e^{tY}) = \exp\{\mu t + \sigma^2 t^2/2\}$ for all values of t (see Meyer, op. cit., p. 210), one can show that

$$E(X) = E(e^Y) = e^{\mu+\sigma^2/2}$$

$$\mathrm{Var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

The median of the distribution of X is e^{μ}. [In general, if Y is a random variable having median m, and if X is an increasing (or decreasing) function of Y, say $X = h(Y)$, then the median of X is $h(m)$.]

The Negative Exponential, Gamma, and Chi-square Distributions. Suppose that events of a certain type are occurring over time in such a way that X_t, the number of events up to time t, has a Poisson distribution with parameter λt for all values of t. Consider the waiting time T for exactly r events to occur. Then the distribution function of T is

$$F(t) = P(T \le t) = P(X_t \ge r) = 1 - \sum_{n=0}^{r-1} e^{-\lambda t}(\lambda t)^n/n! \quad \text{for } t > 0.$$

Therefore, the density function of T is

$$f(t) = F'(t) = -\sum_{n=1}^{r-1} e^{-\lambda t}(\lambda t)^{n-1}\lambda/(n-1)! + \sum_{n=0}^{r-1} (\lambda t)^n \lambda e^{-\lambda t}/n! \quad \text{for } t > 0.$$

Since the terms in the first sum are the negatives of the first r-1 terms

in the second sum, the density reduces to

$$f(t) = \lambda(\lambda t)^{r-1}e^{-\lambda t}/(r-1)! \quad \text{for} \quad t > 0.$$

A random variable having this density is said to have a gamma distribution with parameters r and λ. If $r = 1$, then

$$f(t) = \lambda e^{-\lambda t} \quad \text{for} \quad t > 0.$$

A random variable having this density is said to have a negative exponential distribution with parameter λ.

In general, if events are occurring randomly over time in such a way that the number of occurrences up to time t has a Poisson distribution with parameter t, then not only is it the case that the waiting time for the first occurrence has a negative exponential distribution with parameter λ, but also the waiting times between any two successive occurrences has a negative exponential distribution with the same parameter. Conversely, if the waiting times between successive occurrences are "independent" (see Section VII) and if these waiting times have a negative exponential distribution with parameter λ, then the number of occurrences in any fixed time interval of length t has a Poisson distribution with parameter λt. Thus, to generate a sequence of occurrences for which the Poisson model would apply, it suffices to generate random variables having negative exponential distributions. (See Exercise 2 below.)

The parameter r in the gamma distribution was assumed to be a positive integer above, but the gamma distribution can be defined for all positive values of r by specifying the density as

$$f(t) = \lambda(\lambda t)^{r-1}e^{-\lambda t}/\Gamma(r) \quad \text{for} \quad t > 0,$$

where Γ is the gamma function defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx.$$

It can be shown using integration by parts that

$$\Gamma(r) = (r-1)\Gamma(r-1),$$

and since $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$, it follows that $\Gamma(r) = (r-1)!$ for all positive integers r. It can be shown that $\Gamma(1/2) = \sqrt{\pi}$. Applying the formula above, one can compute $\Gamma(3/2) = \sqrt{\pi}/2$, $\Gamma(5/2) = 3\sqrt{\pi}/4$, etc.

The chi-square distribution with n degrees of freedom, which will be discussed in Section VIII, is a special case of the gamma distribution with parameters $r = n/2$ and $\lambda = 1/2$.

Exercises.

1. Show that, if X has a gamma distribution with parameters r and λ, then $E(X) = r/\lambda$ and $Var(X) = r/\lambda^2$.

2. Show that, if U has a Uniform(0,1) distribution, then $T = -\log U$ has a negative exponential distribution with parameter $\lambda = 1$, and $V = T/\lambda$ has a negative exponential distribution with parameter λ

The Cauchy Distribution. A random variable X is said to have a Cauchy distribution with parameters μ and $\lambda > 0$ if X has density function

$$f(x) = \frac{\lambda}{\pi[\lambda^2 + (x-\mu)^2]}, \quad -\infty < x < \infty .$$

Since the Cauchy distribution has a bell-shaped density function that is symmetric about μ, the median of the distribution is μ. The distribution is of primary interest to statisticians as a source of counterexamples. The expectation and variance of random variables having this distribution do not exist, and certain averages of random variables having Cauchy distributions have peculiar properties that will be discussed in Section VIII.

Laplace Distribution. A random variable X is said to have a Laplace (or double exponential) distribution if it has density function

$$f(x) = \frac{1}{2\lambda} e^{-|x-\mu|/\lambda}, \quad -\infty < x < \infty .$$

This tent-shaped distribution, which is symmetric about its mean μ, is primarily of theoretical interest, in part because of problems related to estimating the parameter μ. The case $\mu = 0$ arises in considering differences of random variables that have negative exponential distributions.

Pareto Distribution. This distribution has density

$$f(x) = (\alpha/c)(c/x)^{\alpha+1} \qquad \text{for } x > c.$$

This arises in considering distributions of characteristics which have been "truncated" from below. For example, consider the distribution of incomes among families that have incomes exceeding \$20,000, or the distribution of rain-gauge readings after storms that yield more than one inch of rain. The parameter c above is the truncation point. Since $P(X > x) = (c/x)^{\alpha}$ for $x > c$ by Exercise 1 below, the parameter α indicates how rapidly the probability in the "tail" of the distribution tends to zero.

Other Truncated Distributions. The distribution of any random variable X can be truncated to the left (or right) at some point c by considering the (conditional) distribution of X on the set $\{X > c\}$ (or $\{X < c\}$). If X has density function $f(x)$, the conditional probability that $X \le x$ given that $X > c$ is

$$P(X \le x | X > c) = \frac{P(c < X \le X)}{P(X > c)} = \int_c^x f(x')dx'/[1-F(c)] \qquad \text{for } x > c.$$

This can be viewed as the "conditional" distribution function of X given that $X > c$. Taking the derivative of $P(X \le x | X > c)$ with respect to x yields the density function

$$g(x) = \begin{cases} 0 & \text{for } x \le c \\ f(x)/[1-F(c)] & \text{for } x > c. \end{cases}$$

This density function, which is zero for $x \le c$ and has the same shape as $f(x)$ for $x > c$, is said to be the density function of the distribution of X truncated to the left at $x = c$. Density functions of distributions truncated to the right and probability functions of distributions truncated to the left (or right) are defined similarly.

Examples. 1. If $X \sim N(\mu,\sigma^2)$, the density of the distribution of X truncated to the left at $x = c$ is

$$g(x) = \frac{K}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

where $K = 1/P(X > c) = [1 - \Phi(\frac{c-\mu}{\sigma})]^{-1} = [\Phi(\frac{\mu-c}{\sigma})]^{-1}$.
It can be shown that the expectation and variance of the truncated distribution are $\mu + \lambda\sigma$ and $(1 - \lambda^2)\sigma^2 + \lambda\sigma(c - \mu)$ where $\lambda = \varphi(\frac{\mu-c}{\sigma})/\Phi(\frac{\mu-c}{\sigma})$.
[See H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946, p. 249. The function $\Phi(t)/\varphi(t)$ is tabulated in D. B. Owen, Handbook of Statistical Tables, Addison-Wesley, Reading, Massachusetts, 1962, pp. 1-10.]

2. Suppose X has a negative exponential distribution with parameter λ. Then

$$P(X > c) = \int_c^\infty \lambda e^{-\lambda x} dx = e^{-\lambda c} \qquad \text{for all } c > 0.$$

The density of the distribution of X truncated to the left at $x = c$ is

$$g(x) = \lambda e^{-\lambda x}/e^{-\lambda c} = \lambda e^{-\lambda(x-c)} \qquad \text{for } x > c > 0.$$

In this case, the truncated density is the same as the original density except that it has been shifted c units to the right. It follows that the expectation and variance of the truncated distribution are $c + 1/\lambda$ and $1/\lambda^2$.

Exercises. 1. Show that, if X has the Pareto density with parameters α and c, then

(a) $P(X > x) = (c/x)^\alpha$ for $x > c$,

(b) $E(X) = \alpha c/(\alpha-1)$ for $\alpha > 1$.

Note that the expectation does not exist if $\alpha \le 1$.

2. Let T be the lifetime in hours of a component chosen at random from electronic components of a certain type. Then the probability $P(T > t)$ can be interpreted as the proportion of components of that type that last for more than t hours. In reliability theory, the function defined by

$$R(t) = P(T > t) \quad \text{for} \quad t > 0$$

is called the reliability function for these components. Clearly, $R(t) = 1 - F(t)$ where F is the distribution function of T. For example, if T has a negative exponential distribution with parameter λ then $R(t) = e^{-\lambda t}$ for $t > 0$.

(a) Suppose n components are chosen at random from components having reliability function $R(t)$, and all of them begin operating at the same time. Let $N(t)$ be the number of these components that are still operating after t hours. Show that $E[N(t)] = n\,R(t)$ and $P\{N(t) = n\} = [R(t)]^n$.

(b) Show that, if T has density function

$$f(t) = \lambda k t^{k-1} \exp(-\lambda t^k) \quad \text{for} \quad t > 0$$

where λ, k are positive parameters, then $R(t) = \exp(-\lambda t^k)$ for $t > 0$. A random variable having this density is said to have a Weibull distribution with parameters k and λ. Note that, if $k = 1$, this is the same as the negative exponential distribution with parameter λ.

(c) Show that the random variable T in part (b) has the same distribution as $X^{1/k}$ where X has a negative exponential distribution with parameter λ, and use this to show that the n^{th} moment of T is

$$E(T^n) = \lambda^{-n/k}\Gamma(n/k + 1).$$

SECTION VII. - JOINT DISTRIBUTIONS, CORRELATION, AND CONDITIONING

References:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1960, Chapter 9 and pp. 144-158.

Seymour Lipschutz, Theory and Problems of Probability, Schaum's Outline Series, McGraw-Hill, New York, 1968, Chapter 5.

Paul E. Pfeiffer, Concepts of Probability Theory, McGraw-Hill, New York, 1965, pp. 142-179.

Let X and Y be two random variables defined on the same sample space S. Just as a single random variable X carries probabilities from S into the line R, thereby determining a probability measure on R called the distribution of X, the pair of random variables (X,Y) carries probabilities into the plane R^2, determining a probability measure on the Borel subsets[1] of R^2 called the joint distribution of X and Y. In particular, the probability carried into the half-open rectangle $(a,b]\times(c,d]$ by (X,Y) is

$$P(a < X \leq b,\ c < Y \leq d) = P\{s:\ a < X(s) \leq b,\ c < Y(s) \leq d\}.$$

Definition. Two random variables X and Y are said to have a discrete joint distribution if there is a countable set $A = \{(x_j,y_k),\ j=1,2,\ldots,\ k=1,2,\ldots\}$ such that $P\{(X,Y) \in A\} = 1$. In this case, the function p defined on A by

$$p(x_j,y_k) = P(X = x_j,\ Y = y_k) \text{ for } j=1,2,\ldots,k=1,2,\ldots$$

is called the joint probability function of X and Y.

Clearly, $p(x,y) \geq 0$ for all (x,y) in A and $\Sigma\, p(x_j,y_k) = 1$. Also, if p_X and p_Y are the probability functions of X and Y, then

$$p_X(x_j) = \Sigma_k\, p(x_j,y_k) \text{ for } j=1,2,\ldots, \text{ and } p_Y(y_k) = \Sigma_j\, p(x_j,y_k) \text{ for } k = 1,2,..$$

---

[1] The class of Borel subsets of R^2 is the smallest collection of sets that contains the rectangles $(a,b]\times(c,d]$ and is closed under countable set operations.

In this context, p_X and p_Y are called the marginal probability functions of X and Y to distinguish them from the joint probability function p. If the joint probability function is given by a two-way table as in the example below, then the marginal probability functions can be obtained by summing the rows and columns in the table.

Example. A fair coin is tossed four times. Let X be the number of heads on the first two tosses, and let Y be the number of heads on all four tosses. Then the joint and marginal probability functions of X and Y are as follows:

y \ x	0	1	2	$p_Y(y)$
0	1/16	0	0	1/16
1	1/8	1/8	0	1/4
2	1/16	1/4	1/16	3/8
3	0	1/8	1/8	1/4
4	0	0	1/16	1/16
$p_X(x)$	1/4	1/2	1/4	1

The joint distribution of any pair of random variables is completely determined by their joint distribution function F, which is defined by

$$F(x,y) = P(X \le x, Y \le y) \quad \text{for all } x \text{ and } y.$$

To distinguish the joint distribution function from the individual distribution functions of X and Y, the latter are referred to as the marginal distribution functions in this context. The marginal distribution functions can be determined from the joint distribution function by

$$F_X(x) = F(x,\infty) \quad \text{and} \quad F_Y(y) = F(\infty,y).$$

The "bivariate" distribution function F has properties analogous to those in the "univariate" case (see page 34).

(a) $0 \le F(x,y) \le 1$ for all (x,y) in R^2.

(b) $F(x,-\infty) = F(-\infty,y) = 0$ for all x and y, and $F(\infty,\infty) = 1$.

(c) F is monotonically increasing and right continuous in each of its arguments.

(d) $P(a < X \le b,\ c < Y \le d) = F(b,d) - F(b,c) - F(a,d) + F(a,c)$.

Although the joint distribution function is of theoretical interest since it characterizes any type of joint distribution, it is hard to visualize and awkward to work with. Therefore, in practice, the joint distribution of a pair of random variables is ordinarily specified by giving either their joint probability function or their joint density function, which is defined as follows:

Definition. Two random variables X and Y are said to have a continuous (or absolutely continuous) joint distribution if there is a nonnegative function f on R^2 (called the joint density function of X and Y) such that for all (x,y)

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x',y')\, dy'\, dx'.$$

This is equivalent to saying that X and Y have a continuous distribution if there is a nonnegative function f on R^2 such that for all real numbers a,b,c, and d with $a < b$ and $c < d$

$$P(a < X < b,\ c < Y < d) = \int_a^b \int_c^d f(x,y)\, dy\, dx.$$

Hence, in this case, the probability $P(a < X < b,\ c < Y < d)$ has the geometrical interpretation as the volume under the surface $z = f(x,y)$ and above the rectangle $(a,b)\times(c,d)$.

If X and Y have joint density function f, then the "marginal" density function of X is $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$, because $F_X(x) = F(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x',y)\, dy\, dx'$ and it follows from the definition of the density function (see page 35) that X has the density function $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$.

Example. Let X and Y be the successive waiting times for two calls coming into a telephone exchange. Suppose that X and Y have joint density function

$$f(x,y) = e^{-(x+y)} \quad \text{for} \quad x > 0,\ y > 0.$$

[Assume here and below that $f(x,y) = 0$ for values of x and y other than those for which the functional form is specified.] In this case, the marginal density function of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_0^{\infty} e^{-(x+y)}\, dy = e^{-x} \quad \text{for} \quad x > 0.$$

By the symmetry of the joint density function, Y has the same density function as X. The following two examples illustrate how the joint density function can be used in computing probabilities of events:

(a) $$P(\min(X,Y) > 2) = P(X > 2,\ Y > 2) = \int_2^{\infty}\int_2^{\infty} e^{-(x+y)}\, dy\, dx$$

$$= \int_2^{\infty} e^{-x}\left[\int_2^{\infty} e^{-y}\, dy\right]dx = e^{-2}\int_2^{\infty} e^{-x}\, dx = e^{-4}.$$

(b) $$P(X + Y < 2) = \int_0^2\int_0^{2-x} e^{-(x+y)}\, dy\, dx$$

$$= \int_0^2 e^{-x}\,[1 - e^{x-2}]\, dx = \int_0^2 (e^{-x} - e^{-2})dx = 1 - 3e^{-2}.$$

Given the joint density of X and Y, one can (in theory) derive the distribution of random variables Z that are functions of X and Y. For example, let $Z = X + Y$. Then the distribution function of Z for $z > 0$ is

$$F(z) = P(Z \le z) = P(X + Y \le z) = 1 - e^{-z} - ze^{-z}.$$

The last expression follows by a calculation like that in (b) above. It follows that Z has the density function

$$f_Z(z) = F'(z) = e^{-z} + ze^{-z} - e^{-z} = ze^{-z} \quad \text{for} \quad z > 0.$$

Exercises.

1. Three balls are placed at random into one of three cells. Let X be the number of balls in cell #1 and Y the number in cell #2.

(a) Verify that the joint and marginal probability functions of X and Y are as follows:

y \ x	0	1	2	3	$p_Y(y)$
0	1/27	1/9	1/9	1/27	8/27
1	1/9	2/9	1/9	0	4/9
2	1/9	1/9	0	0	2/9
3	1/27	0	0	0	1/27
$p_X(x)$	8/27	4/9	2/9	1/27	1

(b) Derive the probability function of $Z = X + Y$ and verify that $E(Z) = E(X) + E(Y)$. Ans. $p(0) = 1/27$, $p(1) = 2/9$, $p(2) = 4/9$, $p(3) = 8/27$.

(c) Show that $Var(X) = Var(Y) = Var(Z) = 2/3$, so that $Var(X + Y) \neq Var(X) + Var(Y)$ in this case.

2. Suppose X and Y have the joint density function

$$f(x,y) = x + y \quad \text{for} \quad 0 < x < 1,\ 0 < y < 1.$$

(a) Show that $P(X < 1/2,\ Y < 1/2) = 1/8$.

(b) Show that X and Y have the same marginal density function $g(x) = x + 1/2$ for $0 < x < 1$, and $P(X < 1/2) = P(Y < 1/2) = 3/8$. [Note that $P(X < 1/2,\ Y < 1/2) \neq P(X < 1/2)P(Y < 1/2)$.]

(c) It can be shown that, if $Z = X + Y$, then Z has the density function

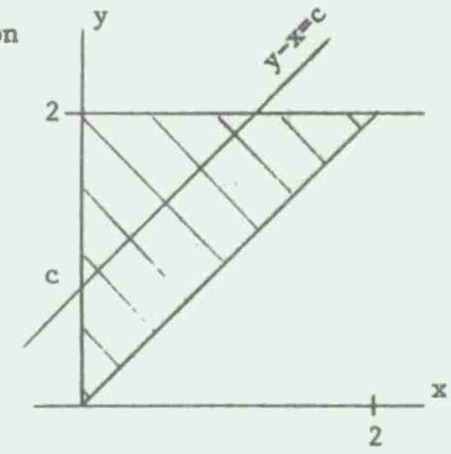
$$h(z) = \begin{cases} z^2 & \text{for } 0 < z < 1 \\ z(2 - z) & \text{for } 1 < z < 2. \end{cases}$$

Show that $E(Z) = 7/6$, $Var(Z) = 5/36$, $E(X) = E(Y) = 7/12$, and $Var(X) = Var(Y) = 11/144$. Thus, $E(X+Y) = E(X) + E(Y)$ but $Var(X+Y) \neq Var(X) + Var(Y)$.

3. Suppose X and Y have joint density function

$$f(x,y) = 1/2 \quad \text{for} \quad 0 < x < y < 2,$$

so that the density function is constant over the shaded region in the figure at the right.



(a) Show that $P(X < 1) = 3/4$.

(b) Show that the marginal density functions of X and Y are $f_X(x) = (2-x)/2$ for $0 < x < 2$ and $f_Y(y) = y/2$ for $0 < y < 2$.

(c) Verify that $Z = Y - X$ has the same density function as X by first noting that $1 - F_Z(c) = P(Y - X > c) = (2 - c)^2/4$ for $0 < c < 2$.

(d) Show that $E(X) = E(Z) = 2/3$ and $E(Y) = 4/3$, verifying that $E(Y - X) = E(Y) - E(X)$.

(e) Show that $Var(X) = Var(Z) = Var(Y) = 2/9$.

The definitions above for the "bivariate" case extend immediately to the "multivariate" case. Let $X_1, X_2, \ldots, X_n$ be n random variables defined on the same sample space. If the random variables X_i are all discrete, then the <u>joint probability function</u> of $X_1,\ldots,X_n$ is defined by

$$p(x_1,x_2,\ldots,x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Whether the random variables are discrete or not, the <u>joint distribution function</u> of $X_1,\ldots,X_n$ is defined by

$$F(x_1,x_2,\ldots,x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n).$$

The random variables are said to have a <u>continuous joint distribution</u> if there is a function f on R^n (called the <u>joint density function</u>) such that

$$P((X_1,X_2,\ldots,X_n) \in B) = \int\int_B \ldots \int f(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$$

for all n-dimensional rectangles $B = (a_1,b_1)\times(a_2,b_2)\times\ldots\times(a_n,b_n)$.

The following theorem is the multivariate analog of Theorems 5-2 and 5-4 in the univariate case. The proof in the discrete case is like that given for Theorem 5-2.

Theorem 7-1. Let $Y = g(X_1,X_2,\ldots,X_n)$ be a random variable such that $E(Y)$ exists. Then

(a) if $X_1, X_2, \ldots, X_n$ are discrete random variables having joint probability function p,

$$E(Y) = \Sigma g(x_1,x_2,\ldots,x_n)p(x_1,x_2,\ldots,x_n)$$

where the summation is over all points $(x_1,x_2,\ldots,x_n)$ for which $p(x_1,x_2,\ldots,x_n) > 0$.

(b) if $X_1,X_2,\ldots,X_n$ have joint density function f,

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,x_2,\ldots,x_n)f(x_1,\ldots,x_n)dx_1\ldots dx_n.$$

Example. A fair die is tossed three times. Let X_i be the result on the i^{th} toss. Then the joint probability function of X_1, X_2, X_3 is $p(x_1,x_2,x_3) = (1/6)^3$ for all (x_1,x_2,x_3), $x_i = 1,2,\ldots,6$. By the theorem above, if $Y = X_1X_2X_3$, then

$$E(Y) = \Sigma x_1x_2x_3/6^3,$$

the summation being over all triples (x_1,x_2,x_3) with $x_i = 1,2,\ldots,6$. But since $\Sigma x_1x_2x_3$ is the expansion of $(1+2+3+4+5+6)^3$, $E(Y) = (21)^3/6^3 = (7/2)^3$. Note that, in this case, $E(X_1X_2X_3) = E(X_1)\cdot E(X_2)\cdot E(X_3)$. It is <u>not</u> true in general that, for any two random variables X and Y, $E(XY) = E(X)\cdot E(Y)$.

Definition. The random variables $X_1, X_2, \ldots, X_n$ are said to be <u>independent</u> if for all Borel subsets $A_1, A_2,\ldots,A_n$ of R^1

$$P(X_1\epsilon A_1, X_2\epsilon A_2, \ldots, X_n\epsilon A_n) = \Pi_{i=1}^{n} P(X_i\epsilon A_i).$$

It can be shown that this relationship holds for all Borel subsets A_i if and only if it holds for all sets A_i of the form $A_i = (-\infty, x]$ for some x.

Hence, $X_1, X_2, \ldots X_n$ are independent if and only if their joint distribution function F is the product of the marginal distribution functions:

$$F(x_1,\ldots,x_n) = F_1(x_1)F_2(x_2)\cdots F_n(x_n) \quad \text{for all } x_1,x_2,\ldots,x_n,$$

where F_i is the distribution function of X_i. In the discrete case, it follows immediately from the definition of independence that the joint probability function must be the product of the marginal probability functions at every point $(x_1,x_2,\ldots,x_n)$:

$$p(x_1,x_2,\ldots,x_n) = p_1(x_1)p_2(x_2)\cdots p_n(x_n).$$

In the continuous case, if $X_1, X_2, \ldots, X_n$ are independent and X_i has the marginal density function f_i, then the joint density can be taken as the product of the marginal density functions:

$$f(x_1,x_2,\ldots,x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n).$$

In a probability model for n independent experiments (see page 25), if X_k depends on the outcome of the k^{th} trial only, then the random variables $X_1, X_2, \ldots, X_n$ are independent, in which case the joint distribution function (or probability function or density function) is the product of the marginal distribution functions (or probability functions or density functions).

Examples.

1. The random variables X_1, X_2, and X_3 in the previous example are independent. Here, the marginal probability function of X_1 is $p_1(x) = 1/6$ for $x=1,2,\ldots,6$.

2. The random variables X and Y having joint density $f(x,y) = e^{-(x+y)}$ for $x > 0$, $y > 0$ are independent. As was shown on page 71, the marginal density functions of X and Y were $f_X(x) = e^{-x}$ for $x > 0$ and $f_Y(y) = e^{-y}$ for $y > 0$. Hence, the joint density is the product of the marginal densities in this case.

3. Consider a sequence of n Bernoulli trials with probability p of success on each trial. Let X_i be 1 or 0 according as the i^{th} trial is a success or not. Then, since the probability function of X_i is

$$p_i(x_i) = p^{x_i}(1-p)^{1-x_i} \quad \text{for } x_i = 0 \text{ or } 1,$$

the joint probability function of $X_1, X_2, \ldots, X_n$ is

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

4. Let $X_1, X_2, \ldots, X_n$ be independent random variables, each having a $N(\mu,\sigma^2)$ distribution. Then the joint density function of $X_1, X_2, \ldots, X_n$ is

$$f(x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$= (2\pi)^{-n/2}\sigma^{-n}\, e^{-\Sigma(x_i-\mu)^2/2\sigma^2}$$

Theorem 7-2. If X and Y are independent random variables, then

(a) so are $U = g(X)$ and $V = h(Y)$,

(b) $E(XY) = E(X)\cdot E(Y)$,

(c) $E[g(X)h(Y)] = E[g(X)]\cdot E[h(Y)]$,

and (d) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$,

provided that the indicated expectations exist.

Proof: (a) $P(U \in A, V \in B) = P(g(X) \in A, h(Y) \in B) = P(X \in g^{-1}(A), Y \in h^{-1}(B))$

$= P(X \in g^{-1}(A))\, P(Y \in h^{-1}(B)) = P(g(X) \in A)P(h(Y) \in B)$

$= P(U \in A)P(V \in B)$. [Note: $g^{-1}(A)$ is defined as $\{x: g(x) \in A\}$.]

(b) In the discrete case, it follows from Theorem 7-1 that

$$E(XY) = \Sigma xy\, p(x,y) = \Sigma xy\, p_X(x)\cdot p_Y(y) = \Sigma x\, p_X(x)\cdot \Sigma y\, p_Y(y)$$

$$= E(X)\cdot E(Y).$$

The proof for the continuous case is similar.

(c) This follows immediately from (a) and (b).

(d) $\mathrm{Var}(X+Y) = E(X + Y - E(X+Y))^2 = E[(X-\mu_X) + (Y-\mu_Y)]^2$

$= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)]$

$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2E[(X - \mu_X)(Y - \mu_Y)].$

By (c), the last term is equal to $E(X - \mu_X)\cdot E(Y - \mu_Y) = 0$.

Note from the proof of (d) above that, in general,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2E[(X - \mu_X)(Y - \mu_Y)].$$

The expectation in the last term on the right, which has value 0 when X and Y are independent, provides a convenient measure of association between two random variables.

Definition. The covariance of two random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

provided the indicated expectations exist. If X and Y have nonzero variances σ_X^2 and σ_Y^2, the correlation coefficient of X and Y, denoted by $\rho(X,Y)$ or just by ρ if no ambiguity results, is defined by

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = E\left[\frac{X-E(X)}{\sigma_X} \cdot \frac{Y-E(Y)}{\sigma_Y}\right].$$

X and Y are said to be uncorrelated if $\mathrm{Cov}(X,Y) = 0$.

Theorem 7-3. Assuming that all the expectations indicated below exist, the following properties hold:

(a) $\mathrm{Cov}(X,Y) = E(XY) - E(X)\cdot E(Y)$

(b) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ and $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

(c) $\mathrm{Cov}(aX + b, cY + d) = ac\cdot\mathrm{Cov}(X,Y)$

(d) $\mathrm{Cov}(\Sigma a_i X_i, Y) = \Sigma a_i \mathrm{Cov}(X_i,Y)$

(e) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

$= \sigma_X^2 + \sigma_Y^2 + 2\rho\,\sigma_X\sigma_Y$

(f) $\mathrm{Var}(X_1+X_2+\ldots+X_n) = \sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\sum_{i<j}\mathrm{Cov}(X_i,X_j)$

(g) If X and Y are independent, $\mathrm{Cov}(X,Y) = \rho(X,Y) = 0$

(h) If $X_1, X_2, \ldots,$ are independent, then $\mathrm{Var}(X_1+X_2+\ldots+X_n) = \sum_{i=1}^{n}\mathrm{Var}(X_i)$

(i) $$\rho(aX+b,\ cY+d) = \begin{cases} \rho(X,Y) & \text{if } ac > 0 \\ -\rho(X,Y) & \text{if } ac < 0 \\ \text{undefined} & \text{if } ac = 0 \end{cases}$$

(j) If $Y = aX + b$ where $a \neq 0$, then $\rho(X,Y) = 1$ if $a > 0$, and $\rho(X,Y) = -1$ if $a < 0$.

The correlation coefficient between two random variables X and Y is a measure of the amount of linear relationship between them. If Y is well approximated (or well "predicted") by a linear function of X, say $a + bX$, then $|\rho|$ is close to 1. Otherwise, ρ is close to zero. This is made precise by the following theorem.

Theorem 7-4. Let X and Y be random variables having nonzero variances.

(a) If α and β are the values of a and b which minimize $S(a,b) = E[Y - (a + bX)]^2$, then $\beta = \rho\sigma_Y/\sigma_X$, $\alpha = E(Y) - \beta E(X)$, and $S(\alpha,\beta) = (1 - \rho^2)\sigma_Y^2$.

(b) $-1 \le \rho(X,Y) \le 1$, and $|\rho| = 1$ if and only if there exist constants α and β such that $P(Y = \alpha + \beta X) = 1$.

Proof: (a) $E[Y - (a+bX)]^2 = E[(Y-\mu_Y) - b(X-\mu_X) - (a-\mu_Y+b\mu_X)]^2$

$= \sigma_Y^2 + b^2\sigma_X^2 + (a-\mu_Y+b\mu_X)^2 - 2b\ \mathrm{Cov}(X,Y)$

$= b^2\sigma_X^2 - 2b\rho\sigma_X\sigma_Y + \rho^2\sigma_Y^2 + (1-\rho^2)\sigma_Y^2 + (a-\mu_Y + b\mu_X)^2$

$= (b\sigma_X - \rho\sigma_Y)^2 + (1-\rho^2)\sigma_Y^2 + (a-\mu_Y + b\mu_X)^2.$

Only the first and third terms depend on a and b, and these can be minimized by setting $\beta = \rho\sigma_Y/\sigma_X$ and $\alpha = \mu_Y - \beta\mu_X$. For these values of a and b, $E[Y - (a+bX)]^2 = (1-\rho^2)\sigma_Y^2$.

Since both $E[Y - (\alpha + \beta X)]^2$ and σ_Y^2 are nonnegative and $\sigma_Y^2 \neq 0$, it follows that $1-\rho^2 \geq 0$, implying that $\rho^2 \leq 1$ or $|\rho| \leq 1$. If $\rho = 1$, then $E[Y - (\alpha + \beta X)]^2 = 0$, which implies that $Y = \alpha + \beta X$, except perhaps on a set of probability zero.

Since $\min_a E(Y - a)^2 = E(Y - \mu_Y)^2 = \sigma_Y^2$ (see Exercise 7, page 52) and $\min_{a,b} E[Y - (a+bX)]^2 = (1 - \rho^2)\sigma_Y^2$, incorporating the random variable X into the "linear predictor" $a + bX$ reduces the lowest attainable mean squared prediction error from σ_Y^2 to $(1 - \rho^2)\sigma_Y^2$. Thus ρ^2 is the proportional reduction in mean squared error that results from including X in the predictor.

The random variable $\alpha + \beta X$ referred to in Theorem 7-4 is sometimes called the best linear predictor of Y based on X. The line $y = \alpha + \beta x$ is called the regression line of Y upon X. This line can be written in the form:

$$\frac{y - E(Y)}{\sigma_Y} = \rho \frac{x - E(X)}{\sigma_X}$$

Examples.

1. A fair coin is tossed 3 times in succession. Let X be 1 or 0 according as the first toss results in heads or not, and let Y be the number of heads on all 3 tosses. Then X and Y have the following joint probability function:

y \ x	0	1	$p_Y(y)$
0	1/8	0	1/8
1	1/4	1/8	3/8
2	1/8	1/4	3/8
3	0	1/8	1/8
$p_X(x)$	1/2	1/2	1

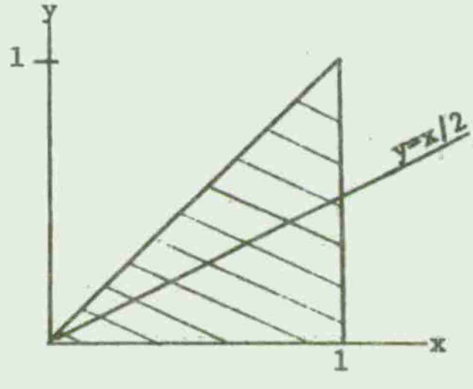
Here, $E(X) = 1/2$, $E(Y) = 3/2$, $Var(X) = 1/4$, and $Var(Y) = 3/4$. Since $E(XY) = 1(1/8) + 2(1/4) + 3(1/8) = 1$, $Cov(X,Y) = E(XY) - E(X)E(Y) = 1 - (1/2)(3/2) = 1/4$, and $\rho(X,Y) = Cov(X,Y)/\sigma_X\sigma_Y = 1/\sqrt{3}$. The regression line of Y upon X is

$$\frac{y - 3/2}{\sqrt{3}/2} = \frac{1}{\sqrt{3}} \left(\frac{x - 1/2}{1/2} \right)$$

which can be written in the form $y = x + 1$. Note the significance of the value of $\rho^2 = 1/3$ in this case.

2. Suppose $Y = X + U$ where X and U are independent. (For example, in Exercise 1, U is the number of heads on the last two tosses.) In this case, $Cov(X,Y) = Cov(X,X + U) = Cov(X,X) + Cov(X,U) = \sigma_X^2$, and $\rho(X,Y) = \sigma_X^2/\sigma_X\sigma_Y = \sigma_X/\sigma_Y$. The best linear predictor of Y based on X turns out to be $X + E(U)$. As a special case, suppose a coin which has probability p of turning up heads is tossed n times. If X is the number of heads on first $m(<n)$ tosses and Y is the number of heads on all n tosses, then $\sigma_X^2 = mpq$ and $\sigma_Y^2 = npq$ so that $\rho = \sigma_X/\sigma_Y = \sqrt{m/n}$. Again note the significance of the value of ρ^2.

3. Suppose X and Y have joint density $f(x,y) = 2$ for $0 < y < x < 1$, so that the density function is constant over the triangular region in the figure at the right. Then



$$E(XY) = \int_0^1 \int_0^x 2xy \, dy \, dx = \int_0^1 x^3 \, dx = 1/4.$$

The marginal density of X in this case is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^x 2 \, dy = 2x \quad \text{for} \quad 0 < x < 1,$$

so that $E(X) = 2/3$ and $Var(X) = 1/18$. The marginal density of Y is

$$f_Y(y) = \int_y^1 2 \, dx = 2(1 - y) \quad \text{for} \quad 0 < y < 1$$

so that $E(Y) = 1/3$ and $Var(Y) = 1/18$. It follows that $Cov(X,Y) = E(XY) - E(X)E(Y) = 1/36$ and $\rho(X,Y) = Cov(X,Y)/\sigma_X\sigma_Y = 1/2$. Thus, the regression

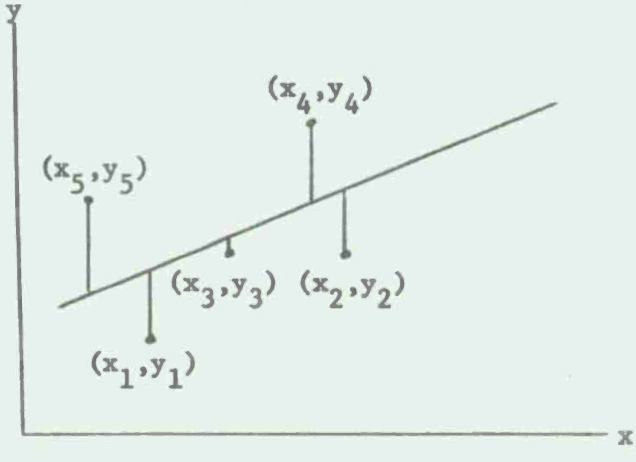
line of Y on X is $y = x/2$.

4. Let (X,Y) have joint probability function $p(x_i,y_i) = 1/n$ where $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$ are any n points on the plane. Since this situation usually arises in a context where the pairs (x_i,y_i) are regarded as being a sample from a larger population, the means and variances of X and Y are called sample means and sample variances in this case, and special notation is introduced: $\bar{x}$ for $E(X)$, s_x^2 for σ_X^2, and r for ρ. Here, $\bar{x} = \Sigma x_i/n$, $s_x^2 = \Sigma(x_i - \bar{x})^2/n = (\Sigma x_i^2/n) - \bar{x}^2$, and similar formulas hold for $\bar{y}$ and s_y^2. Omitting the subscripts i below, we can write $r = Cov(X,Y)/s_x s_y$ where

$$Cov(X,Y) = \Sigma(x-\bar{x})(y-\bar{y})/n = (\Sigma xy/n) - \bar{x}\cdot\bar{y} = [\Sigma xy - n\,\Sigma x\Sigma y]/n,$$

providing a convenient formula for hand calculations.

Since choosing α and β to minimize $E(Y - a - bX)^2 = \Sigma(y_i - a - bx_i)^2/n$ amounts to choosing α and β to minimize $\Sigma(y_i - a - bx_i)^2$ the resulting regression line is called the <u>least squares regression line</u> in this case. Applying the formulas which hold for any regression line, we see that the coefficients of the regression are given by $\beta = rs_y/s_x$ and $\alpha = \bar{y} - \beta\bar{x}$. Since $r = Cov(X,Y)/s_x s_y$, β can be computed using the formula



$$\beta = \frac{Cov(X,Y)}{s_x^2} = \frac{\Sigma(x-\bar{x})(y-\bar{y})/n}{\Sigma(x-\bar{x})^2/n} = \frac{\Sigma xy - (1/n)\Sigma x\Sigma y}{\Sigma x^2 - (1/n)(\Sigma x)^2}$$

For example, the pairs of scores on the left below result from comparing 14 students' diagnostic test scores (x) on a simple algebra test with their final exam scores (y) in a certain statistics course. A plot of the points

and the regression line is given on the next page. Note that the regression line passes through the point $(\bar{x},\bar{y})$.

x	y
36	49
29	47
41	75
34	60
33	55
31	47
43	79
33	60
35	56
23	54
41	65
35	68
37	59
33	65
484	839

$\Sigma x^2 = 17080$, $\Sigma y^2 = 51517$, $\Sigma xy = 29466$

$\bar{x} = \Sigma x/n = 484/14 = 34.57$

$\bar{y} = \Sigma y/n = 839/14 = 59.93$

$\Sigma(x-\bar{x})^2 = \Sigma x^2 - (1/n)(\Sigma x)^2 = 17080 - (484)^2/14 = 347.43$

$\Sigma(y-\bar{y})^2 = \Sigma y^2 - (1/n)(\Sigma y)^2 = 51517 - (839)^2/14 = 1236.93$

$\Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - (1/n)\Sigma x\Sigma y = 29466 - (484)(839)/14 = 460.57$

$s_x^2 = 347.43/14 = 24.82$

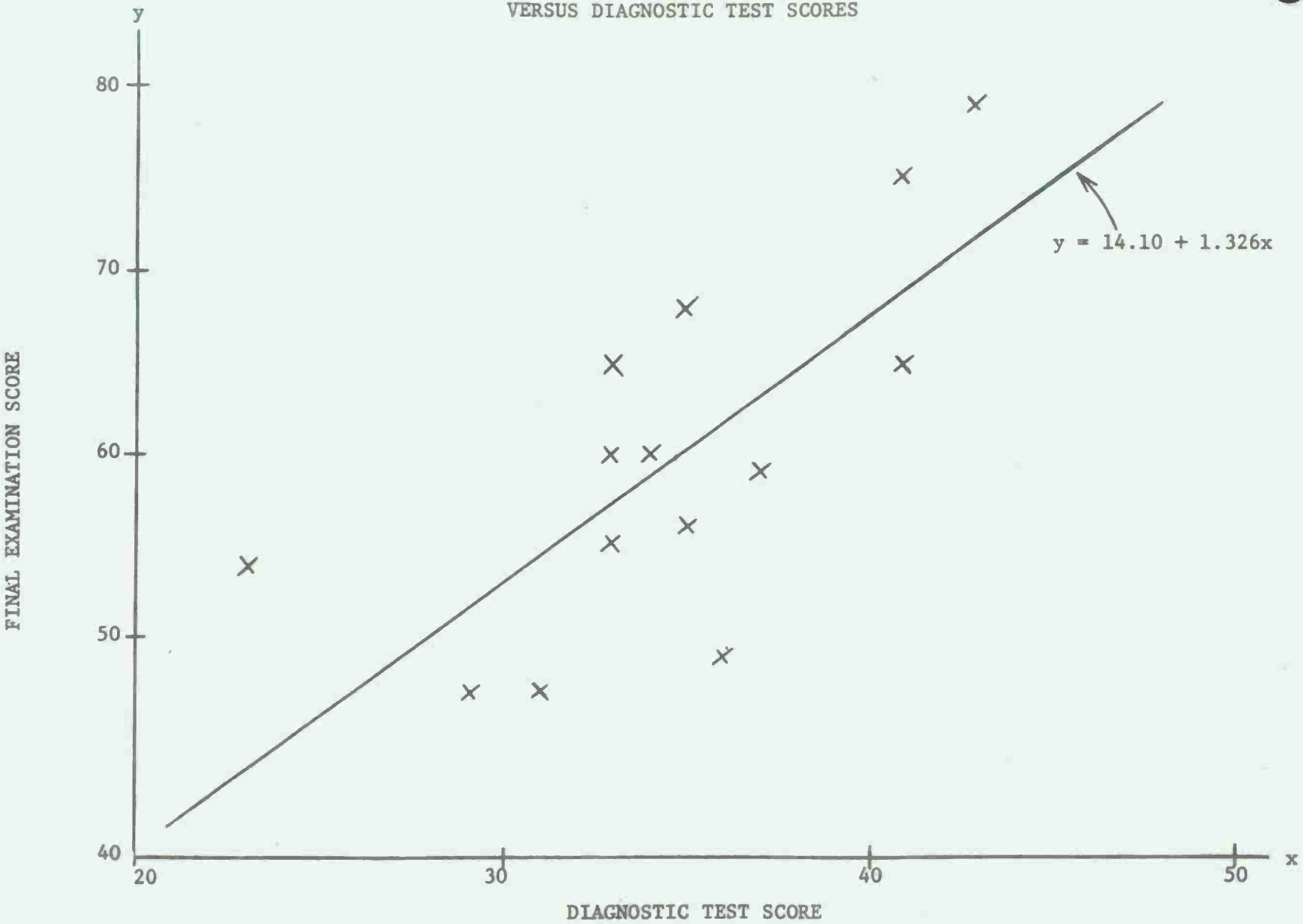
$s_y^2 = 1236.93/14 = 88.35$

$\beta = 460.57/347.43 = 1.326$

$\alpha = 59.93 - 1.326(34.57) = 14.10$

$r = 460.57/\sqrt{(347.43)(1236.93)} = 0.703$

SCATTER DIAGRAM OF FINAL EXAMINATION SCORES
VERSUS DIAGNOSTIC TEST SCORES



Exercises.

1. Let X and Y have the joint probability function given in the example on page 69. (a) Show that $\mathrm{Cov}(X,Y) = 1/2$, $\mathrm{Var}(X) = 1/2$, and $\mathrm{Var}(Y) = 1$, so that $\rho(X,Y) = \sqrt{2}/2$. (b) Show that the regression line of Y on X is $y = x + 1$. (c) Show that the regression line of X on Y is $x = y/2$.

2. Let X and Y have the joint probability function in Exercise 1, page 71. (a) Show that $\mathrm{Cov}(X,Y) = 1/3$ and $\rho(X,Y) = -1/2$. (b) In Exercise 1(c), page 72 you showed that $\mathrm{Var}(X + Y) = \mathrm{Var}(X) = \mathrm{Var}(Y) = 2/3$. Recompute $\mathrm{Var}(X + Y)$ using Theorem 7-3(e). (c) Show that the regression line of Y on X is $y = (3-x)/2$.

3. Let X and Y have the joint density function

$$f(x,y) = x + y \qquad \text{for } 0 < x < 1,\ 0 < y < 1.$$

(See Exercise 2, page 72.) Show that $\mathrm{Cov}(X,Y) = -1/144$, $\rho(X,Y) = -1/11$, and the regression line of Y on X is $y = (7-x)/11$.

4. By Theorem 7-3(g), if X and Y are independent, they are uncorrelated. The converse of this theorem does not hold in general. (a) Show that, if X and Y have joint probability function $p(0,0) = p(-1,1) = p(1,1) = 1/3$, then X and Y are uncorrelated but not independent. (b) Suppose X has a $N(0,1)$ distribution and $Y = X^2$. Show that $\rho(X,Y) = 0$. [Hint: $E(XY) = E(X^3) = 0$ in this case.]

5. Let X and Y have joint density $f(x,y) = 1/2$ for $0 < x < y < 2$ as in Exercise 3, page 73. Show that $\mathrm{Cov}(X,Y) = 1/9$, $\rho(X,Y) = 1/2$, and verify that the regression line of Y on X is $y = (x + 2)/2$.

6. Show that the constant c which minimizes $E(Y - cX)^2$ is $c = E(XY)/E(X^2)$. Use the result to deduce from $E(Y - cX)^2 \geq 0$ that

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}. \quad \text{(Cauchy-Schwarz Inequality.)}$$

Definition. The random variables X and Y are said to have a bivariate normal distribution with parameters ξ, η, σ_x, σ_y, and ρ where $\sigma_x > 0$, $\sigma_y > 0$, and $|\rho| < 1$ if the joint density of X and Y is given by $f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\xi}{\sigma_x})^2 - 2\rho(\frac{x-\xi}{\sigma_x})(\frac{y-\eta}{\sigma_y}) + (\frac{y-\eta}{\sigma_y})^2]\}$. This density function has a maximum at $(x,y) = (\xi,\eta)$, and the coutours of the density function are concentric ellipses centered at (ξ,η). Any plane perpendicular to the (x,y) plane cuts the surface $f(x,y)$ in a curve of the normal form.

Theorem 7-5. If X and Y have the bivariate normal density above, then

(a) $X \sim N(\xi, \sigma_x^2)$, $Y \sim N(\eta,\sigma_y^2)$, and the correlation coefficient of X and Y is ρ;

(b) X and Y are independent if and only if $\rho = 0$;

(c) if $Z = a+bX+cY$ where either $b \neq 0$ or $c \neq 0$, then Z has a normal distribution with mean $a+b\xi+c\eta$ and variance $b^2\sigma_x^2 + c^2\sigma_y^2 + 2bc\rho\sigma_x\sigma_y$.

Proof: (a) The proof that X and Y have the specified marginal distributions follows from the fact that $f(x,y)$ can be written in the form

$$f(x,y) = \frac{1}{\sigma_x}\varphi\left(\frac{x-\xi}{\sigma_x}\right)\frac{1}{\sigma_y\sqrt{1-\rho^2}}\varphi\left(\frac{y-\alpha-\beta x}{\sigma_y\sqrt{1-\rho^2}}\right)$$

where $\beta = \rho\sigma_y/\sigma_x$, $\alpha = \eta - \beta\xi$, and φ is the density function of the standard normal distribution. Integrating out y after a change of variables to $v = (y - \alpha - \beta x)/\sigma_y\sqrt{1-\rho^2}$ yields $f_X(x) = \frac{1}{\sigma_x}\varphi(\frac{x-\xi}{\sigma_x})$, which is the density of a $N(\xi,\sigma_x^2)$ distribution. A similar proof can be used to show that $Y \sim N(\eta,\sigma_y^2)$. The proof that X and Y have correlation coefficient ρ will be given later in this section.

(b) The joint density $f(x,y)$ factors into the marginal densities if and only if $\rho = 0$.

(c) See Alexander M. Mood and Franklin A. Graybill, Introduction to the Theory of Statistics, Second Edition, McGraw-Hill, New York, 1963, p. 211.

Definition. If X and Y have joint probability function $p(x,y)$, then the conditional probability function of Y given $X = x$ is defined by

$$p(y|x) = \frac{p(x,y)}{p_X(x)} \qquad \text{provided } p_X(x) > 0.$$

If X and Y have joint density function $f(x,y)$, then the conditional density function of Y given $X = x$ is defined by

$$f(y|x) = \frac{f(x,y)}{f_X(x)} \qquad \text{provided } f_X(x) > 0.$$

The definition for the discrete case is motivated by the fact that

$$P(Y = y|X = x) = \frac{P(X = x,\ Y = y)}{P(X = x)} \qquad \text{provided that } p_X(x) > 0.$$

The definition for the continuous case is motivated by a consideration of the conditional distribution function of Y, given $X = x$, which can be defined as

$$F(y|x) = \lim_{h \to 0} P(Y \le y | x-h < X < x+h).$$

If X and Y have a continuous joint density function $f(x,y)$, then it can be shown that $F(y|x) = \int_{-\infty}^{y} f(x,y')\, dy' / f_X(x)$. (See H. Cramér, Mathematical Methods of Statistics, Princeton University Press, 1946, p. 268.) Taking the derivative with respect to y yields $f(y|x) = f(x,y)/f_X(x)$. Note that, if X and Y are independent, then the conditional distribution of Y for any value of X is the same as the marginal distribution of Y.

Definition. The conditional distribution of Y, given $X = x$, is the distribution specified by the conditional distribution function $F(y|x)$ defined above [or by $p(y|x)$ or $f(y|x)$ in the discrete or continuous cases]. The conditional expectation of Y given $X = x$, denoted by $E(Y|X = x)$, is defined to be the expectation of the conditional distribution.

The conditional variance, denoted by $\mathrm{Var}(Y|X = x)$, is defined as the variance of the conditional distribution.

In particular, if X and Y have joint probability function $p(x,y)$, then the conditional expectation of Y given $X = x$ is given by

$$E(Y|X = x) = \sum_y y\, p(y|x)$$

for those values of x for which $p_X(x) > 0$ and $\Sigma|y|p(y|x) < \infty$. If X and Y have joint density function $f(x,y)$, then the conditional expectation of Y given $X = x$ is given by

$$E(Y|X = x) = \int_{-\infty}^{\infty} y\, f(y|x)\, dy$$

for those values of x for which $f_X(x) > 0$ and $\int|y|\, f(y|x)\, dy < \infty$.

If X and Y are independent, then the conditional distribution of Y given $X = x$ is the same as the marginal distribution of Y so that $E(Y|X = x) = E(Y)$ and $\mathrm{Var}(Y|X = x) = \mathrm{Var}(Y)$. If Y is some function of X, say $Y = g(X)$, then given that $X = x$, the conditional distribution of Y is entirely concentrated at the point $g(x)$. Hence, in this case, $E(Y|X = x) = E(g(X)|X = x) = g(x)$, and $\mathrm{Var}(Y|X = x) = 0$.

Definition. Assume that $E(Y|X = x)$ exists for all x for which $p_X(x) > 0$ [or $f_X(x) > 0$ in the continuous case]. Then the conditional expectation of Y given X, denoted by $E(Y|X)$, is the random variable having the value $E(Y|X = x)$ when $X = x$.

In particular, if $Y = g(X)$, then $E(Y|X) = g(X)$. If Y and X are independent, then $E(Y|X) = E(Y)$. Other examples will be given below.

Although the definitions above are stated for the case that the conditioning variable X is a random variable, the definitions could just as well have been given for the more general case where X is a "random vector," i.e., $X = (X_1,\ldots,X_n)$ where the random variables X_i are all defined on the same sample space.

Examples.

1. A fair coin is tossed four times. Let X be the number of heads on the first two tosses, and let Y be the number of heads on all four tosses. Then the joint probability function of X and Y is given on page 69. Given $X = 1$, the conditional probability function of Y is: $p(0|1) = 0$, $p(1|1) = 1/4$, $p(2|1) = 1/2$, $p(3|1) = 1/4$, and $p(4|1) = 0$. Thus, $E(Y|X = 1) = 1(1/4) + 2(1/2) + 3(1/4) = 2$. In this case, the random variable $E(Y|X)$ has value 1 when $X = 0$, 2 when $X = 1$, and 3 when $X = 2$, so that $E(Y|X) = 1 + X$.

2. As a generalization of example 1, let Y be the number of successes in $n + m$ Bernoulli trials with probability p of success on each trial, and let X be the number of successes on the first n trials. Then

$$p(x,y) = \binom{n}{x}p^x q^{n-x}\binom{m}{y-x}p^{y-x}q^{m-(y-x)} = \binom{n}{x}\binom{m}{y-x}p^y q^{n+m-y} \quad \text{for}$$

$$x = 0,1,\ldots,n, \quad y = x,x+1,\ldots,x+m.$$

Since X has a binomial distribution with parameters n and p, $p_X(x) = \binom{n}{x}p^x q^{n-x}$, and it follows that the conditional probability function of y for given x is

$$p(y|x) = \binom{m}{y-x}p^{y-x}q^{m-(y-x)} \quad \text{for} \quad y = x,x+1,\ldots,x+m.$$

Therefore the conditional distribution of Y is the same as the distribution of $x + Z$ where Z has a binomial distribution with parameters m and p. It follows that $E(Y|X = x) = x + mp$ and $\text{Var}(Y|X = x) = mpq$. Note that in this case the random variable $E(Y|X) = X + mp$ is the same as the best linear predictor of Y based on X. (See Example 2, page 80.) Theorem 7-6(a) below states the general result that $E[E(Y|X)] = E(Y)$. This can be verified directly in this case as follows:

$$E[E(Y|X)] = E(X) + mp = np + mp = (n + m)\, p = E(Y).$$

3. If X and Y have a bivariate normal distribution, then the marginal distribution of X is $N(\mu_X, \sigma_X^2)$ by Theorem 7-5, and it follows from the representation of the bivariate normal density $f(x,y)$ at the bottom of page VII-18 that the conditional density of Y for $X = x$ is

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sigma_Y\sqrt{1-\rho^2}}\, \varphi\left(\frac{y - \alpha - \beta x}{\sigma_Y\sqrt{1-\rho^2}}\right)$$

where $\beta = \rho\sigma_Y/\sigma_X, \alpha = \mu_Y - \beta\mu_X$, and φ is the standard normal density function. Thus, given $X = x$, the conditional distribution of Y is $N(\alpha + \beta x, \sigma_Y^2(1-\rho^2))$, and the conditional expectation of Y given X is $E(Y|X) = \alpha + \beta X$, which again coincides with the best linear predictor of Y based on X. In this case, the conditional variance of Y given $X = x$ is $\sigma_Y^2(1-\rho^2)$ for all values of x.

Theorem 7-6. Let X and Y be random variables such that the expectations indicated below exist.

(a) $E[E(Y|X)] = E(Y)$.

(b) $E[g(X)|X] = g(X)$.

(c) If X and Y are independent, $E(Y|X) = E(Y)$.

(d) $E[g(X)h(Y)|X] = g(X)\, E[h(Y)|X]$.

(e) For any constants a and b, $E[aY + b|X] = a\, E(Y|X) + b$.

(f) If U and V are random variables having finite expectations, then $E(U + V|X) = E(U|X) + E(V|X)$.

Proof: (a) In the discrete case,

$$E[E(Y|X)] = \sum_x [\sum_y y\, p(y|x)]\, p_X(x) = \sum_x \sum_y y\, p(x,y) = \sum_y y \sum_x p(x,y) = \Sigma y\, p_Y(y) = E(Y).$$

A similar proof can be given in the continuous case.

(b)-(f). Parts (b) and (c) were proved earlier. The proofs of the other parts are omitted.

Theorem 7-4 derived the best linear predictor of Y based on X in the sense of mean squared prediction error and showed that $E(Y - \alpha - \beta X)^2 = (1 - \rho^2)\sigma_Y^2$.

Theorem 7-7. Among all functions $\delta(X)$, $E[Y - \delta(X)]^2$ is minimized by $\delta(X) = E(Y|X)$. The mean square prediction error is given by $E[Y - E(Y|X)]^2 = (1 - \eta^2)\sigma_Y^2$ where $\eta = \rho(Y, E(Y|X))$. [η^2 is called the correlation ratio of Y and X.].

Proof: Let $g(X) = E(Y|X) - \delta(X)$. Then $[Y - \delta(X)]^2 = [Y - E(Y|X) + g(X)]^2 = [Y - E(Y|X)]^2 + 2g(X)[Y - E(Y|X)] + [g(X)]^2$. Since the next to the last term on the right has expectation zero by parts (a) and (d) of the previous theorem,

$$E[Y - \delta(X)]^2 = E[Y - E(Y|X)]^2 + E[g(X)]^2 \geq E[Y - E(Y|X)]^2.$$

The fact that $E[Y - E(Y|X)]^2 = (1 - \eta^2)\sigma_Y^2$ follows immediately from Theorem 7-4(a) by observing that the best linear predictor of Y based on $E(Y|X)$ is $E(Y|X)$.

Since the mean squared prediction error using the best linear function of X is $E[Y - \alpha - \beta X]^2 = (1 - \rho^2)\sigma_Y^2$ where ρ is the correlation coefficient of X and Y, it follows that $1 - \rho^2 \geq 1 - \eta^2$, implying that $\rho^2 \leq \eta^2$. If it happens that $E(Y|X)$ is linear in X, then $\rho^2 = \eta^2$ and $E(Y|X) = \alpha + \beta X$ where $\beta = \rho\sigma_Y/\sigma_X$ and $\alpha = \mu_Y - \beta\mu_X$. In particular, if X and Y have a bivariate normal distribution, then it was shown on page 89 that $E(Y|X) = \alpha + \beta X$ where $\beta = \rho\sigma_Y/\sigma_X$. This proves a result that was stated but not proved earlier--namely, that the parameter ρ in the bivariate normal density function is the correlation coefficient of X and Y.

The following example illustrates how the above theory is sometimes applied to estimate parameters of distributions in those instances where the parameters themselves can be considered to be random variables.

Example. You can observe a sequence of n Bernoulli trials with probability y of success on each trial. Suppose that the value of y is unknown, and you want to guess y based on the number of successes, X, in n trials. In the absence of any information on the values of y, you might guess y using the "estimator" X/n. This estimator has expectation $E(X/n) = y$ and variance $Var(X/n) = Var(X)/n^2 = ny(1-y)/n^2 = y(1-y)/n$.

Now suppose that you are informed (or are willing to assume) that the value of y was randomly generated according to a distribution having density function $f(y)$ on $(0,1)$. That is, y can be regarded as the value of a random variable Y having density function $f(y)$, and the conditional probability function of X given $Y = y$ is

$$p(x|y) = \binom{n}{x} y^x (1-y)^{n-x} \quad \text{for} \quad x = 0,1,\ldots,n.$$

Guessing the value of Y based on X amounts to "predicting" Y using some function of X. If the mean squared prediction error is an appropriate goodness criterion for your estimator, then the theory above suggests using $E(Y|X)$ to estimate y. Here, X is discrete and Y is continuous, so that the joint distribution of X and Y is neither discrete nor continuous. However, using the fact that

$$P(X \in A,\ a < Y < b) = \sum_{x \in A} \int_a^b p(x|y)\, f(y)\, dy,$$

one can show that the conditional density of Y given $X = x$ is

$$f(y|x) = \frac{p(x|y)f(y)}{p_X(x)} = \frac{y^x(1-y)^{n-x}\, f(y)}{\int_0^1 y^x(1-y)^{n-x}\, f(y)\, dy}.$$

Thus, given the density function $f(y)$, one can compute the conditional expectation of Y for any value of X. In particular, if $f(y) = 1$ for $0 < y < 1$, then

$E(Y|X = x) = \int_0^1 y\, f(y|x)\, dy = \int_0^1 y^{x+1} (1-y)^{n-x}\, dy / \int_0^1 y^x (1-y)^{n-x}\, dy.$

Since $\int_0^1 y^\alpha (1-y)^\beta\, dy = \alpha!\, \beta!/(\alpha + \beta + 1)!$ for $\alpha = 0,1,\ldots,\ \beta = 0,1,\ldots$ (see Alexander M. Mood and Franklin A. Graybill, Introduction to the Theory of Statistics, Second Edition, McGraw-Hill, New York, 1963, pp. 129-131), it follows that

$$E(Y|X = x) = \frac{(x+1)!\,(n-x)!}{(n+2)!} \cdot \frac{(n+1)!}{x!\,(n-x)!} = \frac{x+1}{n+2}$$

Thus, if Y is chosen according to a uniform distribution on $(0,1)$, then the estimator $E(Y|X) = (X+1)/(n+2)$ has the smallest mean squared prediction error among all functions of X.

Exercises.

1. Suppose X and Y have joint density function $f(x,y) = 2$ for $0 < y < x < 1$. (See Example 3, page 80.) Show that (a) given $X = x$ the conditional distribution of Y is a uniform distribution on $(0,x)$, (b) $E(Y|X) = X/2$, and (c) $E(X|Y) = (1+Y)/2$. Verify directly that $E[E(X|Y)] = E(X)$.

2. Let X and Y have the joint probability function given at the top of page 72. Verify that (a) $E(Y|X = 0) = 3/2$, (b) $E(Y|X) = (3-X)/2$.

3. Suppose X has a uniform distribution on $(-1,1)$, and $Y = X^2$. Show that (a) X and Y are uncorrelated, (b) the regression line of Y on X is $y = 1/3$, (c) the correlation ratio between X and Y is 1.

4. If the conditional variance of Y given X is defined by

$$\mathrm{Var}(Y|X) = E([Y - E(Y|X)]^2|X),$$

show that

(a) $\mathrm{Var}(Y|X) = E(Y^2|X) - [E(Y|X)]^2$

(b) $\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E(Y|X)]$.

5. Show that if X and Y have joint density $f(x,y) = e^{-y}$ for $0 < x < y < \infty$, then $E(Y|X) = X + 1$, $E(X|Y) = Y/2$, and $\rho(X,Y) = \sqrt{2}/2$.

SECTION VIII. - SOME SAMPLING THEORY

Reference:

Paul L. Meyer, Introductory Probability and Statistical Applications, 2nd Edition, Addison-Wesley, 1960, Chapters 12 and 13.

By a random sample of size n from a population having distribution function F is meant a sequence of n i.i.d. (independent and identically distributed) random variables $X_1, X_2, \ldots, X_n$, each having distribution function F. Such a sample might result from choosing an element at random from some population, observing the value X_1 of some characteristic of the element, replacing the element, choosing a second element, observing the value X_2 of the same characteristic of the second element, and so forth. Alternatively, the sequence $X_1, X_2, \ldots, X_n$ may result from observing n independent trials of the same type. For example, X_i might be the sum of the results on the ith trial when two dice are tossed repeatedly. Or $X_1, X_2, \ldots, X_n$ might be the waiting times between n successive telephone calls coming into an exchange.

In most statistical applications the distribution of the X_i's is unknown and one attempts to make inferences about the distribution based upon the values of the observations $X_1, X_2, \ldots, X_n$. For example, one might want to estimate the mean or standard deviation of the distribution from which the X_i's are drawn. Quite often the statistics commonly used in drawing such inferences involve sums or averages of the X_i's or functions of the X_i's.

Theorem 8-1. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, each having mean μ and finite variance σ^2.

(a) If $\bar{X} = (X_1+X_2+\ldots+X_n)/n$, then $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2/n$.

(b) (Law of Large Numbers) For any $\epsilon > 0$, $P(|\bar{X} - \mu| \geq \epsilon)$ tends to 0 as n becomes infinite.

Proof: (a) $E(\bar{X}) = (1/n)\Sigma E(X_i) = n\mu/n = \mu$.

$$Var(\bar{X}) = (1/n^2)\Sigma Var(X_i) = n\sigma^2/n^2 = \sigma^2/n.$$

(b) By Chebyshev's Inequality (see page 50).

$$P(|\bar{X} - \mu| \geq \epsilon) \leq Var(\bar{X})/\epsilon^2 = \sigma^2/n\epsilon^2.$$

The last member on the right tends to 0 as $n \to \infty$.

Since $Var(\bar{X}) = \sigma^2/n$, as n increases the distribution of $\bar{X}$ becomes more and more concentrated about $E(\bar{X}) = \mu$. If, instead of considering the average of $X_1, X_2, \ldots, X_n$, one considers the sum $S_n = X_1+X_2+\ldots+X_n$, then $E(S_n) = n\mu$ and $Var(S_n) = n\sigma^2$ so that if $\sigma^2 > 0$ the distribution of S_n becomes increasingly spread out as n increases

Consider the "standardized" variable $(S_n - n\mu)/\sigma\sqrt{n}$. This random variable has mean 0 and variance 1 for all values of n. The theorem below states that, no matter what the initial distribution of the X_i's is, the distribution function of $(S_n - n\mu)/\sigma\sqrt{n}$ tends to the distribution function of a standard normal distribution.

Theorem 8-2. (Central Limit Theorem.) Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and finite variance $\sigma^2 > 0$, and let $S_n = X_1+X_2+\ldots+X_n$. For any constants a and b with $-\infty \leq a < b \leq \infty$,

$$\lim_{n\to\infty} P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) = \Phi(b) - \Phi(a)$$

where Φ is the distribution function of a standard normal distribution.

Proof: See Meyer, op. cit., pp. 252-253.

This is a generalization of the DeMoive-LaPlace Central Limit Theorem stated earlier in Section VI for the case where the X_i's have Bernoulli distributions. The theorem suggests that for "large" values of n one can approximate probabilities of the form $P(S_n \le k)$ as follows:

$$P(S_n \le k) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{k - n\mu}{\sigma\sqrt{n}} \right) \doteq \Phi\left(\frac{k - n\mu}{\sigma\sqrt{n}} \right).$$

Depending on the distribution of the X_i's, this "normal approximation" for sums of i.i.d. random variables is usually quite good even for relatively small values of n (say, $n = 25$ if the X_i's have Bernoulli distributions with p close to 1/2, and $n = 10$ if the X_i's have uniform or exponential distributions). If the X_i's have normal distributions, the approximation is exact because in this case it can be shown that S_n also has a normal distribution. (See Theorem 8-6 below.)

Note that, since

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

the Central Limit Theorem could just as well have stated that the average of n i.i.d. random variables has a limiting normal distribution as n becomes infinite.

Example. Suppose that light bulbs have lifetimes in hours that can be assumed to have a distribution with mean 1000 and standard deviation 500. Find the probability that the average lifetime of 100 such lightbulbs will be greater than 1100 hours.

Let $X_1, X_2, \ldots, X_{100}$ be the lifetimes in hours, and let $\bar{X} = \Sigma X_i/100$. Assuming that the X_i's are a random sample from a distribution having mean 1000 and standard deviation 500, it follows that $E(\bar{X}) = 1000$ and $\sigma_{\bar{X}} = 500/\sqrt{100} = 50$.

Hence,

$$P(\bar{X} > 1100) = 1 - P(\frac{\bar{X} - 1000}{50} \leq \frac{1100 - 1000}{50}) \doteq 1 - \Phi(2.0) = 0.023.$$

Exercises.

1. Suppose that 10 storage batteries $B_1, B_2, \ldots, B_{10}$ are used in the following way. First, B_1 is used until it fails, at which time it is replaced by B_2. Then, when B_2 fails, it is replaced by B_3, etc. If these batteries are chosen at random from a population having mean lifetime 12 hours and variance 2.5 hours, what is the approximate probability that the total time of operation of the batteries will exceed 110 hours? Ans. 0.98.

2. Suppose 100 random digits are generated. That is, 100 independent trials are conducted in which one of the digits 0,1,2,...,9 is chosen at random. Approximate the probability that (a) the digit 0 occurs more than 15 times among the 100 random digits, (b) the sum of the 100 digits exceeds 500, (c) the average of the 100 digits lies between 4.0 and 5.0. Ans. 0.03, 0.04, 0.92.

3. Suppose that, when the heights of 300 plants are measured to the nearest inch, the rounding errors are independent and uniformly distributed over (-0.5, 0.5). If the 300 heights are averaged after rounding, what is the probability that the magnitude of the total error due to rounding exceeds 0.02? Ans. 0.23.

4. In pari-mutuel wagering, the racetrack (or gambling house) takes a fixed percentage of the total amount bet and returns the rest to those who have bet on the winning horse. For example, suppose that the total amount bet on a certain race is $6000, of which $2000 is bet on Horse #1, including your $2 bet. If the track "take" is $1000, then the remaining $5000 is divided up among the holders of winning tickets on Horse #1. Thus

the "betting odds" on Horse #1 are said to be "5-to-2"--i.e., a \$2 bet will yield a return of \$5 for a net gain of \$3. Perhaps a reasonable assessment of the probability that Horse #1 will win is the proportion of the total amount of the money that is bet on Horse #1, which is 1/3 in this case. If you repeatedly play a game like this, on each play you either win \$3 with probability 1/3 or lose \$2 with probability 2/3, (a) find the expectation and variance of your "net gain" after 18 plays of the game, and (b) find the approximate probability that you will be ahead after 18 plays.

Ans. -6, 100, 0.23.

Theorems 8-1 and 8-2 above are usually applied in situations where the random variables $X_1, X_2, \ldots, X_n$ are the values of the observations themselves. However, the theorems apply equally well to transformations of the observations in the following sense.

Theorem 8-3. Let $Y_i = g(X_i)$ where $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, and let $\overline{Y} = \Sigma Y_i/n$ and $T_n = \Sigma Y_i$. If Y_i has mean $\eta = E[g(X)]$ and variance $\tau^2 < \infty$, then

(a) $E(\overline{Y}) = \eta$ and $Var(\overline{Y}) = \tau^2/n$,

(b) for any $\epsilon > 0$, $P(|\overline{Y} - \eta| \geq \epsilon) \to 0$ as $n \to \infty$,

(c) for any constants a and b with $-\infty \leq a < b \leq \infty$

$$\lim_{n\to\infty} P\left(a < \frac{T_n - n\eta}{\tau\sqrt{n}} < b\right) = \Phi(b) - \Phi(a).$$

Proof: Since the random variables $Y_1, Y_2, \ldots, Y_n$ are i.i.d. [the independence of the Y_i's is an obvious generalization of Theorem 7-2(a)], these results follow immediately from Theorems 8-1 and 8-2.

The above theorems are of fundamental importance in statistics. Given a random sample $X_1, X_2, \ldots, X_n$ from a distribution having unknown mean

μ and variance σ^2, one can estimate μ using the value of the estimator $\bar{X}$. That is, if the observed values of $X_1, X_2, \ldots, X_n$ are $x_1, x_2, \ldots, x_n$, then the estimated value of μ for that particular sample is $\bar{x} = \Sigma x_i/n$, the observed value of $\bar{X}$. The goodness of an estimator is usually measured by the extent to which the distribution of the estimator is concentrated around the parameter being estimated. As we shall see later, there may be other estimators that are better than $\bar{X}$ in particular instances, but $\bar{X}$ has certain appealing properties. By Theorem 8-1, the distribution of $\bar{X}$ is "centered" at μ in the sense that $E(\bar{X}) = \mu$ for all values of μ. Also the variance of $\bar{X}$ is only $1/n$ times as large as the variance of each of the original X_i's. If n is large enough, $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$ so that, if σ^2 is known, one can use the normal approximation to approximate the probability that $\bar{X}$ will deviate from μ by more than any prespecified amount $\epsilon > 0$:

$$P(|\bar{X} - \mu| > \epsilon) = P\left(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} > \frac{\epsilon}{\sigma/\sqrt{n}}\right) \approx 2\Phi\left(\frac{-\epsilon}{\sigma/\sqrt{n}}\right).$$

If σ^2 is unknown, one can derive an estimator of σ^2 using the fact that $\sigma^2 = E(X^2) - \mu^2$ where X has the same distribution as the X_i's. By Theorem 8-3, $\Sigma X_i^2/n$ has expectation $E(X^2)$. Therefore, one can estimate σ^2 using

$$\hat{\sigma}^2 = (\Sigma X_i^2/n) - \hat{\mu}^2$$

where $\hat{\mu}$ is some estimator of μ. If one uses $\hat{\mu} = \bar{X}$, the resulting estimator is the "sample variance"

$$S^2 = (\Sigma X_i^2/n) - \bar{X}^2 = \Sigma(X_i - \bar{X})^2/n.$$

However, S^2 is a "biased" estimator of σ^2 [i.e., $E(S^2) \neq \sigma^2$] because $E(\bar{X}^2) = \mathrm{Var}(\bar{X}) + E^2(\bar{X}) = (\sigma^2/n) + \mu^2$, implying that

$$E(S^2) = E(X^2) - (\sigma^2/n) - \mu^2 = \sigma^2 - \sigma^2/n = \sigma^2(n-1)/n.$$

To obtain an unbiased estimator, we can multiply S^2 by $n/(n-1)$, yielding the following alternative estimator:

$$\hat{\sigma}^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1} = \frac{\Sigma X_i^2 - n\bar{X}^2}{n-1}$$

Theorem 8-4. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having mean μ and variance σ^2.

(a) The sample variance $S^2 = \Sigma(X_i - \bar{X})^2/n$ has expectation $E(S^2) = \sigma^2(n-1)/n$.

(b) An unbiased estimator of σ^2 is $\hat{\sigma}^2 = \Sigma(X_i-\bar{X})^2/(n-1)$.

(c) An unbiased estimator of $\mathrm{Var}(\bar{X})$ is $\hat{\sigma}^2/n$.

Part (c) above enables us to attach a measure of reliability to $\bar{X}$ as an estimator of μ even if σ^2 is unknown. If σ^2 is known, the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$. If σ^2 is unknown, the variance of $\bar{X}$ can be estimated by σ^2/n (or S^2/n). The square root of the estimated variance (called the standard error of $\bar{X}$) is an estimate of the standard deviation of $\bar{X}$.

The reader should not infer from the above that the estimators $\bar{X}$ and $\hat{\sigma}^2$ are necessarily good estimators in all circumstances. Nor is it the case that the unbiased estimator $\hat{\sigma}^2$ is necessarily preferable to the biased estimator S^2. In the next section, examples will be given to indicate that both $\bar{X}$ and $\hat{\sigma}^2$ can often be improved upon, depending on the nature of the distribution from which the random sample is taken. Also, S^2 is a better estimator than $\hat{\sigma}^2$ according to a certain goodness criteria that will be introduced later.

Definition. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population having distribution function F. The order statistics corresponding to the random sample are defined by

$X_{(1)} = \min(X_1, X_2, \ldots, X_n)$,

$X_{(2)}$ = next largest of the X_i's,

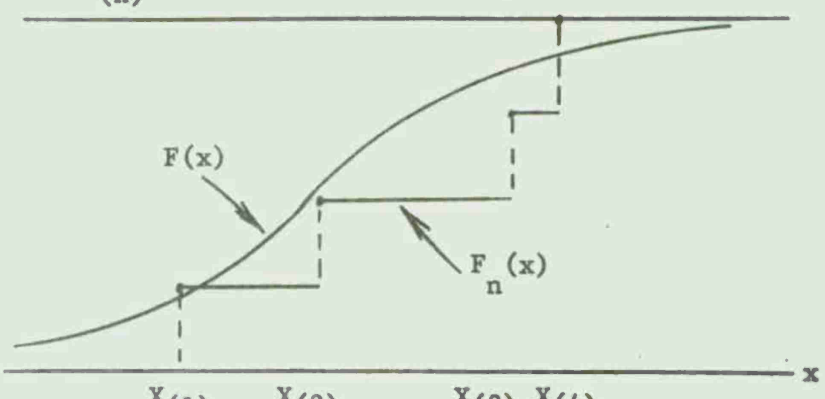
. . .

$X_{(n)} = \max(X_1, X_2, \ldots, X_n)$.

The sample range is defined by $R = X_{(n)} - X_{(1)}$, and the sample median by $X_{[(n+1)/2]}$ if n is odd and $(1/2)[X_{(n/2)} + X_{(n/2+1)}]$ if n is even. The sample (empirical) distribution function of $X_1, X_2, \ldots, X_n$ is defined for all x by

$$F_n(x) = \frac{\text{number of } X_i\text{'s having value } \leq x}{n}.$$

For given values of the X_i's, the sample c.d.f. (cumulative distribution function) is a step function having jumps of size $1/n$ at the values $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The figure below depicts the case $n = 4$.



Note that F_n is a discrete distribution function that varies from sample to sample. Let χ be a random variable having this (conditional) distribution function. Then the conditional expectation of χ given the sample random variables $X_1, X_2, \ldots, X_n$ is $\Sigma X_{(i)}/n = \Sigma X_i/n = \bar{X}$, and the conditional variance of χ is the sample variance $\Sigma(X_i - \bar{X})^2/n$.

Just as the sample mean $\bar{X}$ and the sample variance S^2 can be considered as estimators of the population mean μ and variance σ^2, the sample c.d.f. can be considered as an estimator of the population

distribution function. The following theorem shows that F_n is an unbiased estimator of F, and as n becomes infinite, F_n tends to F for all values of x.

Theorem 8-5. Let F_n be the empirical distribution function of a random sample $X_1, X_2, \ldots, X_n$ from a population having distribution function F. Then

(a) $E[F_n(x)] = F(x)$ for all x,

(b) $\text{Var}[F_n(x)] = F(x)[1 - F(x)]/n$,

(c) $P(|F_n(x) - F(x)| \geq \epsilon) \to 0$ as $n \to \infty$ for all values of x and any $\epsilon > 0$.

Proof: Note that $F_n(x) = \overline{Y}$ where $Y_i = 1$ or 0 according to whether $X_i \leq x$ or $X_i > x$. Here Y_i has a Bernoulli distribution with parameter $p = P(X_i \leq x) = F(x)$. By Theorem 8-3, $E[F_n(x)] = E(\overline{Y}) = F(x)$ and $\text{Var}[F_n(x)] = F(x)[1-F(x)]/n$, which tends to zero as $n \to \infty$.

Exercises.

1. A random sample of size 25 is taken with the result that $\Sigma x_i = 50$ and $\Sigma x_i^2 = 200$. Compute the values of (a) $\overline{X}$, (b) S^2, (c) $\hat{\sigma}^2$, (d) $\hat{\sigma}/\sqrt{n}$. Ans. (a) 2, (b) 4, (c) 25/6, (d) 0.41.

2. Show that if $X_1, X_2, \ldots, X_n$ are independent, Bernoulli random variables with parameter p, then the formula for S^2 in this case reduces to $S^2 = \overline{X}(1 - \overline{X})$, and the standard error of $\overline{X}$ is $\hat{\sigma}/\sqrt{n} = \sqrt{\overline{X}(1 - \overline{X})/(n-1)}$.

3. Let $X_1, X_2, \ldots, X_{100}$ be a random sample of 100 IQ scores from a normal distribution having unknown mean μ but a known standard deviation $\sigma = 16$. In this case $\overline{X}$ has a normal distribution by Theorem 8-6 below.

(a) Compute $P(|\overline{X} - \mu| \leq 2)$. (b) Suppose you can choose a larger sample size to increase the reliability of $\overline{X}$. How large a sample would you need to assure that $P(|\overline{X} - \mu| \leq 2) \geq 0.95$? Ans. (a) 0.79, (b) 246.

4. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots Y_n$ be two independent random samples such that $E(X_i) = \xi$, $Var(X_i) = \sigma^2$, $E(Y_j) = \eta$, and $Var(Y_j) = \tau^2$. Such a model might arise if the Y_j's correspond to the responses of n individuals who had received a special treatment of some kind, and the X_i's are the corresponding responses for individuals in the control group. Let $\delta = \eta - \xi$ be the average effect of the treatment.

(a) Show that an unbiased estimator of the treatment effect δ is $\hat{\delta} = \overline{Y} - \overline{X}$, and its variance is $Var(\hat{\delta}) = (\tau^2/n) + (\sigma^2/m)$.

(b) Show that an unbiased estimator of $Var(\hat{\delta})$ is $(\hat{\tau}^2/n) + (\hat{\sigma}^2/m)$ where $\hat{\sigma}^2 = \Sigma(X_i - \overline{X})^2/(m-1)$ and $\hat{\tau}^2 = \Sigma(Y_j - \overline{Y})^2/(n-1)$.

(c) If $\sigma^2 = \tau^2$, show that an unbiased estimator of σ^2 is the "pooled" estimator

$$\hat{\sigma}^2 = \frac{\Sigma(X_i - \overline{X})^2 + \Sigma(Y_j - \overline{Y})^2}{n+m-2} .$$

Let $X_1, X_2, \ldots, X_n$ be n random variables defined on the same sample space. For certain statistical applications, it is necessary to derive the exact distributions of certain functions of the X_i's, such as $\overline{X}$, $\Sigma a_i X_i$, $\max(X_1, X_2, \ldots, X_n)$, etc. There are certain standard techniques for deriving such distributions that are treated in most statistics texts. (See, for example, Robert V. Hogg and Allen T. Craig, Introduction to Mathematical Statistics, Second Edition, The Macmillan Company, New York, Chapter 4.) You have already used one general technique several times in deriving density functions of transformed variables by first finding their distribution functions and then taking derivatives. With only a few exceptions below, we shall not need the other standard techniques for the distribution theory in this course, and appropriate references will be cited when results are given without proof.

The following theorem states some results about the exact distributions of sums of random variables. Many of these results are somewhat obvious from the discussion of the models in Section VI. For notational convenience below we shall use abbreviations such as "$X \sim$ Binomial(n,p)" to denote "X has a binomial distribution with parameters n and p."

Theorem 8-6. In each of the following, assume that $X_1, X_2, \ldots, X_n$ are independent random variables.

(a) If $X_i \sim \text{Binomial}(n_i,p)$, then $\Sigma X_i \sim \text{Binomial}(\Sigma n_i,p)$.

(b) If $X_i \sim \text{Poisson}(\lambda_i)$, then $\Sigma X_i \sim \text{Poisson}(\Sigma\lambda_i)$.

(c) If $X_i \sim \text{Geometric}(p)$, then $\Sigma X_i \sim \text{Negative Binomial}(n,p)$.

(d) If $X_i \sim \text{Negative Binomial}(r_i,p)$, then $\Sigma X_i \sim \text{Negative Binomial}(\Sigma r_i,p)$.

(e) If $X_i \sim N(\mu_i,\sigma_i^2)$, then $\Sigma a_i X_i \sim N(\Sigma a_i\mu_i,\Sigma a_i^2\sigma_i^2)$.

(f) If $X_i \sim \text{Gamma}(r_i,\lambda)$, then $\Sigma X_i \sim \text{Gamma}(\Sigma r_i,\lambda)$.

(g) If $X_i \sim \text{Cauchy}(\mu_i,\lambda_i)$ then $\Sigma a_i X_i \sim \text{Cauchy}(\Sigma a_i\mu_i,\Sigma a_i\lambda_i)$.

Proof:

(a) Consider a sequence of Bernoulli trials with probability p of success on each trial. Let X_1 be the number of successes on the first n_1 trials, X_2 the number of successes on the next n_2 trials, and so forth. Then $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim \text{Binomial}(n_i,p)$. Since ΣX_i is the total number of successes on all Σn_i trials, ΣX_i has a binomial distribution with parameters Σn_i and p.

(b) Consider $Y = X_1 + X_2$ where X_1 and X_2 are independent, $X_1 \sim \text{Poisson}(\lambda)$, $X_2 \sim \text{Poisson}(\mu)$. It suffices to show that $Y \sim \text{Poisson}(\lambda+\mu)$, since the result for the sum of n random variables then follows by mathematical induction. For $y = 0,1,2,\ldots$

$$P(Y = y) = \Sigma_{x=0}^{y} P(X_1 = x, X_2 = y-x) = \Sigma_{x=0}^{y} \frac{e^{-\lambda}\lambda^x}{x!} \frac{e^{-\mu}\mu^{y-x}}{(y-x)!}$$

$$= \frac{e^{-(\lambda+\mu)}}{y!} \Sigma_{x=0}^{y} \frac{y!}{x!\,(y-x)!} \lambda^x \mu^{y-x} = \frac{e^{-(\lambda+\mu)}}{y!} (\lambda+\mu)^y$$

(c)-(d) These can be proved as in (b) above.

(e)-(g) In general, if $T = U + V$ where U and V are independent random variables having density functions $g(u)$ and $h(v)$ then T has density

$$f(t) = \int_{-\infty}^{\infty} g(u)\, h(t-u)\, du,$$

because the distribution function of T is

$$F(t) = P(T \le t) = P(U + V \le t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-u} g(u)\, h(v)\, dv\, du,$$

and the derivative of the double integral on the right is $f(t)$. With this simplification, the derivation of parts (e)-(g) is straightforward but tedious, and the proofs are omitted. The proof of (e) for the case $n = 2$ is a special case of Theorem 7-5(c), which states that linear functions of random variables having a bivariate normal distribution have normal distributions.

Note that by part (f) of the theorem that if $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim \text{Cauchy}(\mu,\lambda)$, then $\bar{X}$ has exactly the same distribution as each of the individual X_i's. Hence, in this case, the distribution of $\bar{X}$ does not become more and more concentrated as n increases nor does the distribution of $\bar{X}$ become increasingly normal as $n \to \infty$. Why is this not a counter-example to Theorems 8-1 and 8-2?

Definition. A random variable X is said to have a chi-square(χ^2) distribution with n degrees of freedom [abbreviated $X \sim \chi^2(n)$] if X has the same distribution as $\Sigma_{i=1}^{n} Z_i^2$ where $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables.

Random variables having chi-square distributions occur frequently in statistical applications. In particular, in sampling from a $N(\mu,\sigma^2)$ distribution, the estimators S^2 and $\hat{\sigma}^2$ introduced earlier in this section are both multiples of chi-square distributed random variables. This application will be discussed later in this section. The reason for calling the parameter n the number of "degrees of freedom" will become clear later. For now the student should ignore this peculiar terminology and merely regard the parameter n as the number of terms in the sum ΣZ_i^2.

Theorem 8-7. If $X \sim \chi^2(n)$, then

(a) X has a gamma distribution with parameters $r = n/2$ and $\lambda = 1/2$,

(b) $E(X) = n$, $Var(X) = 2n$.

Proof:

(a) If $n = 1$, the distribution function of X for $x > 0$ is

$$F(x) = P(X \le x) = P(Z^2 \le x) = P(-\sqrt{x} < Z < \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1.$$

Therefore, the density of X is

$$f(x) = F'(x) = 2\varphi(\sqrt{x})/(2\sqrt{x}) = (1/\sqrt{2\pi x})\, e^{-x/2} \qquad \text{for } x > 0.$$

Comparing this with the density of a Gamma(1/2,1/2) distribution (see Table 1, Section VI) and recalling that $\Gamma(1/2) = \sqrt{\pi}$, we see that $X \sim$ Gamma(1/2,1/2). It follows from Theorem 8-6(f) that $\Sigma Z_i^2 \sim$ Gamma(n/2,1/2).

(b) This follows from the fact that the expectation and variance of a Gamma(r,λ) distribution are r/λ and r/λ^2. (See Exercise 1, page 64.)

The figure on the next page shows the graphs of the chi-square density functions for 1, 2, 4, and 6 degrees of freedom. As the number of degrees of freedom increases, the density function becomes more symmetric about its mean. Since the chi-square distribution is the distribution of a sum of i.i.d. random variables, it has a limiting normal distribution by the Central Limit Theorem. The normal approximation to the chi-square distribution becomes quite good for $n \geq 20$.

Table 1 on the following page gives the values of x for which the distribution function $F(x)$ of a chi-square distribution has certain specified values. Suppose $n = 20$. Then the entry 31.4 in the 20th row under the column headed .950 means that if $X \sim \chi^2(20)$, then $P(X < 31.4) = 0.95$. An equivalent way of saying the same thing is to say that 31.4 is the 95th percentile (or percentage point) of a chi-square distribution with 20 degrees of freedom.

How well does the normal approximation work in this case? Since $E(X) = 20$ and $\mathrm{Var}(X) = 40$, the normal approximation of $P(X < 31.4)$ is given by

$$P(X < 31.4) \doteq \Phi\left(\frac{31.4 - 20}{\sqrt{40}}\right) = \Phi(1.80) = 0.96.$$

This is within 0.01 of the actual probability 0.95 in this case.

Theorem 8-8. If $X_1, X_2, \ldots, X_n$ are i.i.d., each $N(\mu, \sigma^2)$, then

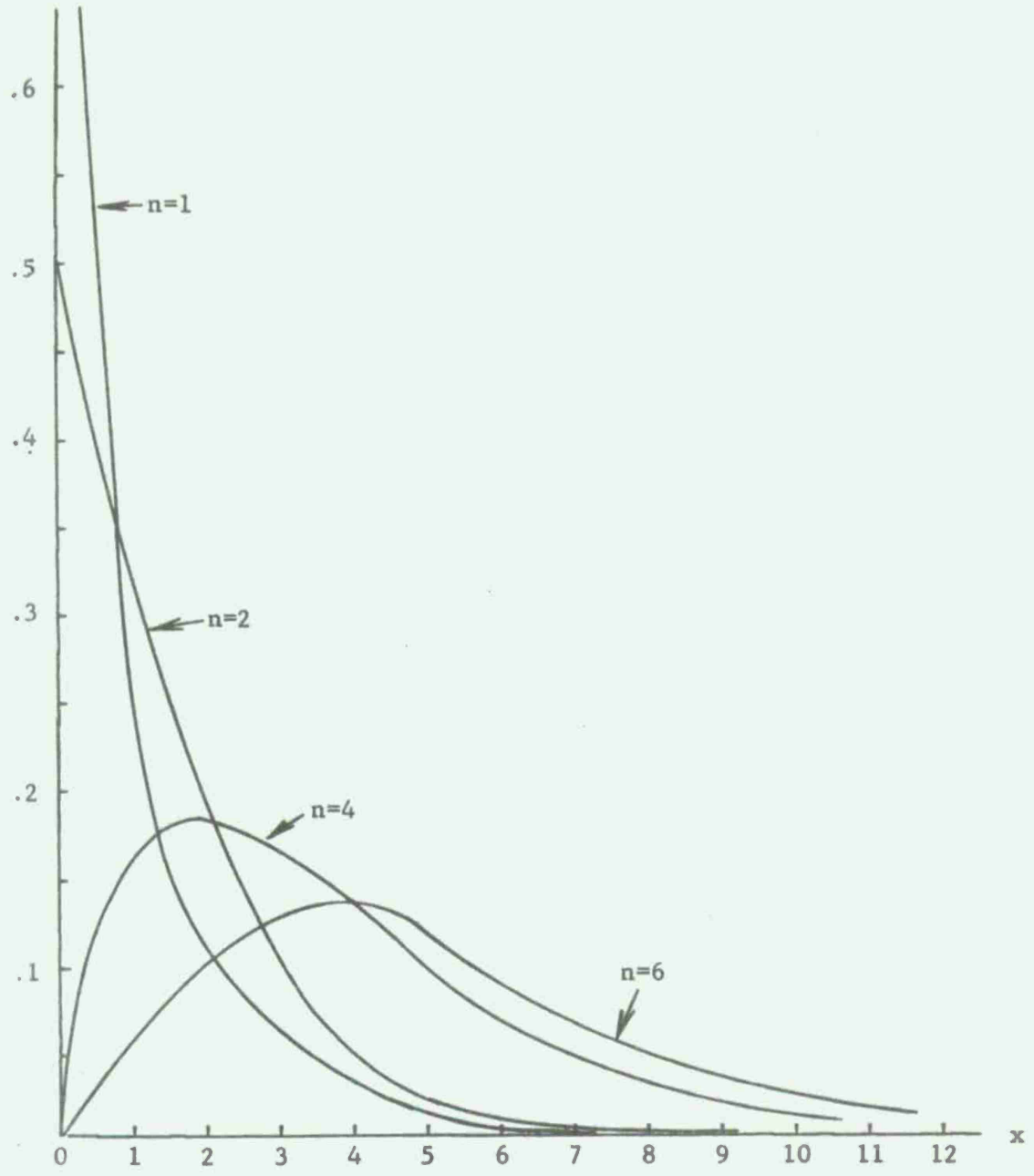
(a) $\Sigma_{i=1}^{n}(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$

(b) $\Sigma_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2$ has a $\chi^2(n-1)$ distribution and is independent of $\bar{X}$.

Proof: (a) Set $Z_i = (X_i - \mu)/\sigma$. Then $Z_1, Z_2, \ldots, Z_n$ are i.i.d., each $N(0,1)$. Therefore, $\Sigma Z_i^2 = \Sigma(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.

(b) The proof of (b) will be omitted, but its plausibility is clear from the following considerations. First, one can verify directly that

$$\Sigma(X_i - \mu)^2 = \Sigma(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$



Chi-square Density Functions.

Table 1

PERCENTAGE POINTS OF A CHI-SQUARE DISTRIBUTION

n \ P	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	$.0^4393$	$.0^3157$	$.0^3982$	$.0^2393$	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

Dividing both sides by σ^2 and rewriting the last term yields

$$\frac{\Sigma(X_i - \mu)^2}{\sigma^2} = \frac{\Sigma(X_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

By part (a), the left member has a $\chi^2(n)$ distribution. The second term on the right has a $\chi^2(1)$ distribution since it is the square of a standard normal random variable. This suggests (but does not prove) that the first term on the right has a $\chi^2(n-1)$ distribution. As for the independence of $\Sigma(X_i - \bar{X})^2$ and $\bar{X}$, this is plausible since $X_i - \bar{X}$ and $\bar{X}$ are independent for each i. The reason is that $X_i - \bar{X}$ and $\bar{X}$ can be shown to have a bivariate normal distribution. Hence, to check their independence it suffices to show they are uncorrelated:

$$\mathrm{Cov}(X_i - \bar{X}, \bar{X}) = \mathrm{Cov}(X_i, \bar{X}) - \mathrm{Var}(\bar{X}) = \mathrm{Cov}(X_i, \Sigma x_j/n) - \sigma^2/n$$

$$= (1/n)\mathrm{Cov}(X_i, X_i) - \sigma^2/n = \sigma^2/n - \sigma^2/n = 0.$$

For a rigorous proof of this theorem, see H. Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, N. J., 1946, Chapter 29.

Example. Suppose you have a random sample of size 30 from a $N(\mu,\sigma^2)$ distribution. If $\sigma^2 = 10$, what is the probability that the sample variance $S^2 = \Sigma(X_i - \bar{X})^2/30$ will exceed 15?

Solution: $P(S^2 > 15) = P\{\Sigma(X_i - \bar{X})^2 > (15)(30)\} = P\{\Sigma(X_i - \bar{X})^2/10 > 45\}$. From Table 1, we see that 45 is between the 95th and 97.5th percentage points (42.6 and 45.7) of a chi-square distribution with 29 degrees of freedom. Using linear interpolation, $P(S^2 > 15) \doteq 1 - 0.97 = 0.03$.

Exercise. 1. (a) Given a random sample of size 30 from a $N(\mu,\sigma^2)$ distribution, find values c and d such that $P(c\sigma^2 < S^2 < d\sigma^2) = 0.95$. (Ans. 0.53, 1.52.) Note that it follows from this that $P(S^2/d < \sigma^2 < S^2/c) = 0.95$. That is, the unknown parameter value σ^2 lies between the random endpoints of

the interval $(S^2/d, S^2/c)$ with probability 0.95.

(b) Using the fact that $\Sigma(X_i - \bar{X})^2/\sigma^2 \sim \chi^2(n-1)$, show that $\mathrm{Var}(S^2) = 2(n-1)\sigma^4/n^2$.

2. Show that if U and V are independent with $U \sim \chi^2(n)$ and $V \sim \chi^2(m)$, then $U+V \sim \chi^2(n+m)$. It follows that if $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples from two normal distributions that have the same variance σ^2, then

$$[\Sigma(X_i - \bar{X})^2 + \Sigma(Y_j - \bar{Y})^2]/\sigma^2 \sim \chi^2(n+m-2).$$

SECTION IX - PARAMETER ESTIMATION

In many statistical applications, the experimental data consist of observations $x_1,x_2,\ldots,x_n$ which, according to some mathematical model, can be regarded as values of random variables $X_1,X_2,\ldots,X_n$ having a joint distribution which depends on a vector of unknown parameters $\theta = (\theta_1,\theta_2,\ldots,\theta_k)$. Quite often the purpose of the experiment is to use the observations to estimate the values of one or more of the parameters θ_j or perhaps some function of the parameters $g(\theta)$. At this stage, we shall not question the appropriateness of the mathematical model, but the student should be aware of the fact that the goodness of certain estimators to be considered below depends critically on the assumption that the joint distribution of the observations $X_1,X_2,\ldots,X_n$ is correctly specified. Although there are many statistical techniques for testing the appropriateness of statistical models, a comprehensive discussion of model-building and methods for assessing appropriateness of models is beyond the scope of this course.

To simplify notation below, let $X^{(n)}$ denote the vector of observations $(X_1,X_2,\ldots,X_n)$, and let $x^{(n)} = (x_1,x_2,\ldots,x_n)$ denote the value of $X^{(n)}$ for a particular experimental outcome. We recall that an estimator $\delta = \delta(X^{(n)})$ is some function of the observations used to estimate a parameter. It is implicit in this definition that δ is a random variable that depends only on the observations X_i and the values of known constants. This is meant to exclude those functions of the observations that depend on the unknown parameters themselves. The value $\delta(x^{(n)})$ of an estimator for a particular experimental outcome is called an estimate of the parameter.

For example, suppose that $X_1, X_2, \ldots, X_n$ are i.i.d., each $N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. Here, the vector of parameters specifying the distribution of $X^{(n)}$ is $\theta = (\mu, \sigma)$, and the joint density of the observations is

$$f(x^{(n)}; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x_i-\mu)^2/2\sigma^2}.$$

Examples of parameters of interest in this case are (a) μ, (b) σ^2, (c) σ, (d) $\mu + 1.28\sigma$, the 90th percentile of the distribution, and (e) $P(X < c) = \Phi(\frac{c-\mu}{\sigma})$, the proportion of the population having x-values below c. Some estimators of μ are:

(1) $\bar{X}$, the sample mean,

(2) X_1, the first observation only,

(3) $[X_{(1)} + X_{(n)}]/2$, the average of the smallest and largest values in the sample,

(4) $\text{mdn}(X^{(n)})$, the sample median,

(5) $[X_{(k)} + X_{(n-k+1)}]/2$, where k is some integer between 1 and n/2,

(6) $[X_{(2)}+X_{(3)}+\ldots+X_{(n-1)}]/(n-2)$, the average of the observations that remain after "trimming" the smallest and largest observations in the sample,

(7) δ_c, the estimator which ignores the observations and estimates μ to be equal to some preassigned constant c,

(8) $pc + (1-p)\bar{X}$ where p is some value between 0 and 1,

(9) $\max(\bar{X}, 0)$, the estimator which estimates μ using $\bar{x}$ if $\bar{x} > 0$ but estimates μ to be equal to 0 if $\bar{x} \leq 0$.

Clearly, in any particular instance, there are infinitely many estimators that can be proposed, and the values of these estimators will have wildly different values for the same experimental outcome. To narrow down the class of estimators that might be considered in a

particular instance, one can impose various criteria that seem reasonable under the circumstances, and then eliminate those estimators that perform poorly according to the standards that are adopted. The difficulty in providing a general theory of estimation is that the goodness criteria can vary widely from application to application. It is not hard to conceive of applications in which each of the nine estimators listed above for μ would be best under certain circumstances. Thus, for example, X_1 would be "better" than $\bar{X}$ if the value of X_1 is available now, but it would cost \$1000 to get each additional observation, and the increased precision is not worth the added cost. If the problem of estimating μ is repeated several or even hundreds of times a day, then perhaps computation time or the difficulty of doing calculations by hand would be factors to be considered. In many applications one needs to worry about the possibility of large recording errors or highly unusual observations in the data, in which case one of the estimators (4)-(6) above might be chosen. Also, there may be considerable evidence from previous experiments (or from experiments taking place concurrently) that ought to be considered in the estimation process.

The presentation that follows will be restricted primarily to considering properties of estimators that are commonly used and are of relevance in a wide number of applications. As we shall see below, the imposition of certain goodness criteria leads to unique "best" estimators of many of the parameters of the distributions introduced in the previous sections. Although the sense in which these estimators are best is narrowly defined and does not include such factors as ease of computation and cost of sampling, these estimators have been widely adopted in practice, and

many of these estimators satisfy other goodness criteria that are not listed here.

Definition. Let $\delta = \delta(X^{(n)})$ be an estimator of a parameter $g(\theta)$. The bias of δ is defined by

$$B(\theta) = E_\theta(\delta) - g(\theta).$$

If $E_\theta(\delta) = g(\theta)$ for all values of θ, δ is said to be an unbiased estimator of $g(\theta)$.

The subscript θ on the expectation sign is included to remind the reader that the distribution, and hence the expectation, of δ depends on θ. Note that unbiasedness requires that $E_\theta(\delta) = g(\theta)$ for all possible values of θ. In order for this definition to be meaningful, the set of possible values of θ must be specified. In the absence of any explicit specification of the parameter set, we shall assume that the set of possible values of θ is the "usual" parameter set for that model. For example, in considering estimates of μ and σ in the $N(\mu,\sigma^2)$ case, the "usual" parameter set is $\{(\mu,\sigma): -\infty < \mu < \infty, 0 < \sigma < \infty\}$. However, in certain applications one may want to restrict the possible values of μ to some subset of the line, e.g., to the nonnegative real numbers. The usual parameter sets for many of the other distributions that will be considered in this section are given in Table 1, Section VI.

Other things equal, we would ordinarily prefer unbiased estimators or, at least, those for which the bias $B(\theta)$ is small for those values of θ that are deemed most likely. As an indication that criteria other than unbiasedness are of more importance in choosing estimators, consider choosing between an unbiased estimator that has large variance and one

that is biased but has a distribution that is much more concentrated about the parameter being estimated for all values of θ. Clearly, what is needed in choosing among estimators are measures of how close the values of the estimators are to the parameters being estimated. Some simple measures that have been proposed in the past are: (a) mean squared error $E_\theta[\delta - g(\theta)]^2$, (b) mean absolute error $E_\theta(|\delta - g(\theta)|)$, and (c) $P_\theta(|\delta - g(\theta)| > c)$, the probability that δ misestimates the parameter $g(\theta)$ by more than c units. Each of these measures of closeness is an instance of $E_\theta\, L(\delta, g(\theta))$ where L is a "loss function" that specifies the loss suffered if the estimated value is δ and the parameter value is $g(\theta)$. Although this more general approach would seem to apply in more situations, in actual practice loss functions can rarely be specified precisely, and we shall not pursue this approach. For the purposes of this presentation, we shall concentrate most of our attention on the first of the three measures of closeness above. It is the easiest to work with, since the mean squared error of an estimator bears a simple relationship to its bias and its variance.

Theorem 9-1. If δ is an estimator of $g(\theta)$ with bias $B(\theta)$, the mean squared error of δ satisfies

$$E_\theta[\delta - g(\theta)]^2 = \mathrm{Var}_\theta(\delta) + [B(\theta)]^2.$$

Proof: This theorem is merely a restatement of the easily verified fact that, for any random variable Y having mean μ and finite variance σ_Y^2,

$$E(Y - c)^2 = \sigma_Y^2 + (\mu - c)^2.$$

The verification is left as an exercise.

It follows from the theorem that, if δ is unbiased for $g(\theta)$, then the mean squared error of δ is just the variance of δ. In many cases, there is a unique unbiased estimator of a parameter that has minimum variance for every possible parameter value θ.

Definition. An unbiased estimator δ^* is said to be the uniformly minimum variance unbiased (UMVU) estimator of a parameter $g(\theta)$ if δ^* has minimum variance for all values of θ. In this case the efficiency of any other estimator δ relative to δ^* is defined to be the ratio $\mathrm{Var}_\theta(\delta^*)/\mathrm{Var}_\theta(\delta)$.

For example, suppose $X_1, X_2, \ldots, X_n$ are i.i.d., each $N(\mu, \sigma^2)$. Of the nine estimators of μ that were listed earlier, the first six are all unbiased. Estimators (3)-(6) are all instances of weighted averages $\Sigma w_i X_{(i)}$ of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ where $\Sigma w_i = 1$ and $w_k = w_{n-k+1}$ for $k = 1,2,\ldots,n$. That is, $X_{(1)}$ and $X_{(n)}$ receive the same weight, $X_{(2)}$ and $X_{(n-1)}$ receive the same weight, etc.

If $n = 20$, the variances of the first six estimators of μ are

(1) $\mathrm{Var}(\bar{X}) = \sigma^2/20 = 0.05\sigma^2$.

(2) $\mathrm{Var}(X_1) = \sigma^2$.

(3) $\mathrm{Var}([X_{(1)} + X_{(n)}]/2) = 0.143\sigma^2$.

(4) $\mathrm{Var}(\mathrm{mdn}(X^{(n)})) = 0.073\sigma^2$.

(5) $\mathrm{Var}([X_{(6)} + X_{(15)}]/2) = 0.061\sigma^2$.

(6) $\mathrm{Var}([X_{(2)} + X_{(3)} + \ldots + X_{(19)}]/18) = 0.051\sigma^2$.

(See W. J. Dixon and Frank J. Massey, Jr., Introduction to Statistical Analysis, Second Edition, McGraw-Hill, New York, p. 406.) Thus, among these unbiased estimators, $\bar{X}$ has smallest variance.

It will be shown later in this section that $\bar{X}$ is the UMVU estimator of μ in this case. The efficiency of the median relative to $\bar{X}$ is $0.05/0.073 = 0.68$. It can be shown that for large values of n the variance of the sample median is approximately $(\pi/2)\sigma^2/n$, so that for large n the efficiency of the median relative to $\bar{X}$ is approximately $2/\pi = 0.64$. The implication of this is that $\bar{X}$ achieves approximately the same precision as the sample median using only 64 percent as many observations.

We note in passing that the "trimmed mean" estimator (6) above has efficiency 0.98. This estimator is almost as efficient as $\bar{X}$, and it affords some protection against gross recording errors and "wild shots" in the data by eliminating the largest and smallest observation from the calculation of the estimate.

The other estimators (7)-(9) are biased estimators of μ, but each of them has smaller mean squared error than $\bar{X}$ for certain values of μ. The mean squared error of δ_c, the estimator which estimates μ to be equal to c for all values of the observations, is equal to $E_\theta(\delta_c - \mu)^2 = (c - \mu)^2$, which is less than the mean squared error of $\bar{X}$, namely σ^2/n, for values of μ close to c.

The estimator $\delta = pc + (1-p)\bar{X}$ has bias

$$E(\delta) - \mu = pc + (1-p)\mu - \mu = p(c-\mu),$$

and its variance is

$$\mathrm{Var}(\delta) = (1-p)^2\mathrm{Var}(\bar{X}) = (1-p)^2\sigma^2/n.$$

Therefore, by Theorem 9-1, the mean-squared error of δ is

$$E_\theta(\delta - \mu)^2 = (1-p)^2\sigma^2/n + p^2(c-\mu)^2.$$

Note that this estimator has smaller variance than $\bar{X}$, so that if the bias of δ is not too large (i.e., if c is close to μ), then δ has smaller

mean squared error than $\bar{X}$.

The biased estimator $\delta = \max(\bar{X},0)$ has smaller mean squared error than $\bar{X}$ for all positive values of μ since

$$(\delta - \mu)^2 \le (\bar{X} - \mu)^2$$

for all possible sample values with strict inequality holding whenever $\bar{X} < 0$. It follows that $E_\theta(\delta - \mu)^2 < E_\theta(\bar{X} - \mu)^2$ for all positive values of μ.

Exercises. 1. Show that, for any random variable Y having mean μ and variance $\sigma_Y^2 < \infty$, $E(Y - c)^2 = \sigma_Y^2 + (\mu - c)^2$.

2. Let $X_1, X_2, \ldots, X_{25}$ be i.i.d., each Bernoulli(p). Compute the bias, variance, and mean squared error of each of the following estimators of p: (a) $\bar{X}$, (b) $\delta_{1/2}$, the estimator having value 1/2 for all values of the X_i's, (c) the "constant risk" estimator $\frac{1 + 10\bar{X}}{12}$. Plot the mean squared error as a function of p for each of the three estimators. Ans. (a) 0, p(1-p)/25, p(1-p)/25; (b) (1-2p)/2, 0, $(1-2p)^2/4$; (c) (1-2p)/12, p(1-p)/36, 1/144.

3. Let $X_1, X_2, \ldots, X_n$ be i.i.d., $N(\mu,\sigma^2)$. Consider the estimators $S^2 = SS/n$ and $\hat{\sigma}^2 = SS/(n-1)$ where $SS = \Sigma(X_i - \bar{X})^2$.
(a) Show that, as estimators of σ^2, S^2 has smaller mean squared error than the unbiased estimator $\hat{\sigma}^2$. (b) Among estimators of the form $\tilde{\sigma}^2 = cSS$, determine the value of c for which $\tilde{\sigma}^2$ has smallest mean squared error. Ans. $c = 1/(n+1)$.

Definition. Suppose $X_1, X_2, \ldots, X_n$ have joint density (or probability function) $f(x^{(n)}; \theta)$ where θ is a vector of unknown parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. If the observed value of $X^{(n)} = (X_1, X_2, \ldots, X_n)$ is $x^{(n)}$, the function

$$L(\theta) = f(x^{(n)}; \theta)$$

considered as a function of θ is called the likelihood function.

If, for each possible value of $X^{(n)}$, there is a unique value of θ, say $\hat{\theta}(X^{(n)})$, that maximizes the likelihood function, then the estimator $\hat{\theta} = \hat{\theta}(X^{(n)})$ determined in this way is called the maximum likelihood estimator (MLE) of θ.

In discrete cases, using the maximum likelihood estimator amounts to choosing θ to maximize the probability of what was observed. As we shall see later, maximum likelihood estimators are usually very good estimators for the parameters in the models that we have discussed so far.

For example, suppose that $X_1, X_2, \ldots, X_n$ are i.i.d., Bernoulli(θ). Then

$$L(\theta) = f(x^{(n)}; \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t}$$

where $t = \Sigma x_i$. We distinguish three cases:

(a) $t = 0$. In this case, the likelihood function is $L(\theta) = (1-\theta)^n$, a strictly decreasing function of θ on the unit interval $[0,1]$ that achieves its maximum at $\hat{\theta} = 0$.

(b) $t = n$. Here, the likelihood function is $L(\theta) = \theta^n$, which achieves its maximum at $\hat{\theta} = 1$.

(c) $0 < t < n$. In this case, the likelihood function is $L(\theta) = \theta^t(1-\theta)^{n-t}$, which is a polynomial in θ that has value 0 at the end points of the unit interval and is positive for $0 < \theta < 1$. The MLE

$\hat{\theta}$ can be determined by setting the derivative of $L(\theta)$ equal to 0 and solving for θ. However, it is easier to determine the maximum of the logarithm of $L(\theta)$:

$$\log L(\theta) = t \log \theta + (n-t) \log(1-\theta).$$

Since $\log x$ is an increasing function of x, the value of θ that maximizes $\log L(\theta)$ will also maximize $L(\theta)$. Setting the derivative of $\log L(\theta)$ equal to 0 yields

$$\frac{t}{\theta} - \frac{n-t}{1-\theta} = 0.$$

Solving for θ gives $\hat{\theta} = t/n = \Sigma x_i/n = \bar{x}$. We conclude that the MLE of θ is $\hat{\theta} = \bar{X}$. This estimator is unbiased and has variance $\theta(1-\theta)/n$. It will be shown later that $\bar{X}$ is the UMVU estimator of θ.

When the joint density or probability function $f(x^{(n)}; \theta)$ has several unknown parameters, one can usually find the MLE's of the parameters by setting the partial derivatives of $\log L(\theta)$ with respect to the parameters θ_i equal to zero and solving the resulting equations.

For example, if $X_1, X_2, \ldots, X_n$ are i.i.d., each $N(\mu, \sigma^2)$, then

$$L(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = (\sqrt{2\pi}\,\sigma)^{-n} \exp[-\Sigma(x_i-\mu)^2/2\sigma^2].$$

Since $\Sigma(x_i-\mu)^2 = \Sigma(x_i-\bar{x})^2 - n(\bar{x}-\mu)^2$,

$$\log L(\mu,\sigma) = -n \log \sqrt{2\pi} - n \log \sigma - \frac{\Sigma(x_i-\bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}.$$

Here we could set the partial derivatives of $\log L(\mu,\sigma)$ with respect to μ and σ equal to zero and solve the resulting two equations for μ and σ. However, we observe that μ only occurs in the last term on the right, and this term is maximized by setting $\hat{\mu} = \bar{x}$. Thus the problem reduces to choosing σ to minimize the sum of the other terms.

Setting $\partial \log L(\mu,\sigma)/\partial\sigma = 0$ and solving for σ^2 yields

$$\sigma^2 = \Sigma(x_i - \bar{x})^2/n.$$

Thus, the MLE's of μ and σ are $\bar{X}$ and $S = [\Sigma(X_i - \bar{X})^2/n]^{1/2}$. The MLE of functions of μ and σ are the corresponding functions of $\bar{X}$ and S. For example, the MLE of σ^2 is S^2, and the MLE of $\mu + 1.28\sigma$ is $\bar{X} + 1.28S$. As we saw earlier, S^2 is a biased estimator of σ^2, but it has smaller mean squared error than the usual unbiased estimator $\hat{\sigma}^2 = \Sigma(X_i - \bar{X})^2/(n-1)$. The MLE of σ is also a biased estimator. An unbiased estimator of σ can be obtained by taking $\tilde{\sigma} = c_n\hat{\sigma}$ where $\hat{\sigma}$ is the square root of $\hat{\sigma}^2$ and

$$c_n = [(n-1)/2]^{1/2}\Gamma[(n-1)/2]/\Gamma(n/2).$$

The values of c_n for $n \leq 10$ are

n	2	3	4	5	6	7	8	9	10
c_n	1.253	1.128	1.085	1.064	1.051	1.042	1.036	1.032	1.028

For $n > 5$, c_n is well approximated by $1 + 1/4(n-1)$. Another way of representing this unbiased estimator is in the form $\tilde{\sigma} = [\Sigma(X_i - \bar{X})^2/k_n]^{1/2}$, where $k_n = 2\{\Gamma(n/2)/\Gamma[(n-1)/2]\}^2$. The values of k_n for $n \leq 10$ are:

n	2	3	4	5	6	7	8	9	10
k_n	0.637	1.571	2.546	3.534	4.527	5.522	6.519	7.517	8.515

For $n > 10$, k_n is approximately equal to $n - 3/2$. (See John Gurland and Ram C. Tripathi, "A Simple Approximation for Unbiased Estimation of the Standard Deviation," The American Statistician, October 1971, pp. 30-32.)

Example. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d., each uniformly distributed on $(0,\theta)$, so that the density of each of the X_i's can be specified as $f(x_i;\theta) = 1/\theta$ for $0 < x_i \leq \theta$.

Here, the likelihood function is

$$L(\theta) = \begin{cases} 0 & \text{if } x_i < 0 \text{ or } x_i > \theta \text{ for some } i, \\ 1/\theta^n & \text{if } 0 < x_i \le \theta \text{ for } i = 1,2,\ldots,n. \end{cases}$$

Since $L(\theta)$ is a decreasing function of θ for $\theta \ge x_{(n)} = \max(x_1,\ldots,x_n)$ and $L(\theta)$ is zero for $\theta < x_{(n)}$, it follows that $L(\theta)$ is maximized by $\hat{\theta} = x_{(n)}$, and the MLE of θ is $\hat{\theta} = X_{(n)}$. By Exercise 1 below, $\hat{\theta}$ is a biased estimator of θ with expectation $E(\hat{\theta}) = n\theta/(n+1)$ and $\mathrm{Var}(\theta) = n\theta^2/(n+1)^2(n+2)$.

Next consider estimating $\mu = \theta/2$, the mean of the X_i's. The MLE of μ is $\hat{\mu} = \hat{\theta}/2$, which is again a biased estimator. The corresponding unbiased estimator of μ that depends on $X_{(n)}$ is $\tilde{\mu} = (n+1)\hat{\theta}/2n$, which has variance

$$\mathrm{Var}(\tilde{\mu}) = (n+1)^2\mathrm{Var}(\hat{\theta})/4n^2 = \theta^2/4n(n+2).$$

How does this compare with the sample mean $\bar{X}$? Since each X_i has variance $\theta^2/12$, $\mathrm{Var}(\bar{X}) = \theta^2/12n$. Hence, the efficiency of $\bar{X}$ relative to $\tilde{\mu}$ is $\mathrm{Var}(\tilde{\mu})/\mathrm{Var}(\bar{X}) = 3/(n+2)$. Note how poorly $\bar{X}$ performs relative to $\tilde{\mu}$ in this case. For example, if $n = 28$, the variance of $\tilde{\mu}$ is only one tenth as large as the variance of $\bar{X}$.

Exercises. 1. Let $Y = X_{(n)}$ where $X_1, X_2,\ldots,X_n$ are i.i.d., each Uniform(0, θ). Show that (a) the density of Y is $f(y) = ny^{n-1}/\theta^n$ for $0 < y < \theta$, (b) $E(Y) = n\theta/(n+1)$, and (c) $\mathrm{Var}(Y) = n\theta^2/(n+1)^2(n+2)$.

2. Show that, if $X_1, X_2,\ldots,X_n$ are i.i.d., each having a negative exponential distribution with parameter λ, then the MLE of λ is $\hat{\lambda} = 1/\bar{X}$.

3. Assume that $Y_1,Y_2,\ldots,Y_n$ are independent random variables, and $Y_i \sim N(\alpha x_i,\sigma^2)$ where $\sigma, x_1, x_2,\ldots,x_n$ are known constants. Show that (a) the MLE of α is $\hat{\alpha} = \Sigma x_i Y_i/\Sigma x_i^2$, (b) $\hat{\alpha}$ is an unbiased estimator of α with variance $\sigma^2/\Sigma x_i^2$.

4. Show that, if $X_1,X_2,\ldots,X_n$ are i.i.d., each Poisson (λ), then the MLE of λ is $\hat{\lambda} = \bar{X}$.

The reader should note that no optimality properties for maximum likelihood estimators were stated in the previous section. There is a good reason for this omission--namely, the fact that some MLE's are poor estimators. Sometimes it is asserted that MLE's are good estimators because they have desirable "asymptotic" properties. To see that the reasoning behind this assertion is shaky, let us first define our terms.

Definition. Suppose the vector of observations $X^{(n)}$ has joint density or probability function $f(x^{(n)}; \theta)$, and let $\hat{\gamma}_n = \hat{\gamma}(X^{(n)})$ be an estimator (or, more precisely, a sequence of estimators) of a parameter $\gamma = g(\theta)$. The sequence $\hat{\gamma}_n$ is said to be

(a) consistent if $\hat{\gamma}_n$ tends to γ in probability [i.e., for any $\epsilon > 0$, $P_\theta(|\hat{\gamma}_n - \gamma| \geq \epsilon)$ tends to 0 as n becomes infinite];

(b) asymptotically normal with mean γ and variance σ^2/n if the distribution of $\sqrt{n}(\hat{\gamma}_n - \gamma)/\sigma$ tends to a standard normal distribution,

(c) best asymptotically normal (BAN) if $\hat{\gamma}_n$ is asymptotically normal with mean γ and variance σ^2/n and, if $\tilde{\gamma}_n$ is any other asymptotically normal sequence with mean γ and variance τ^2/n, then $\sigma^2 \leq \tau^2$.

Since $P_\theta(|\hat{\gamma}_n - \gamma| \geq \epsilon) \leq E_\theta(\hat{\gamma}_n - \gamma)^2/\epsilon^2$ by Theorem 5-6, and

$$E_\theta(\hat{\gamma}_n - \gamma)^2 = \mathrm{Var}_\theta(\hat{\gamma}_n) + B_n^2(\theta)$$

where $B_n(\theta)$ is the bias of $\hat{\gamma}_n$, to prove consistency it suffices to show that $\mathrm{Var}(\hat{\gamma}_n) \to 0$ and $E(\hat{\gamma}_n) \to \gamma$ for all values of θ. Thus for example, if $X_1, X_2, \ldots, X_n$ are i.i.d., each $N(\mu, \sigma^2)$, then $\bar{X}$ is a consistent, asymptotically normal estimator of μ, but so are the following ridiculous estimators:

(a) the average of X_1 and every thousandth observation thereafter,

(b) $\tilde{\mu} = \begin{cases} 17 & \text{if } n < 10^{10} \\ \bar{X} & \text{if } n \geq 10^{10} \end{cases}$,

(c) $\mu^* = (n\bar{X} + 10^{10})/(n+1)$.

The point of these examples is that consistency says nothing about the goodness of an estimator for small samples or even for very large ones. Conversely, inconsistent estimators may still be good in small samples. As a frivolous example in the normal case above, consider

$$\hat{\mu} = \begin{cases} \bar{X} & \text{if } n < 10^{10} \\ 0 & \text{if } n \geq 10^{10} \end{cases}.$$

The reason for citing asymptotic properties of estimators is that in many cases it is difficult to determine the properties of estimators in small samples, but methods exist for determining their asymptotic distributions. A second reason is based on the wishful thinking that those estimators that have desirable asymptotic properties will also prove to be good in small samples.

For what it is worth, if $X_1, X_2, \ldots, X_n$ are i.i.d., each having density or probability function $f(x_i; \theta)$, and if $\hat{\theta}$ is the MLE of θ, then $\hat{\theta}$ is a consistent, BAN estimator of θ provided that $f(x_i; \theta)$ satisfies certain regularity conditions.[1] On the other hand, examples exist to show that MLE's need not be consistent.[2]

Since the method of maximum likelihood sometimes leads to poor estimators, the reader may wonder why we have devoted so much space to this topic. The reason is that there is no single method for deriving good

---

[1] For a comprehensive discussion of maximum likelihood estimation, see M. G. Kendall, and Alan Stuart, The Advanced Theory of Statistics, Vol. 2, Hafner Publishing Company, New York, 1961, Chapter 18.

[2] See Kendall and Stuart, ibid., p. 61. Also, R. R. Bahadur, "Examples of Inconsistency of Maximum Likelihood Estimates," Sankhya, December 1958, pp. 207-210.

estimators, and the maximum likelihood estimators provide a reasonable starting point. Another reason is that, if $T = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ where $\hat{\theta}_i$ is the MLE of θ_i, then it often happens that, if $g(\theta)$ is a parameter for which a UMVU estimator δ^* exists, then δ^* is usually either $g(T)$ or some multiple of $g(T)$. Moreover, T is usually a "sufficient" statistic for θ.

Definition. Let $X^{(n)}$ have joint density or probability function $f(x^{(n)}; \theta)$. A statistic T is said to be sufficient for $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ if the conditional distribution of $X^{(n)}$, given $T = t$, does not depend on θ.

The importance of a sufficient statistic, which may be a single random variable or a vector of random variables $T = (T_1, T_2, \ldots, T_m)$, is that it summarizes all the information about θ that is contained in the sample values. Since the conditional distribution of $X^{(n)}$, given $T = t$, does not depend on θ, it follows that the conditional distribution of any other statistic $U = u(X^{(n)})$ does not depend on θ either. Since the conditional distribution of U is the same for all θ, knowing the value of U cannot provide any additional information about the value of θ

Example. Let $X_1, X_2, \ldots, X_n$ be i.i.d., Bernoulli(θ). To see that $T = \Sigma X_i$ is sufficient for θ, consider

$$P_\theta(X^{(n)} = x^{(n)} | T = t) = P_\theta(X^{(n)} = x^{(n)}, T = t)/P_\theta(T = t).$$

The numerator on the right is zero unless $t = \Sigma x_i$, in which case

$$P_\theta(X^{(n)} = x^{(n)}, T = t) = \theta^{\Sigma x_i}(1 - \theta)^{n - \Sigma x_i} = \theta^t(1-\theta)^{n-t}.$$

Since $P_\theta(T = t) = \binom{n}{t} \theta^t(1-\theta)^{n-t}$,

$$P_\theta(X^{(n)} = x^{(n)} | T = t) = \begin{cases} 0 & \text{if } t \neq \Sigma x_i \\ 1/\binom{n}{t} & \text{if } t = \Sigma x_i. \end{cases}$$

The expression on the right is free of θ, completing the proof that T is sufficient for θ.

In general, it is hard to establish sufficiency directly from the definition as was done in this case. Fortunately, the following theorem enables us to spot sufficient statistics easily from the joint density or probability function of $X^{(n)}$.

Theorem 9-2. (Fisher-Neyman Factorization Theorem.) A statistic $T = t(X^{(n)})$ is a sufficient statistic for θ if and only if the joint density or probability function of $X^{(n)}$ can be factored into two parts

$$f(x^{(n)}; \theta) = g(t,\theta)\, h(x^{(n)}),$$

where $g(t,\theta)$ depends only on $t = t(x^{(n)})$ and the parameter(s) θ, and $h(x^{(n)})$ does not depend on θ.

Example. In the Bernoulli case above,

$$f(x^{(n)}; \theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}.$$

Here, we can apply the Factorization Theorem by setting $h(x^{(n)}) = 1$ and $g(t,\theta) = \theta^t(1-\theta)^{n-t}$ where $t = \Sigma x_i$. It follows that $T = \Sigma X_i$ is a sufficient statistic for θ.

Example. Let $X_1, X_2, \ldots, X_n$ be i.i.d., each $N(\mu,\sigma^2)$ where both μ and σ^2 are unknown. Then

$$f(x^{(n)}; \mu,\sigma) = (\sqrt{2\pi}\,\sigma)^{-n}\, e^{-\Sigma(x_i-\mu)^2/2\sigma^2}.$$

Since $\Sigma(x_i-\mu)^2 = \Sigma(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$, it follows from the Factorization Theorem by setting $h(x^{(n)}) = 1$ that $T = (\bar{X}, \Sigma(X_i-\bar{X})^2)$ is a (set of) sufficient statistic(s). If σ^2 is known, then using the factorization

$$f(x^{(n)}; \mu) = e^{-n(\bar{x}-\mu)^2/2\sigma^2} \cdot (\sqrt{2\pi}\,\sigma)^{-n}\, e^{-\Sigma(x_i-\bar{x})^2},$$

we see that $\bar{X}$ is sufficient for μ. If μ is known, then it follows from the first representation above that $\Sigma(X_i-\mu)^2$ is sufficient for σ^2.

Note that it follows from the Factorization Theorem that, if T is sufficient for θ, and $U = u(T)$ is some one-to-one function of T, then U is also sufficient for θ. For example, in the Bernoulli case above, knowing that $T = \Sigma X_i$ is sufficient for θ implies that $\bar{X}$ is also sufficient for θ. In the normal case, $(\bar{X}, S^2)$, $(\bar{X}, \hat{\sigma}^2)$, and $(\Sigma X_i, \Sigma X_i^2)$ are all sufficient statistics for μ and σ.

Theorem 9-3. (Rao-Blackwell Theorem.) Let T be a sufficient statistic for θ, and let δ be any unbiased estimator of $g(\theta)$. Then $\delta^* = E(\delta|T)$ is also an unbiased estimator of $g(\theta)$ and $\text{Var}_\theta(\delta^*) \le \text{Var}_\theta(\delta)$ with strict inequality holding unless δ is a function of T. If δ is a biased estimator of $g(\theta)$, then δ^* has the same bias as δ for all values of θ, and $\text{Var}(\delta^*) \le \text{Var}(\delta)$, implying that the mean squared error of δ^* is at least as small as that for δ for all θ.

Proof: $\delta^* = E(\delta|T)$ is a function of T alone (and does not depend on θ) since, given $T = t$, the conditional distribution of $\delta = \delta(X^{(n)})$ is independent of θ. Hence, the conditional expectation of δ, given T, is independent of θ. The estimator δ^* has the same bias as δ, because δ^* and δ have the same expectation for all values of θ by Theorem 7-6(a):

$$E_\theta(\delta^*) = E_\theta(E(\delta|T)) = E_\theta(\delta).$$

The fact that $\text{Var}(\delta^*) \le \text{Var}(\delta)$ follows from Exercise 4(b), page 92:

$$\text{Var}_\theta(\delta) = E_\theta[\text{Var}(\delta|T)] + \text{Var}_\theta[E(\delta|T)],$$

and the fact that $\text{Var}(\delta|T) \ge 0$. Note that $\text{Var}_\theta(\delta^*) < \text{Var}_\theta(\delta)$ unless $\text{Var}(\delta|T) = 0$, which would imply that δ is a function of T.

An implication of the theorem is that any estimator that is not a function of a sufficient statistic can always be improved upon by an estimator that is a function of a sufficient statistic. For example, suppose $X_1, X_2, \ldots, X_n$ are i.i.d., each having a uniform distribution on $(0,\theta)$. By writing the

joint density of the observations in the form

$$f(x^{(n)};\ \theta) = (1/\theta^n)\ I(x_{(n)} \le \theta)$$

where $I(x_{(n)} \le \theta)$ is 1 if $x_{(n)} \le \theta$ and 0 otherwise, we see from the Factorization Theorem that $T = X_{(n)}$ is a sufficient statistic for θ. Consider estimating $\mu = \theta/2$, the mean of the X_i's. Two unbiased estimators of μ are X_1 and $\bar{X}$, neither of which are functions of the sufficient statistic. It follows that $E(X_1|T)$ and $E(\bar{X}|T)$ are unbiased estimators of μ having smaller variance than either X_1 and $\bar{X}$. It turns out that $E(X_1|T) = E(\bar{X}|T) = \tilde{\mu}$, where $\tilde{\mu} = (n+1)T/2n$. This is the same estimator of μ that was derived on page 122.

The Rao-Blackwell Theorem would seem to provide a useful tool for improving upon estimators. However, the tool is rarely used since, in the commonly used statistical models in which the density or probability function $f(x^{(n)};\ \theta)$ is known except for the parameter values $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, the "standard" estimators are either MLE's or functions of the MLE's, and it follows easily from the Factorization Theorem that, if T is sufficient for θ, then the MLE δ of $g(\theta)$ is a function of T, say $\delta(T)$. Since $E(\delta(T)|\ T) = \delta(T)$ by Theorem 7-6(b), MLE's are unaffected by conditioning on a sufficient statistic.

Although it is not true in general, it often happens that if $T = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ where $\hat{\theta}_i$ is the MLE of θ_i, then T is a "minimal" sufficient statistic for θ, i.e., all other sufficient statistics are functions of T. Moreover, it often happens that the unbiased estimator of a parameter $g(\theta)$ that depends on T is <u>unique</u> in the sense that if $\delta_1(T)$ and $\delta_2(T)$ are two unbiased estimators of $g(\theta)$, then $\delta_1(T) = \delta_2(T)$ except perhaps on a subset of the sample space that has probability zero for all values of θ. Under these circumstances, it then follows from the

Rao-Blackwell Theorem that, if one can find a single estimator δ^* that is a function of T, then δ^* is the UMVU estimator of $g(\theta)$. Any other unbiased estimator δ can be improved upon by $E(\delta|T)$, but this is an unbiased estimator that depends on T, and by assumption δ^* is the unique unbiased estimator that is a function of T.

We shall now define a property of sufficient statistics T that assures uniqueness of unbiased estimators that are functions of T.

Definition. A statistic T is said to be complete if the only real-valued functions $h(T)$ satisfying $E_\theta[h(T)] = 0$ for all values of θ are those for which $P_\theta\{h(T) = 0\} = 1$ for all θ.

To see that unbiased estimators that depend on a complete, sufficient statistic are unique in the sense specified above, suppose $\delta_1(T)$ and $\delta_2(T)$ are two unbiased estimators of the same parameter $g(\theta)$. Then $E_\theta[\delta_1(T) - \delta_2(T)] = 0$ for all values of θ. By the definition of completeness, it follows that $\delta_1(T) = \delta_2(T)$ except perhaps on a set probability zero.

Theorem 9-4. (Lehmann-Scheffé Theorem.) Suppose T is a complete, sufficient statistic for θ, and $g(\theta)$ has at least one unbiased estimator. Then $g(\theta)$ has a unique UMVU estimator that depends on T.

Proof: Let δ be any unbiased estimator of $g(\theta)$. Then, by the Rao-Blackwell Theorem, $\delta^* = E(\delta|T)$ is again unbiased for $g(\theta)$, and $\text{Var}(\delta^*) \leq \text{Var}(\delta)$ with equality holding if and only if δ is a function of T. Since T is complete, δ^* is the unique unbiased estimator of $g(\theta)$ depending on T.

Many of the statistical models that we have considered have complete, sufficient statistics T that can be determined by setting $T = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ where $\hat{\theta}_i$ is the MLE of θ_i. The sufficiency of the statistics T in

the examples below can be verified by the Factorization Theorem. The proofs of the completeness of many of these statistics are special cases of a theorem that can be found in E. L. Lehmann, Testing Statistical Hypotheses, John Wiley & Sons, New York, p. 132.

Bernoulli. If $X_1, X_2, \ldots, X_n$ are i.i.d., Bernoulli(θ), then $\bar{X}$ is complete and sufficient for θ. Since $\bar{X}$ is unbiased for θ, it is the UMVU estimator of θ. Let $g(\theta) = \theta(1-\theta)/n$, which is the variance of $\bar{X}$. Since $\bar{X}(1 - \bar{X})/(n-1)$ is an unbiased estimator of $g(\theta)$ that depends on $\bar{X}$, it follows that $\bar{X}(1 - \bar{X})/(n-1)$ is the UMVU estimator of $g(\theta)$.

Poisson. If $X_1, X_2, \ldots, X_n$ are i.i.d., Poisson(λ), then the MLE of λ is $\bar{X}$, which is a complete and sufficient statistic. Since $\bar{X}$ is unbiased for λ, it is the UMVU estimator of λ.

Geometric. If $X_1, X_2, \ldots, X_n$ are i.i.d., each Geometric(p), then the MLE of p is $1/\bar{X}$. This is a biased estimator, but it is complete and sufficient for p. The UMVU estimator of p is $(n-1)/(\Sigma X_i - 1)$ if $n > 1$. If $n = 1$, the UMVU estimator of p has value 1 if $X_1 = 1$ and 0 if $X_1 > 1$. (In this case, the UMVU estimator is absurd.)

Normal. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d., each $N(\mu, \sigma^2)$.

(a) If both μ and σ^2 are unknown, the MLE of $\theta = (\mu, \sigma)$ is $T = (\bar{X}, S)$, which is complete and sufficient for θ. Let $\hat{\mu} = \bar{X}, \hat{\sigma}^2 = \Sigma(X_i - \bar{X})^2/(n-1)$, and $\tilde{\sigma} = c_n\hat{\sigma}$ where c_n is defined on page 121. Since these are unbiased estimators that are functions of the complete, sufficient statistic T, they are the UMVU estimators of μ, σ^2, and σ.

(b) If σ^2 is known, $\bar{X}$ is complete and sufficient for μ. Hence, $\bar{X}$ is the UMVU estimator of μ.

(c) If μ is known, the MLE of σ^2 is $\hat{\sigma}^2 = \Sigma(X_i - \mu)^2/n$, which is complete and sufficient. Since $\hat{\sigma}^2$ is unbiased, it is the UMVU estimator of σ^2.

Two-sample Normal. Suppose $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ are independent, $X_i \sim N(\xi, \sigma^2)$, $Y_j \sim N(\eta, \tau^2)$.

(a) If all parameters are unknown, the MLE's $(\bar{X}, S_X^2, \bar{Y}, S_Y^2)$ are complete and sufficient for $\theta = (\xi, \sigma^2, \eta, \tau^2)$. Hence, the UMVU estimators of ξ, σ^2, η, τ^2, and $\xi - \eta$ are $\bar{X}$, $\hat{\sigma}^2$, $\bar{Y}$, $\hat{\tau}^2$, and $\bar{X} - \bar{Y}$.

(b) If $\tau^2 = \sigma^2$ (by assumption) and ξ, η, and σ^2 are all unknown, then the MLE's $\bar{X}$, $\bar{Y}$, and $S^2 = [\Sigma(X_i - \bar{X})^2 + \Sigma(Y_j - \bar{Y})^2]/(m+n)$ are complete and sufficient. It follows that $\bar{X}$, $\bar{Y}$, and $(m+n)S^2/(n+m-2)$ are the UMVU estimators of ξ, η, and σ^2.

Bivariate Normal. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ. The MLE's of these parameters are $(\bar{X}, \bar{Y}, S_X^2, S_Y^2, r)$ where $S_X^2 = \Sigma(X_i - \bar{X})^2/n$ and r is the sample correlation coefficient. Since these statistics are complete and sufficient, the unbiased estimators $\bar{X}$, $\bar{Y}$, $\hat{\sigma}_X^2$, and $\hat{\sigma}_Y^2$ are the UMVU estimators of μ_X, μ_Y, σ_X^2, and σ_Y^2. It can be shown that r is a biased estimator of ρ with mean approximately equal to $\rho[1 - (1-\rho^2)/2n]$. Although the UMVU estimator exists [see I. Olkin and J. W. Pratt, "Unbiased estimation of Certain Correlation Coefficients," Annals of Mathematical Statistics, Vol. 29 (1958), p. 201], it is a complicated function of r. Olkin and Pratt recommend using the approximation $r[1 - (1-r^2)/2(n-4)]$.

Exercise. It can be shown that, if $X_1, X_2, \ldots, X_n$ are i.i.d., each Negative Exponential(λ), the MLE of λ is complete and sufficient. Determine the UMVU estimators of $1/\lambda$ and $1/\lambda^2$, the mean and variance of the X_i's. Ans. $\bar{X}$, $n\bar{X}^2/(n+1)$.

Assume that $Y_1, Y_2, \ldots, Y_n$ are independent random variables such that $E(Y_i) = \eta_i$ and $Var(Y_i) = \sigma_i^2 < \infty$. Let θ be some parameter such that $\theta = \Sigma c_i \eta_i$ for some choice of constants c_i. Then θ has at least one unbiased "linear" estimator--namely, $\Sigma c_i Y_i$.

Definition. Given the situation above, we say that $\hat{\theta}$ is the best linear unbiased estimator (BLUE) of θ if $\hat{\theta}$ is a linear function of the Y_i's (i.e., $\hat{\theta} = \Sigma a_i Y_i$) and $\hat{\theta}$ has minimum variance among all unbiased linear estimators of θ.

The expectation and variance of any linear estimator $\hat{\theta} = \Sigma a_i Y_i$ are given by $E(\hat{\theta}) = \Sigma a_i \eta_i$ and $Var(\hat{\theta}) = \Sigma a_i^2 \sigma_i^2$. Note that these characteristics of θ depend only on the means and variances of the Y_i's, but no further assumptions about the distributions of the Y_i's are needed. Thus, one can determine BLUE's without specifying the exact distributions of the observations Y_i.

For example, suppose $Y_1, Y_2, \ldots, Y_n$ have a common mean θ but possibly different variances σ_i^2, and one wants to find the BLUE of θ. This situation applies if the $Y_1, Y_2, \ldots, Y_n$ are a random sample from any distribution having finite variance, in which case the Y_i's have a common mean θ and a common variance σ^2. More generally, it applies in any situation where $Y_1, Y_2, \ldots, Y_n$ are independent unbiased estimators of the same parameter θ. For example, Y_i may be the average of n_i i.i.d. random variables having mean θ and variance σ^2, in which case $E(Y_i) = \theta$ and $Var(Y_i) = \sigma^2/n_i$.

Theorem 9-5. If $Y_1, Y_2, \ldots, Y_n$ are independent with $E(Y_i) = \theta$ and $Var(Y_i) = \sigma_i^2 < \infty$, the BLUE of θ based on the Y_i's is $\hat{\theta} = \Sigma w_i Y_i$ where the weights w_i satisfy $\Sigma w_i = 1$ and are inversely proportional to the

variances σ_i^2 (i.e., $w_i = \gamma_i/\Sigma\gamma_i$ where $\gamma_i = 1/\sigma_i^2$). In particular, if the Y_i's have the same variance, the BLUE of θ is $\hat{\theta} = \bar{Y}$.

Proof: Let $\tilde{\theta} = \Sigma a_i Y_i$ be any unbiased estimator of θ. Since $E(\tilde{\theta}) = \Sigma a_i \theta$ and $Var(\tilde{\theta}) = \Sigma a_i^2 \sigma_i^2$, the unbiasedness condition implies that $\Sigma a_i = 1$, and the problem reduces to finding a vector of constants $a = (a_1, a_2, \ldots, a_n)$ to minimize $f(a) = \Sigma a_i^2 \sigma_i^2$ subject to the condition that $\Sigma a_i = 1$. In minimizing $f(a)$ on the set $A = \{a: \Sigma a_i = 1\}$, one can just as well consider minimizing

$$g(a) = \Sigma a_i^2 \sigma_i^2 + \lambda (\Sigma a_i - 1)$$

where λ is any constant, since the functions f and g are equal on A. The trick, called the "method of Lagrange multipliers," is to use differentiation to find the value a^* that minimizes $g(a)$ over all values of a in R^n. In general, the components of a^* will depend on λ, but one can determine a value of λ for which a^* is in A. Since this value of a^* minimizes g over R^n and since a^* is in A, a^* also minimizes f over A. Here,

$$\frac{\partial g(a)}{\partial a_i} = 2a_i\sigma_i^2 + \lambda \qquad \text{for } i = 1,2,\ldots,n,$$

and it follows that $g(a)$ is minimized over R^n by $a_i^* = -\lambda\gamma_i/2$ where $\gamma_i = 1/\sigma_i^2$. In order to have $\Sigma a_i^* = 1$, we choose $\lambda = -2/\Sigma\gamma_i$, in which case $a_i^* = \gamma_i/\Sigma\gamma_i$.

Now suppose that $Y_1, Y_2, \ldots, Y_n$ are independent random variables with means $E(Y_i) = \alpha + \beta x_i$ and $Var(Y_i) = \sigma^2$ where $x_1, x_2, \ldots, x_n$ are known constants such that not all of them are equal, and the "regression coefficients" α and β are parameters to be estimated

from the observations. In the absence of specific assumptions about the exact distributions of the Y_i's, we search for the BLUE's of α and β.

Let us assume for the moment that the Y_i's have normal distributions, i.e., $Y_i \sim N(\alpha+\beta x_i, \sigma^2)$. Led by the hope that the MLE's of α and β will turn out to be linear, consider the likelihood function in this case:

$$L(\alpha,\beta,\sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) e^{-(y_i-\alpha-\beta x_i)^2/2\sigma^2} = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-SS/2\sigma^2}$$

where

$$SS = \Sigma_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

Note that the values a and b of α and β that maximize the likelihood function are the values of a and b that minimize the sum of squares SS. Hence, the MLE's a and b are called the <u>least-squares estimators</u> in this case. The partial derivatives of SS with respect to α and β are:

$$\frac{\partial SS}{\partial \alpha} = -2\Sigma(y_i - \alpha - \beta x_i)$$

$$\frac{\partial SS}{\partial \beta} = -2\Sigma x_i(y_i - \alpha - \beta x_i)$$

Setting these partial derivatives equal to zero and solving for α and β yields the MLE's

$$b = \Sigma(x_i-\bar{x})Y_i/\Sigma(x_i-\bar{x})^2$$

$$a = \bar{Y} - b\bar{x}.$$

Note that a and b are both linear estimators of α and β. Are they unbiased? To verify that b is unbiased, we compute

$$E(b) = \frac{\Sigma(x_i-\bar{x})(\alpha+\beta x_i)}{\Sigma(x_i-\bar{x})^2} = \frac{\alpha\Sigma(x_i-\bar{x})}{\Sigma(x_i-\bar{x})^2} + \frac{\beta\Sigma(x_i-\bar{x})x_i}{\Sigma(x_i-\bar{x})^2}.$$

The first term on the right is zero because $\Sigma(x_i-\bar{x}) = 0$; the second term reduces to β because $\Sigma(x_i-\bar{x})x_i = \Sigma x_i^2 - n\bar{x}^2$, which is another way of writing the denominator. Hence, $E(b) = \beta$. The estimator a is also unbiased, because

$$E(a) = E(\bar{Y}) - E(b)\bar{x} = \Sigma(\alpha+\beta x_i)/n - \beta\bar{x} = \alpha.$$

Incidentally, the MLE of σ^2 in this case is $\hat{\sigma}^2 = SS_e/n$ where

$$SS_e = \Sigma(Y_i - a - bx_i)^2.$$

Although $\hat{\sigma}^2$ is a biased estimator of σ^2, the estimator $\tilde{\sigma}^2$ obtained by dividing the "residual sum of squares" SS_e by $n-2$ can be shown to be unbiased. It also turns out that SS_e is independent of a and b, and SS_e/σ^2 has a chi-square distribution with $n-2$ degrees of freedom.

The clincher in this example is that a, b, and $\tilde{\sigma}^2$ can be shown to be complete and sufficient statistics for the parameters α, β, and σ^2. It follows that a, b, and $\tilde{\sigma}^2$ are the UMVU unbiased estimators of the parameters.

What is the implication of this for the original problem of finding the BLUE's of α and β? Since the calculations of the expectations and variances of the linear estimators a and b do not use the normality assumptions, a and b are unbiased linear estimators of α and β whether the Y_i's have normal distributions or not. Moreover, they must be the BLUE's of α and β, because if there were another unbiased linear estimator, say b_1, which had smaller variance than b, then b_1 would be a better unbiased estimator than b in the normal case, contradicting the fact that b is UMVU in the normal case. This completes the proof of the following theorem:

Theorem 9-6. If $Y_1, Y_2, \ldots, Y_n$ are independent observations with $E(Y_i) = \alpha + \beta x_i$ and $Var(Y_i) = \sigma^2$ where $x_1, x_2, \ldots, x_n$ are given constants, then the BLUE's of α and β are the least squares estimators $a = \bar{Y} - b\bar{x}$ and $b = \Sigma(x_i - \bar{x})Y_i / \Sigma(x_i - \bar{x})^2$. Moreover, if the observations Y_i are normally distributed, then a and b are the UMVU estimators of α and β.

It follows from the derivation above that, if $\gamma = c_1\alpha + c_2\beta$ is any linear function of the parameters α and β, then the BLUE of γ is $\hat{\gamma} = c_1 a + c_2 b$. (In the normal case, $\hat{\gamma}$ is the UMVU estimator of γ.) In particular, the BLUE's of the expected values $E(Y_i) = \alpha + \beta x_i$ are the "fitted values" $a + bx_i$. Sometimes the primary purpose of estimating α and β is to predict the expected value of a future value of Y at $x = x_0$. The BLUE of $E(Y) = \alpha + \beta x_0$ is $a + bx_0$. Its variance is given in Exercise 2 below.

Exercises. 1. Show that, if $Y_1, Y_2, \ldots, Y_n$ are independent random variables having possibly different means but the same variance σ^2, then $Var(\Sigma c_i Y_i) = \sigma^2 \Sigma c_i^2$ and $Cov(\Sigma c_i Y_i, \Sigma d_i Y_i) = \sigma^2 \Sigma c_i d_i$.

2. Use part (a) to show that, if a and b are the BLUE's in the theorem above, then $Var(b) = \sigma^2/SS(x)$, $Var(a) = \sigma^2[(1/n) + \bar{x}^2/SS(x)]$, $Cov(a,b) = -\bar{x}\sigma^2/SS(x)$, and $Var(a+bx_0) = \sigma^2[(1/n) + (x_0 - \bar{x})^2/SS(x)]$ where $SS(x) = \Sigma(x_i - \bar{x})^2$.

3. The "residual" $\hat{e}_i$ corresponding to Y_i is defined by $\hat{e}_i = Y_i - a - bx_i$. Show that (a) $\Sigma\hat{e}_i = 0$, (b) $E(\hat{e}_i) = 0$, and (c) $Cov(\hat{e}_i, a) = Cov(\hat{e}_i, b) = 0$.

As a generalization of the "simple linear regression" model considered above, let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with means

$$E(Y_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

where $(x_{1i}, x_{2i}, \ldots, x_{pi})$ are given constants and the β_j's are unknown parameters. In addition, assume that the Y_i's have the same variance σ^2, and the columns of the matrix X below are linearly independent:[1]

$$X = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ & \cdot \quad \cdot \quad \cdot & & \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix}$$

To see that this model includes simple linear regression as a special case, set $x_{1i} = 1$ and $x_{2i} = x_i$ for $i = 1,2,\ldots,n$ where $x_1, x_2, \ldots, x_n$ are the values of the "independent variable." In this case the condition that the columns of X be linearly independent amounts to requiring that not all of the x_i's have the same value.

Let $b_1, b_2, \ldots, b_p$ denote the least-squares estimators of the parameters β_j, i.e., the b_j's are the values of the β_j's that minimize

$$SS = \sum_i (Y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_p x_{pi})^2.$$

Theorem 9-7. (Gauss-Markov Theorem.) Under the above assumptions the least squares estimators b_j are the BLUE's of the regression coefficients β_j, and if $\gamma = \Sigma c_j \beta_j$ is any linear combination of the β_j's, the BLUE of γ is $\hat{\gamma} = \Sigma c_j b_j$.

---

[1] The columns of X are said to be linearly dependent if there exist constants $a_1, a_2, \ldots, a_p$, not all of which are zero, such that $\Sigma_{j=1}^{p} a_j x_{ji} = 0$ for $i=1,2,\ldots,n$. In this case, one of the columns of X is a linear combination of the others.

A proof of this theorem, which is of fundamental importance in many applications, can be given by mimicking the proof above for the case of a single "independent variable" x. If the Y_i's have normal distributions, then the b_j's are the UMVU estimators of the β_j's and $\hat{\gamma}$ is the UMVU estimator of γ. In this case, the residual sum of squares SS_e, obtained by substituting the b_j's for the β_j's in SS above, is independent of the b_j's, and $SS_e/\sigma^2 \sim \chi^2(n-p)$. Whether the Y_i's are normally distributed or not, the estimator $SS_e/(n-p)$ is unbiased for σ^2.

Exercises. 1. Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random variables with mean β and variance σ^2. Show that the least-squares estimator of β is $\hat{\beta} = \bar{Y}$ and the residual sum of squares is $SS_e = \Sigma(Y_i - \bar{Y})^2$.

2. Consider the problem of comparing the means $\beta_1, \beta_2, \ldots, \beta_I$ of I populations on the basis of independent random samples of sizes $n_1, n_2, \ldots, n_I$ from the respective populations. Let Y_{ij} denote the j^{th} observation from the i^{th} population. Assuming that the observations Y_{ij} have the same variance, show that the least-squares estimators of the means are $\hat{\beta}_i = \bar{Y}_i$ where $\bar{Y}_i$ is the sample mean of the observations in the i^{th} group. Also, show that the BLUE of any linear combination of the means, $\Sigma c_i\beta_i$, is $\Sigma c_i\bar{Y}_i$ and find its variance. Ans. $\sigma^2\Sigma c_i^2/n_i$.

3. Consider the same situation as in Exercise 2 except that $E(Y_{ij}) = \beta_i + \gamma z_{ij}$ where the values z_{ij} are known constants. Show that the least-squares estimators of the parameters are

$$\hat{\gamma} = \sum_{i,j} (z_{ij} - \bar{z}_i)Y_{ij} / \sum_{i,j} (z_{ij} - \bar{z}_i)^2$$

and

$$\hat{\beta}_i = \bar{Y}_i - \hat{\gamma}\bar{z}_i.$$

Also, show that the variances of these estimators are given by

$$\mathrm{Var}(\hat{\gamma}) = \sigma^2/\Sigma(z_{ij} - \bar{z}_i)^2,$$

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2 \left\{ \frac{1}{n_i} + \frac{\bar{z}_i^2}{\Sigma(z_{ij} - \bar{z}_i)^2} \right\}.$$