NASA/TP—1998–207194

Probability and Statistics in Aerospace Engineering

M.H. Rheinfurth and L.W. Howell
Marshall Space Flight Center, Marshall Space Flight Center, Alabama

March 1998

The NASA STI Program Office…in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and mission, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results…even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***http://www.sti.nasa.gov***

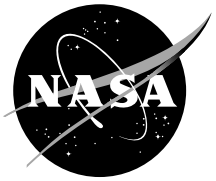
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA Access Help Desk at (301) 621–0134

- Telephone the NASA Access Help Desk at (301) 621–0390

- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  800 Elkridge Landing Road
  Linthicum Heights, MD 21090–2934

NASA/TP—1998–207194



Probability and Statistics in Aerospace Engineering

M.H. Rheinfurth and L.W. Howell
Marshall Space Flight Center, Marshall Space Flight Center, Alabama

National Aeronautics and
Space Administration

Marshall Space Flight Center

March 1998

Available from:

NASA Center for AeroSpace Information
800 Elkridge Landing Road
Linthicum Heights, MD 21090-2934
(301) 621–0390

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
(703) 487–4650

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TECHNICAL PUBLICATION

PROBABILITY AND STATISTICS IN AEROSPACE ENGINEERING

I. INTRODUCTION

A. Preliminary Remarks

Statistics is the science of the collection, organization, analysis, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. In engineering work this includes such different tasks as predicting the reliability of space launch vehicles and subsystems, lifetime analysis of spacecraft system components, failure analysis, and tolerance limits.

A common engineering definition of statistics states that statistics is the science of guiding decisions in the face of uncertainties. An earlier definition was statistics is the science of making decisions in the face of uncertainties, but the verb *making* has been moderated to *guiding*.

Statistical procedures can vary from the drawing and assessment of a few simple graphs to carrying out very complex mathematical analysis with the use of computers; in any application, however, there is the essential underlying influence of "chance." Whether some natural phenomenon is being observed or a scientific experiment is being carried out, the analysis will be statistical if it is impossible to predict the data exactly with certainty.

The theory of probability had, strangely enough, a clearly recognizable and rather definitive start. It occurred in France in 1654. The French nobleman Chevalier de Mere had reasoned falsely that the probability of getting at least one six with 4 throws of a single die was the same as the probability of getting at least one "double six" in 24 throws of a pair of dice. This misconception gave rise to a correspondence between the French mathematician Blaise Pascal (1623–1662) and his mathematician friend Pierre Fermat (1601–1665) to whom he wrote: "Monsieur le Chevalier de Mere is very bright, but he is not a mathematician, and that, as you know, is a very serious defect."

B. Statistical Potpourri

This section is a collection of aphorisms concerning the nature and concepts of probability and statistics. Some are serious, while others are on the lighter side.

"The theory of probability is at bottom only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it. It is remarkable that this science, which originated in the consideration of games of chance, should have become the most important object of human knowledge." (P.S. Laplace, 1749–1827)

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." (H.G. Wells, 1946)

From a file in the NASA archives on "Humor and Satire:" *Statistics is a highly logical and precise method of saying a half-truth inaccurately*.

A statistician is a person who constitutionally cannot make up his mind about anything and under pressure can produce any conclusion you desire from the data.

There are three kinds of lies: white lies, which are justifiable; common lies, which have no justification; and statistics.

From a NASA handbook on shuttle launch loads: "The total load will be obtained in a rational manner or by statistical analysis."

Lotteries hold a great fascination for statisticians, because they cannot figure out why people play them, given the odds.

There is no such thing as a good statistical analysis of data about which we know absolutely nothing.

Real-world statistical problems are almost never as clear-cut and well packaged as they appear in textbooks.

Statistics is no substitute for good judgment.

The probability of an event depends on our state of knowledge (information) and not on the state of the real world. Corollary: There is no such thing as the "intrinsic" probability of an event.

C. Measurement Scales

The types of measurements are usually called *measurement scales*. There exist four kinds of scales. The list proceeds from the "weakest" to the "strongest" scale and gives an example of each:

- Nominal Scale: Red, Green, Blue
- Ordinal Scale: First, Second, Third
- Interval Scale: Temperature
- Ratio Scale: Length.

Most of the nonparametric (distribution-free) statistical methods work with interval or ratio scales. In fact, all statistical methods requiring only a weaker scale may also be used with a stronger scale.

D. Probability and Set Theory

The formulation of modern probability theory is based upon a few fundamental concepts of set theory. However, in probability theory these concepts are expressed in a language particularly adapted to probability terminology. In order to relate the notation commonly used in probability theory to that of set theory, we first present a juxtaposition of corresponding terms, shown in table 1.

TABLE 1.—*Set theory versus probability theory terms.*

Set Vocabulary	Probability Vocabulary
(1) Element	Outcome (*E)* (Sample Point, Elementary Event)
(2) Subset	Event (*A*)
(3) Universal Set	Sample Space (*S*)
(4) Empty Set	Null Event (Φ)
(5) Disjoint	Mutually Exclusive
(6) Union $A \cup B$	"OR" Probability
(7) Intersection $A \cap B$	"AND" Probability

The probability theory terms in table 1 are defined as follows:

(1) An **outcome *E*** is defined as each possible result of an actual or conceptual experiment. Each experiment terminates with an outcome. An outcome is sometimes called a sample point or elementary event.

(2) An **event *A*** is defined as a set of outcomes. One declares that an event *A* has occurred if an outcome *E* of an experiment belongs to an element of *A*.

(3) The **sample space *S*** is defined as the set of all possible outcomes. It is also called the certain event.

(4) The **null event** Ø is defined as the set consisting of no outcomes. It is also called the impossible event.

(5) Two events *A* and *B* are called **mutually exclusive** if they have no common element. Note that outcomes are by definition mutually exclusive.

(6) The **union** of events *A* and *B* is the event that occurs if *A* occurs or/and *B* occurs.

(7) The **intersection** of events *A* and *B* is the event that *A* occurs and *B* occurs.

Two more definitions are used in probability theory with notations that are identical to that of set theory:

(8) The **complement** of an event *A*, written as $\overline{A}, A^c$, or A', is the event that occurs if *A* does not occur.

(9) The **difference** of events *A* and *B*, written as *A*–*B*, is the event that occurs if *A* occurs but *B* does not occur: $(A–B)=A \cap B'$.

EXAMPLE: Toss a die and observe the number that appears facing up. Let *A* be the event that an even number occurs, and *B* the event that a prime number occurs. Then we have in table 2:

TABLE 2.—*Examples of set theory.*

Sample Space	S={1,2,3,4,5,6}
Outcome	E={1},{2},{3},{4},{5},{6}
Event A	A={2,4,6}
Event B	B={2,3,5}
Union	$A \cup B$={2,3,4,5,6}
Intersection	$A \cap B$={2}
Complement	A'={1,3,5}
Difference	$A - B$ = {4,6}

1. Venn Diagrams

When considering operations on events it is often helpful to represent their relationships by so-called Venn Diagrams, named after the English logician John Venn (1834–1923).

2. Principle of Duality (De Morgan's Law)

The Principle of Duality is also known as De Morgan's Law after the English mathematician (1871). Any result involving sets is also true if we replace unions by intersections, intersections by unions, and sets by their complements. For example, $(A \cup B)' = A' \cap B'$.

II. PROBABILITY

A. Definitions of Probability

1. Classical (a Priori) Definition

The classical (a priori) definition of probability theory was introduced by the French mathematician Pierre Simon Laplace in 1812. He defined the probability of an event A as the ratio of the favorable outcomes to the total number of possible outcomes, provided they are equally likely (probable):

$$P(A)=\frac{n}{N} \tag{1}$$

where n=number of favorable outcomes and N=number of possible outcomes.

2. Empirical (a Posteriori) Definition

The empirical (a posteriori) definition was introduced by the German mathematician Richard V. Mises in 1936. In this definition, an experiment is repeated M times and if the event A occurs $m(A)$ times, then the probability of the event is defined as:

$$P(A)=\lim_{M\to\infty}\frac{m(A)}{M} \quad . \tag{2}$$

Empirical Frequency. This definition of probability is sometimes referred to as the *relative frequency*. Both the classical and the empirical definitions have serious difficulties. The classical definition is clearly circular because we are essentially defining probability in terms of itself. The empirical definition is unsatisfactory because it requires the number of observations to be infinite; however, it is quite useful in practice and has considerable intuitive appeal. Because of these difficulties, statisticians now prefer the axiomatic approach based on set theory.

3. Axiomatic Definition

The axiomatic definition was introduced by the Russian mathematician A.N. Kolmogorov in 1933:

- Axiom 1: $P(A)\geq 0$
- Axiom 2: $P(S)=1$
- Axiom 3: $P(A\cup B)=P(A)+P(B)$ if $A\cap B=\emptyset$.

It follows from these axioms that for any event A, then:

$$0\leq P(A)\leq 1 \quad . \tag{3}$$

Probabilities and Odds. If the probability of event A is p, then the *odds* that it will occur are given by the ratio of p to $1–p$. Odds are usually given as a ratio of two positive integers having no common factors. If an event is more likely to *not* occur than to occur, it is customary to give the odds that it will not occur rather than the odds that it will occur.

EXAMPLE:

- Probability: $P=A/B$ where A and B are any two positive numbers and $A \le B$.

  Odds: A: $(B–A)$.

  If the probability of an event is $p=3/4$, we say that the odds are 3:1 in its favor.

- Given the odds are A to B, then the probability is $A/(A+B)$.

Criticality Number. For high reliability systems it is often preferable to work with the probability of failure multiplied by 10^6. This is called the *criticality number*. For instance, if the system has a probability of success of $P=0.9999$, then the criticality number is $C=100$.

B. Combinatorial Analysis (Counting Techniques)

In obtaining probabilities using the classical definition, the enumeration of outcomes often becomes practically impossible. In such cases use is made of combinatorial analysis, which is a sophisticated way of counting.

1. Permutations

A permutation is an *ordered* selection of k objects from a set S having n elements.

- Permutations *without* repetition:

$$P_0(n,k) = {}_nP_k = n\times(n-1)\times(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!} \tag{4}$$

- Permutations *with* repetition:

$$P_1(n,k) = n^k \quad . \tag{5}$$

. Combinations

A combination is an *unordered* selection of k objects from a set S having n elements.

- Combinations *without* repetition:

$$C_0(n,k) = {}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{(called “binomial coefficient”)} \ . \tag{6}$$

- Combinations *with* repetition:

$$C_1(n,k)=\binom{n+k-1}{k}=\frac{(n+k-1)!}{k!(n-1)!} \quad . \tag{7}$$

EXAMPLES: Selection of two letters from {a, b, c}:

(1) $P_0(n,k)=P_0\,(3,2)=3\times2=6$ Without repetition

ab, *ac*, *ba*, *bc*, *ca*, *cb*

(2) $P_1(n,k)=P_1(3,2)=3\times3=9$ With repetition

aa, *ab*, *ac*, *ba*, *bb*, *bc*, *ca*, *cb*, *cc*

(3) $C_0(n,k)=C_0\,(3,2)=\frac{3\times2}{1\times2}=3$ Without repetition

ab, *ac*, *bc*

(4) $C_1(n,k)=C_1\,(3,2)=\frac{3\times4}{1\times2}=6$ With repetition

aa, *bb*, *cc*, *ab*, *ac*, *bc*

PROBLEM: Baskin-Robbins ice-cream parlors advertise 31 different flavors of ice cream. What is the number of possible triple-scoop cones without repetition, depending on whether we are interested in how the flavors are arranged or not?

SOLUTION: ${}_{31}P_3=26{,}970$ and ${}_{31}C_3=4{,}495$.

3. Permutations of a Partitioned Set

Suppose a set consists of n elements of which n_1 are of one kind, n_2 are of a second kind, $\ldots n_k$ are of a k^{th} type. Here, of course, $n=n_1+n_2+\ldots n_k$. Then the number of permutations is:

$$P_2(n,n_k)=\frac{n!}{n_1!n_2!n_3!\ldots n_k!} \quad . \tag{8}$$

An excellent reference for combinatorial methods is M. Hall's book *Combinatorial Analysis*.

PROBLEM: Poker is a game played with a deck of 52 cards consisting of four suits (spades, clubs, hearts, and diamonds) each of which contains 13 cards (denominations 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, and A.) When considered sequentially, the A may be taken to be 1 or A but not both; that is, 10, J, Q, K, A is a five-card sequence called a "straight," as is A, 2, 3, 4, 5; but Q, K, A, 2, 3 is not sequential, that is, not a "straight."

A poker hand consists of five cards chosen at random. A winning poker hand is the one with a higher "rank" than all the other hands.

A “flush” is a five-card hand all of the same suit. A “pair” consists of two, and only two, cards of the same kind, for example (J^S, J^C). “Three-of-a-kind” and “four-of-a-kind” are defined similarly. A “full house” is a five-card hand consisting of a “pair” and “three-of-a-kind.” The ranks of the various hands are as follows with the highest rank first:

- Royal flush (10, J, Q, K, A of one suit)
- Straight flush (consecutive sequence of one suit that is not a royal flush)
- Four-of-a-kind
- Full house
- Flush (not a straight flush)
- Straight
- Three-of-a-kind (not a full house)
- Two pairs (not four of a kind)
- One pair
- No pair (“bust”).

(1) Show that the number of possible poker hands is 2,598,960.

(2) Show that the number of possible ways to deal the various hands are:

(a) 4	(b) 36	(c) 624	(d) 3,744	(e) 5,108
(f) 10,200	(g) 54,912	(h) 123,552	(i) 1,098,240	(j) 1,302,540

SOLUTIONS: Poker is a card game with a deck of 52 cards consisting of four suits (spades, clubs, hearts, and diamonds). Each suit contains 13 denominations (2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, and A). A poker hand has five cards and the players bet on the ranks of their hands. The number of possible ways to obtain each of these ranks can be determined by combinatorial analysis as follows:

(a) Royal Flush. This is the hand consisting of the 10, J, Q, K, and A of one suit. There are four of these, one for each suit, and hence, $N=4$.

(b) Straight Flush. All five cards are of the same suit and in sequence, such as the 6, 7, 8, 9, and 10 of diamonds. Their total number is 10 for each suit. However, we have to exclude the four royal flushes contained in this set. Therefore, the total number of straight flushes is $N=10\times4-4=36$.

(c) Four-of-a-Kind. This hand contains four cards of the same denomination such as four aces or four sixes and then a fifth unmatched card. The total number of possible ways to choose this rank is obtained by:

(1) Choosing the denomination, 13 ways.

(2) Choosing the suit, 4 ways.

(3) Choosing the remaining unmatched card, 12 ways.

The result is: $N=13\times4\times12=624$.

(d) Full House. This hand consists of three cards of one denomination and two cards of another, as 8–8–8–K–K. The total number of possible ways is given by the following sequence of selections:

(1) Choosing denomination of the first triplet of cards, 13 ways.

(2) Selecting three out of the four suits for this triplet, $\binom{4}{3}=4$ ways.

(3) Choosing the denomination of the second doublet of cards, 12 ways.

(4) Selecting two out of the four suits for this doublet, $\binom{4}{2}=6$ ways.

The result is then $N=13\times4\times12\times6=3,744$.

(e) Flush. This hand contains five cards of the same suit, but not all in sequence. To obtain the number of possible ways, select 5 out of 13 denominations, $\binom{13}{5}=1,287$ ways, and select one of the four suits, 4 ways for a total of $N=1,287\times4=5,148$. Here we have to consider again that this number contains the number of straight flushes and royal flushes which have to be subtracted from it. The result is $N=5,148-36-4=5,108$.

(f) Straight. This hand contains a five-card sequence as defined above. We observe that there are 10 possible five-card sequences. Each face value in this sequence can come from any of the four denominations, which results in 4^5 different ways of creating a particular five-card sequence. The result is: $N=10\times4^5=10,240$. Again, it must be noted that this number contains the number of straight flushes and royal flushes which have to be subtracted for the final answer: $N=10,240-36-4=10,200$.

(g) Three-of-a-Kind. This hand contains three cards of one denomination and two different cards, each of different denominations. The total number of ways is obtained by:

(1) Choose the denomination of the three cards, 13 ways.

(2) Select three of the four suits in 4 ways.

(3) Select 2 of the remaining 12 denominations $\binom{12}{2}=66$ ways.

(4) Each of the two remaining cards can have any of the four denominations for $4\times4=16$.

The total number for this rank is, therefore, $N=13\times4\times66\times16=54,912$.

(h) Two Pairs. To obtain the number of possible ways for this rank, we take the following steps:

(1) Select the denomination of the two pairs in $\binom{13}{2}=78$ ways.

(2) Select the two suits for each pair in $\binom{4}{2}=36$ ways.

(3) Select the denomination of the remaining card. There are 11 face values left.

(4) The remaining card can have any of the four suits.

The total number is, therefore, $N=78\times36\times11\times4=123{,}552$.

(i) One Pair. The number of possible ways for this rank is obtained according to the following steps:

(1) Select denomination of the pair in 13 ways.

(2) Select suit in $\binom{4}{2}=6$ ways.

(3) Select denomination of the other three cards from the remaining 12 denominations in $\binom{12}{3}=220$ ways.

(4) Each of these three cards can have any suit, resulting in $4^3=64$ ways.

The total number is then $N=13\times6\times220\times64=1{,}098{,}240$.

(j) No Pair. The number of ways for this "bust" rank is obtained according to the following steps:

(1) Select five cards from 13 denominations as $\binom{13}{5}=1{,}287$.

(2) Each card can have any suit, giving $4^5=1{,}024$.

The result is $N=1{,}287\times1{,}024=1{,}317{,}888$. Again, we note that this number contains the number of royal flushes, straight flushes, flushes, and straights. Thus, we obtain as the answer:

$$N=1{,}317{,}888-4-36-5{,}108-10{,}200=1{,}302{,}540 \quad .$$

QUESTION: The Florida State Lottery requires the player to choose 6 numbers without replacement from a possible 49 numbers ranging from 1 to 49. What is the probability of choosing a winning set of numbers? Why do people think a lottery with the numbers 2, 13, 17, 20, 29, 36 is better than the one with the numbers 1, 2, 3, 4, 5, 6? (Hint: use hypergeometric distribution.)

C. Basic Laws of Probability

1. Addition Law ("OR" Law; "AND/OR")

The Addition Law of probability is expressed as:

$$P(A\cup B)=P(A)+P(B)-P(A\cap B) \quad . \tag{9}$$

The Venn diagram in figure 1 is helpful in understanding this probability law. A formal proof can be found in any standard textbook. In Venn diagrams, the universal set S is depicted by a rectangle and the sets under consideration by closed contours such as circles and ellipses.

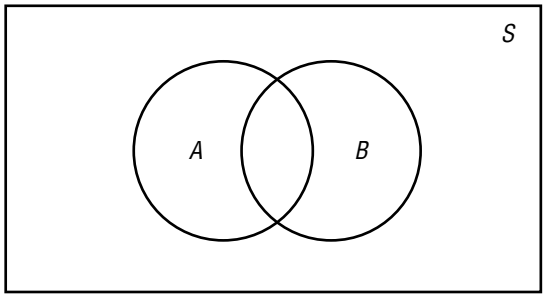


FIGURE 1.—Venn diagram.

GENERAL RULE:

n=3: $$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C) \quad (10)$$

n arbitrary (obtainable by mathematical induction):

$$P(A_1\cup A_2\cup\ldots\cup A_k)=\sum_1^k P(A_i)-\sum_{i<j=2}^{k} P(A_i\cap A_j)+\sum_{i<j<r=3}^{k} P(A_i\cap A_j\cap A_r)$$

$$+(-1)^{k-1}P(A_1\cap A_2\cap A_3\cap\ldots\cap A_k) \quad . \quad (11)$$

2. Conditional Probability

Since the choice of sample space is not always self-evident, it is often necessary to use the symbol $P(A\,|\,B)$ to denote the *conditional probability* of event A relative to the sample space B, or the probability of A given B. Assuming equal probability for the outcomes in A and B, we can derive the relationship shown in figure 2.

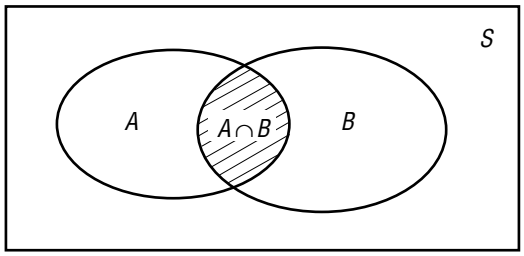


FIGURE 2.—Conditional probability.

Given the number of outcomes in sample space B as $N(B)$, the number of outcomes in sample space $A\cap B$ as $N(A\cap B)$, and the number of outcomes in the sample space S as $N(S)$, we obtain $P(A\,|\,B)$ using the classical definition of probability:

$$P(A\,|\,B)=\frac{N(A\cap B)}{N(B)}=\frac{N(A\cap B)/N(S)}{N(B)/N(S)} \quad . \quad (12)$$

The second term, on the right-hand side, was obtained by dividing the numerator and denominator by $N(S)$. The sample space B is called the *reduced sample space*.

We can now write the above equation in terms of two probabilities defined on the total sample space S:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad . \tag{13}$$

Generalizing from this example, we introduce the following formal definition of conditional probability: If A and B are any two events in a sample space S and $P(B) \neq 0$, the conditional probability of A given B is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad . \tag{14}$$

3. Multiplication Rule ("AND" Law)

Multiplying the above expression of the conditional probability by $P(B)$, we obtain the following multiplication rule:

$$P(A \cap B) = P(B)\, P(A \mid B) \quad . \tag{15}$$

The second rule is obtained by interchanging letters A and B. This rule can be easily generalized to more than two events; for instance, for three events A, B, and C we have:

$$P(A \cap B \cap C) = P(A)\, P(B \mid A)\, P(C \mid A \cap B) \quad . \tag{16}$$

If A and B are two events, we say that A is *independent* of B if and only if $P(A|B)=P(A)$. In other words, the occurrence of event B does not affect the probability of event A. It can be shown that B is independent of A whenever A is independent of B. Therefore, we can simply state that A and B are independent events.

EXAMPLE: Two people often decide who should pay for the coffee by flipping a coin. To eliminate the problem of a biased coin, the mathematician John V. Neumann (1903–1957) devised a way to make the odds more "even" using the multiplication law.

The coin is flipped twice. If it comes up heads both times or tails both times, the process is repeated. If it comes up heads-tail, the first person wins, and if it comes up tail-heads the second person wins. The probabilities of these outcomes are the same even if the coin is biased.

For example, if the probability of heads is $P(H)=0.6$ and that of tails is $P(T)=0.4$, the intersection of the two events $P(H \cap T)=P(T \cap H)=0.6 \times 0.4=0.24$.

4. Event-Composition Method

This approach calculates the probability of an event A by expressing it as a composition (unions and/or intersections) of two or more other events.

PROBLEM 1: Use the addition and multiplication laws of probability to simplify the expression:

$$P[(B \cap C) \cup (D \cap E)] \quad . \tag{17}$$

SOLUTION: Applying the addition law, we obtain:

$$P[(B\cap C)\cup(D\cap E)]=P(B\cap C)+P(D\cap E)-P(B\cap C\cap D\cap E)\ . \tag{18}$$

Observe that the events on the right are intersections and this calls for the application of the multiplication law. Therefore,

$$P[(B\cap C)\cup(D\cap E)]=P(B)\ P(C|B)+P(D)\ P(E|D)-P(B)\ P(C|B)\ P(D|B\cap C)\ P(E|B\cap C\cap D)\ . \tag{19}$$

It is frequently desirable to form compositions of mutually exclusive or independent events, because they simplify the addition and multiplication laws.

PROBLEM 2: It is known that a patient will respond to a treatment of a particular disease with a probability of 0.9. If three patients are treated in an independent manner, determine the probability that at least one will respond.

SOLUTION: Define the events:

A=At least one patient will respond.
$A_i=i^{\text{th}}$ patient will respond (i=1, 2, 3).

The event $A=A_1\cup A_2\cup A_3$.

Now we observe by the law of duality that the complementary event A' is $A_1'\cap A_2'\cap A_3'$ and that $S=A\cup A'$. Then, because $P(S)$=1 and the independence of the events A_i we have:

$$\begin{gathered}P(A)=1-P(A')=1-P(A_1')\times P(A_2')\times P(A_3')\\ P(A)=1-0.1\times 0.1\times 0.1=0.999\ .\end{gathered} \tag{20}$$

This result is of importance because it is often easier to find the probability of the complementary event A' than of the event A itself. This is always the case for problems of the "at-least-one" type, as is this one.

PROBLEM 3: Birthday Problems: (a) What is the probability that in a randomly selected group of n persons, two or more of them will share a birthday?

SOLUTION: In solving this problem, we assume that all birthdays are equally probable (uniform distribution). We also discount leap years. These assumptions are, of course, not quite realistic. Again, it is advantageous to first find the complementary event that all persons have different birthdays.

The first of the n persons has, of course, some birthday with probability 365/365=1. Then, if the second person is to have a different birthday, it must occur on one of the other 364 days. Thus the probability that the second person has a birthday different from the first is 364/365. Similarly the probability that the third person has a different birthday from the first two is 363/365 and so on.

The probability of the complementary event A' is, therefore:

$$P(A')=(365/365)\times(364/365)\times\ldots((365-n+1)/365)\ . \tag{21}$$

The desired probability of the event A is, then:

$$P(A)=1-P(A')\ . \tag{22}$$

For n=23: $P(A)$=0.5073 and for n=40: $P(A)$=0.891.

(b) What is the probability that in a randomly selected group of n persons, at least one of them will share a birthday *with you*?

SOLUTION: The probability that the second person has a birthday different from you is, of course, the same as above, namely 364/365. However, the probability that the third person has a different birthday from yours is, again, 364/365 and so on.

The probability of the complementary event A' is, therefore:

$$P(A')=(364/365)^{n-1}\ . \tag{23}$$

The desired probability of event A is, then, again:

$$P(A)=1-P(A')\ . \tag{24}$$

For n=23: $P(A)$=0.058 and for n=40: $P(A)$=0.101.

PROBLEM 4: Three cards are drawn from a deck of 52 cards. Find the probability that two are jacks and one is a king.

SOLUTION: p=(3!/2!)×(4/52×3/51)×(4/50)=6/5525.

5. Total Probability Rule

Let B_1, B_2, …, B_k form a partition of the sample space S (see fig. 3). That is, we have

$$B_i \cap B_j = \emptyset \text{ for all } i \neq j \tag{25}$$

and

$$B_1 \cup B_2 \cup \ldots \cup B_k = S\ . \tag{26}$$

The events B_i are mutually exclusive and exhaustive (see fig. 3).

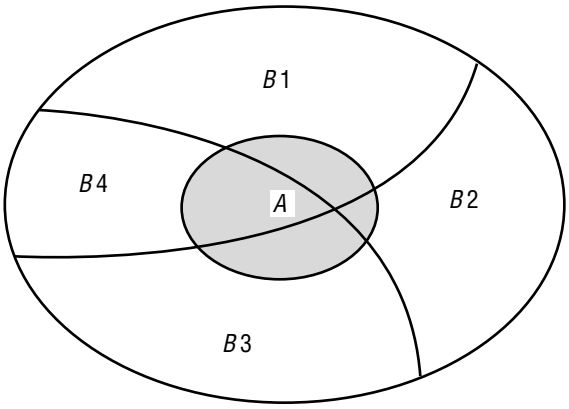


FIGURE 3.—Partitioned sample space.

The total probability for any event A in S is:

$$P(A)=\sum_{i=1}^{k} P(A\cap B_i)=\sum_{i=1}^{k} P(B_i)\times P(A\mid B_i)\ . \tag{27}$$

6. Bayes' Rule

Bayes' Rule was published in 1763 by Thomas Bayes. Bayes' formula (see fig. 4) finds the probability that the event A was "caused" by the events B_i (inverse probabilities).

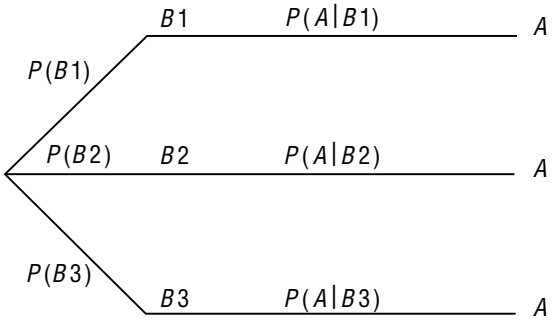


FIGURE 4.—Bayes' Rule.

Let B_i form a partition of S:

$$P(B_i\mid A)=\frac{P(A\cap B_i)}{P(A)}=\frac{P(A\mid B_i)\times P(B_i)}{P(A)}\ . \tag{28}$$

Substituting in the denominator:

$$P(A)=\sum P(A\cap B_i)=\sum P(A\,|\,B_i)\times P(B_i)\ \ , \tag{29}$$

we obtain Bayes' formula as:

$$P(B_i\,|\,A)=\frac{P(A\,|\,B_i)\,P(B_i)}{\sum_{i=1}^{k} P(A\,|\,B_i)\,P(B_i)}\ \ . \tag{30}$$

The unconditional probabilities $P(B_i)$ are called the "prior" probabilities. The conditional probabilities $P(B_i\,A)$ are called the "posterior" probabilities.

PROBLEM 1: Suppose a person with AIDS is given an HIV test and that the probability of a positive result is 0.95. Furthermore, if a person without AIDS is given an HIV test, the probability of an incorrect diagnosis is 0.01. If 0.5 percent of the population has AIDS, what is the probability of a healthy person being falsely diagnosed as being HIV positive?

SOLUTION: Define the events:

A=Person has AIDS
B=HIV test is positive

$P(B\,|\,A)$=0.95 $\qquad$ $P(B\,|\,A')=0.01$
$P(A)$=0.005 $\qquad$ $P(A')=0.995$

$$P(A'\,|\,B)=\frac{P(B\,|\,A')\times P(A')}{P(B\,|\,A')\times P(A')+P(B\,|\,A)\times P(A)}=\frac{(0.01)(0.995)}{(0.01)(0.995)+(0.95)(0.005)}=0.6768\ \ . \tag{31}$$

MESSAGE: False positive tests are more probable than the true positive tests when the overall population has a low incidence of the disease. This is called the false-positive paradox.

NOTE 1: In the medical profession $P(A)$ is called the base rate and the event $B\,|\,A'$ is called a false positive test (later to be defined as a type I error) and $B'\,|\,A$ is called a false negative test (type II error).

NOTE 2: Often it is said that a test proves to be 95 percent reliable or accurate. This statement means that $P(B|A)=P(B'\,|\,A')$=0.95, but it is more precise in this case to call the test 95 percent accurate in both directions.

AUXILIARY QUESTION: What is the probability of a person with AIDS being incorrectly diagnosed as not being HIV positive (type II error)?

$$P(A\,|\,B')=\frac{P(B'\,|\,A)\times P(A)}{P(B'\,|\,A)\times P(A)+P(B'\,|\,A')\times P(A')}=\frac{(0.05)(0.005)}{(0.05)(0.005)+(0.99)(0.995)}=2.537\times 10^{-4}\ \ . \tag{32}$$

SOLUTION: False negative tests are highly improbable.

PROBLEM 2: Suppose you are on a game show (Let's Make a Deal), and you are given the choice of three doors. Behind one door is a car; behind the other two, goats. You pick a door, say #1, and the host

(Monty Hall), who knows what is behind the doors, opens another door, say #3, which has a goat. He then says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice?

SOLUTION: Define events:

- D_i=Car is behind door D_i(i=1, 2, 3)
- H_3=Host opens door #3.

Define conditional probabilities:

- $P(D_2\,|\,H_3)$=Probability that car is behind door #2, given host has opened door #3
- $P(H_3\,|\,D_i)$=Probability that host opens door #3, given that car is behind door #1.

Apply Bayes' formula:

$$P(D_2\,|\,H_3)=\frac{P(H_3\,|\,D_2)\times P(D_2)}{P(H_3\,|\,D_1)\times P(D_1)+P(H_3\,|\,D_2)\times P(D_2)+P(H_3\,|\,D_3)\times P(D_3)}\ . \tag{33}$$

Prior probabilities that car is behind door #1:

$$P(D_1)=P(D_2)=P(D_3)=1/3\ . \tag{34}$$

Conditional probabilities:

$$P(H_3\,|\,D_1)=1/2 \text{ (some set this to unknown } q\text{)},\ P(H_3\,|\,D_2)=1,\ P(H_3\,|\,D_3)=0\ . \tag{35}$$

Posterior probability:

$$P(D_2\,|\,H_3)=\frac{(1)\times(1/3)}{(1/2)\times(1/3)\times(1)\times(1/3)+(0)\times(1/3)}=2/3\ . \tag{36}$$

Therefore, it is of advantage to switch.

PROBLEM 3: We consider 10 successive coin tosses. If the coin is "fair" and the events are assumed to be independent, then the probability of obtaining 10 heads in a row is clearly given as $P_{10}=(1/2)^{10}=1/1024$. What is the probability that the 11th toss will be a head?

SOLUTION: With the above assumptions, the answer is obviously one-half. Sometimes it is said that the coin has no "memory." (There are some gamblers who will bet on tails because of the law of averages, thinking it acts like a rubber band: the greater the deviation, the greater the restoring force towards the mean. This is also known as the "gambler's fallacy".) However, it is more natural to think that the longer the run of heads lasts, the more likely it is that our assumption of the coin being "fair" is not true. Let us expand this train of thought by defining the following situation:

Definition of events:

F=Coin is unbiased
B=Coin is biased
H=Next toss will be heads.

Definition of probabilities:

$$\begin{aligned} P(F)&=0.90 \qquad & P(B)&=1-P(F)=0.10 \\ P(H\mid F)&=0.50 \qquad & P(H\mid B)&=0.70 \ . \end{aligned} \tag{37}$$

Applying the total probability rule we obtain:

$$P(H)=P(H|F)\times P(F)+P(H|B)\times P(B)=(0.50)(0.90)+(0.70)(0.10)=0.52 \ . \tag{38}$$

Applying Bayes' theorem, we can update our prior probabilities $P(F)$ and $P(B)$ after we have observed 10 consecutive heads as follows:

Define event of 10 consecutive heads as H_{10}. Thus we obtain:

$$P(B\mid H_{10})=\frac{P(H_{10}\mid B)P(B)}{P(H_{10}\mid B)P(B)+P(H_{10}\mid F)P(F)}=\frac{(0.7^{10})(0.10)}{(0.7^{10})(0.1)+(0.5^{10})(0.9)}=0.763 \ . \tag{39}$$

Similarly, we obtain for:

$$P(F\mid H_{10})=1-P(B\mid H_{10})=0.237 \ . \tag{40}$$

We observe that the experiment has resulted in an increase of the probability that the coin is biased and a corresponding decrease that it is "honest." As mentioned before, the real problem lies in the assignment of the prior probabilities.

NOTE: Many objections to Bayes' theorem are actually attacks on Bayes' postulate (also called the Principle of Indifference or Principle of Insufficient Reason), that says if we have no knowledge of the prior probabilities, we may assume them to be equally probable. In our case we would set $P(B)=P(F)=0.5$, which is, of course, a very dubious assumption.

AUXILIARY QUESTIONS: (Solutions are left as a challenge to the reader.)

(1) A military operation consists of two independent phases. Phase A has a 10-percent risk and phase B, a 20-percent risk. (Risk is defined as the probability of failure.) What is the probability of mission failure? Answer: 0.28 .

(2) In a straight poker hand, what is the probability of getting a full house? Answer: $\frac{6}{4165}$.

(3) Two prizes are awarded in a lottery consisting of 200 tickets. If a person buys two tickets, what is the probability of winning the first prize or the second prize, but not both (exclusive "OR")? Answer: 0.0198995 .

(4) A student recognizes five different questions that may be asked on a quiz. However, he has time to study only one of them that he selects randomly. Suppose the probability that he passes the test if "his" question appears is 0.90, but the probability that he passes the test if it does not appear is only 0.30. The test contains only one question and it is one of the five.

(a) What are the chances that he will pass the test? Answer: 0.42 .

(b) If the student passed the test, what is the probability that "his" question was asked on the test? Answer: 0.4286 .

(5) A man has two pennies—one "honest" and one two-headed. A penny is chosen at random, tossed, and observed to come up heads. What is the probability that the other side is also a head? Answer: 2/3 .

D. Probability Distributions

In order to define probability distributions precisely, we must first introduce the following auxiliary concepts.

1. Set Function

We are all familiar with the concept of a function from elementary algebra. However, a precise definition of a function is seldom given. Within the framework of set theory, however, it is possible to introduce a generalization of the concept of a function and identify some rather broad classification of functions.

An *ordered pair* consists of two elements, say a and b, in which one of them, say a, is designated as the first element and the other the second element.

The *Cartesian product* $A \times B$ of two sets A and B is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$. For instance, if A=(a, b, c) and B=(1, 2, 3, 4) then the *Cartesian product* $A \times B$ is illustrated by the following area. Only the enclosed area F represents a function as defined below in figure 5.

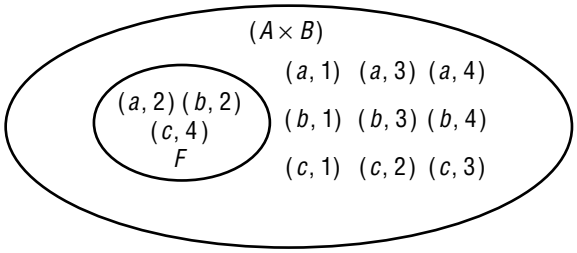


FIGURE 5.—Cartesian product.

A *relation* R from a set A into a set B is a subset of $A \times B$.

A *function* F from a set A into a set B is a subset of $A \times B$ such that *each* $a \in A$ appears only *once* as the first element of the subset (see fig. 6).

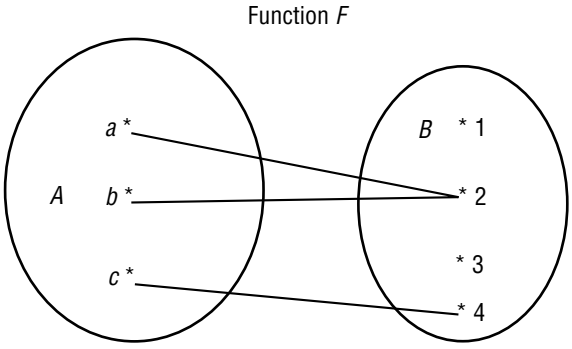


FIGURE 6.—Function $A \rightarrow B$.

The *domain* is the set of all elements of A which are related to at least one element of B.

The *range* is the set of all elements of B which are related to at least one element of A.

EXAMPLE: Supermarket:

A=Products (domain): x
B=Price (range): $y \Rightarrow y=y\,(x)$.

A *set function* is a function where the elements of the domain are sets and the elements of the range are numbers.

2. Random Variable

It is often desirable to assign numbers to the nonnumerical outcomes of a sample space. This assignment leads to the concept of a *random variable*. A random variable X is a set function which assigns to each outcome $E \in S$ a real number $X(E)=x$.

The domain of this set function is the sample space S and its range is the set of real numbers. In general, a random variable has some specified physical, geometrical, or other significance. It is important to observe the difference between the random variable itself and the value it assumes for a typical outcome.

It has become common practice to use capital letters for a random variable and small letters for the numerical value that the random variable assumes after an outcome has occurred. Informally speaking, we may say that a random variable is the name we give an outcome *before* it has happened. It may be that the outcomes of the sample space are themselves real numbers, such as when throwing a die. In such a case, the random variable is the identity function $X(E)=E$.

Note that, strictly speaking, when we are throwing a die the outcomes are not numerical, but are the "dot patterns" on the top face of the die. It is, of course, quite natural, but really not necessary, to associate the face values with the corresponding real number.

3. Probability Function and Cumulative Distribution Function

EXAMPLE: A coin is tossed three times. The sample space then consists of eight equally probable outcomes: $S=\{HHH\ldots TTT\}$. Let X be the random variable that counts the number of heads in each outcome.

Thus, X has range $\{0, 1, 2, 3\}$. Figure 7 lists the value x that the random variable X assumes for each outcome and the probability associated with each value x.

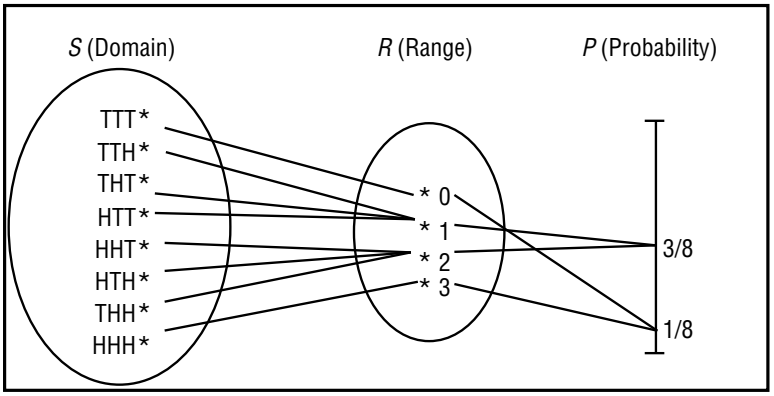


FIGURE 7.—Coin-tossing experiment.

Note that different outcomes may lead to the same value of x.

A *probability function* is a function that assigns to each value of the random variable X the probability of obtaining this value:

$$f(x)=P(X=x) \qquad 0\le f(x)\le 1 \ . \tag{41}$$

For example, $f(1)=P(X=1)=3/8$.

The mathematical function for $f(x)$ is $f(x)=\frac{1}{8}\binom{3}{x}$ (see fig. 8).

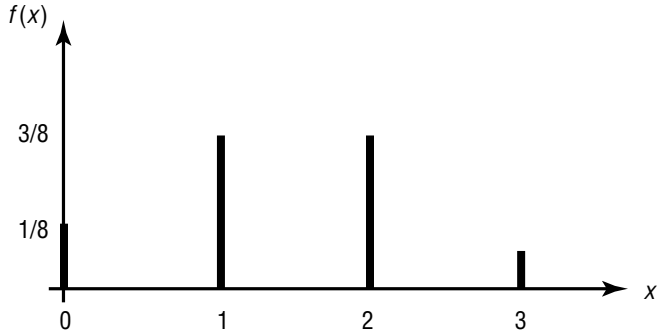


FIGURE 8.—Probability function diagram.

A cumulative distribution function (c.d.f.) is a function that assigns to each value of the random variable X the probability of obtaining *at most* this value. It is sometimes called "or less" cumulative distribution:

$$P(X\le x)=\sum_{t\le x} f_X(t)=F_X(x) \ . \tag{42}$$

Occasionally the complementary cumulative distribution function (c.c.d.f.) is also called the "or more" c.d.f. is used (i.e., nuclear power industry). One is easily obtained from the other.

Because of its appearance, the step function is also called *staircase function* (see fig. 9). Note that the value at an integer is obtained from the higher step, thus the value at 1 is 4/8 and not 1/8. As we proceed from left to right (i.e., going "upstairs") the distribution function either remains the same or increases, taking values between 0 and 1. Mathematicians call this a *monotonically increasing* function.

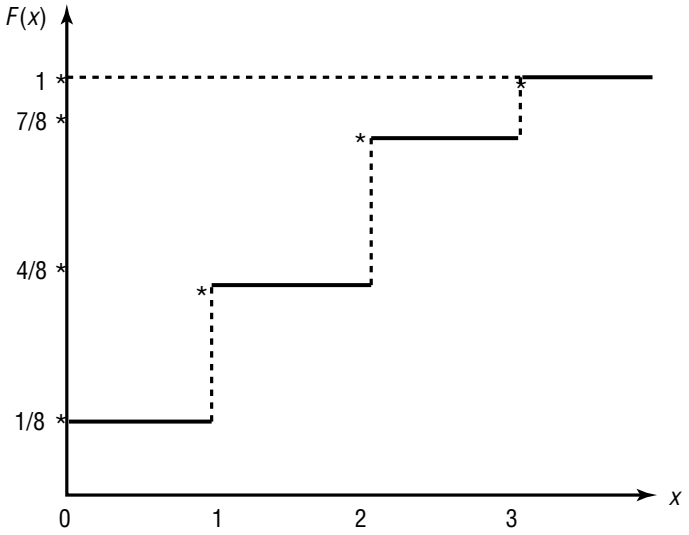


FIGURE 9.—Cumulative distribution function.

In general $f(x)$ is a probability function if:

$$0 \le f(x) \le 1 \tag{43}$$

$$\sum_{\{x\}} f_x = 1 \ . \tag{44}$$

4. Continuous Random Variable and Probability Density Function

Examples of continuous random variable and probability density function are temperature, age, and voltage. These ideas can be extended to the case where the random variable X may assume a continuous set of values. By analogy we define the cumulative distribution function for a continuous random variable by:

$$P(X \le x) = \int_{-\infty}^{x} f_x(t)dt = F_x(x) \ . \tag{45}$$

The rate of change of $P(X \le x)$ with respect to the random variable (probability/interval) is called the probability density function (p.d.f.):

$$\frac{dP(X \le x)}{dx} = \frac{dF(x)}{dx} = f(x) \ . \tag{46}$$

Note that the probability density function is *not* a probability. The probability density function has the following properties:

$$f(x) \ge 0 \tag{47}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \ . \tag{48}$$

Also note that the probability of a null event (impossible event) is zero. The converse is not necessarily true for a continuous random variable, i.e., if X is a continuous random variable, then an event having probability zero can be a possible event. For example, the probability of the possible event that a person is exactly 30 years old is zero.

Alternate definitions:

$$P(x<X<x+dx)=f(x)\,dx \tag{49}$$

$$P(a<X<b)=\int_a^b f(x)dx=F(b)-F(a)\ . \tag{50}$$

E. Distribution (Population) Parameters

One of the purposes of statistics is to express the relevant information contained in the mass of data by means of a relatively few salient features that characterize the distribution of a random variable. These numbers are called distribution or population parameters. We generally distinguish between a known parameter and an estimate thereof, based on experimental data, by placing a hat symbol (^) above the estimate. We usually designate the population parameters by Greek letters. Thus, $\hat{\mu}$ denotes an estimate of the population parameter μ. In the following, the definitions of the population parameters will be given for continuous distributions. The definitions for the discrete distribution are obtained by simply replacing integration by appropriate summation.

1. Measures of Central Tendency (Location Parameter)

a. Arithmetic Mean (Mean, Average, Expectation, Expected Value). The mean is the most often used measure of central tendency. It is defined as:

$$\mu=\int_{-\infty}^{\infty} xf(x)dx\ . \tag{51}$$

The definition of the mean is mathematically analogous to the definition of the center of mass in dynamics. That is the reason the mean is sometimes referred to as the first moment of a distribution.

b. Median (Introduced in 1883 by Francis Galton). The median m is the value that divides the total distribution into two equal halves, i.e.,

$$F(m)=\int_{-\infty}^{m} f(x)dx=1/2\ . \tag{52}$$

c. Mode (Thucydides, 400 B.C., Athenian Historian). This is also called the most probable value, from the French “mode,” meaning “fashion.” It is given by the maximum of the distribution, i.e., the value of x for which:

$$\frac{df(x)}{dx}=0\ . \tag{53}$$

If there is only such maximum, the distribution is called unimodal; if there are several, it is called multimodal.

In a symmetrical unimodal distribution, the mean, median, and mode coincide (see fig. 10). The median and the mode are useful measure for asymmetric (skewed) distributions. The median is commonly used to define the location for the distribution of family incomes, whereas the mode is used by physicians to specify the duration of a disease.

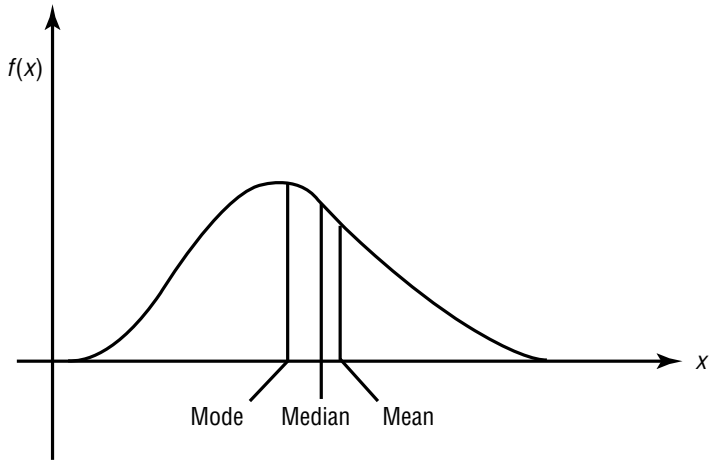


FIGURE 10.—Location of mean, median, and mode.

It is useful to remember that the mean, median, and mode of a unimodal distribution occur in reverse alphabetical order for a distribution that is skewed to the right, as in figure 10, or in alphabetical order for a distribution that is skewed to the left. What gives the mean the great importance in statistical theory is its mathematical tractability and its sampling properties.

2. Measures of Dispersion

The degree to which the data tend to spread about the mean value is called the dispersion or variability. Various measures of this dispersion are given below:

(a) Variance (Introduced by R.A. Fisher in 1918):

$$\sigma^2 = \int (x-\mu)^2 f(x)dx = \int x^2 f(x)dx - \mu^2 = E(x^2) - \{E(x)\}^2 \quad . \tag{54}$$

(b) Standard deviation (Introduced by K. Pearson in 1890): The standard deviation is simply the positive square root of the variance, denoted by σ.

(c) Mean deviation:

$$\sigma_d = \int |x-\mu| f(x)dx \quad . \tag{55}$$

For a normal distribution, the mean deviation =4/5 σ.

(d) Coefficient of variation (c.o.v.): The c.o.v. gives the standard deviation relative to the mean μ as:

$$\text{c.o.v.} = \sigma/\mu \quad . \tag{56}$$

It is a nondimensional number and is often expressed as a percentage. Note that it is independent of the units used.

3. Quantiles

The quantile of order p is the value ξ_p such that $P(X \le \xi_p) = p$.

For p=0.5 Median
p=0.25 Quartile
p=0.10 Decile
p=0.01 Percentile.

The jth quantile is obtained by solving for x:

$$j/n = \int_{-\infty}^{x} f(x)dx \quad . \tag{57}$$

Quantiles are sometimes used as measures of dispersion:

(a) Interquartile range $Q = \xi_{0.75} - \xi_{0.25}$

(b) Semi-interquartile range $Q_2 = 0.5 \times (\xi_{0.75} - \xi_{0.25})$

(c) Interdecile range $Q_{10} = \xi_{0.90} - \xi_{0.10}$

For a normal distribution, the semi-interquartile range =2/3 σ.

4. Higher Moments

Other important population parameters referred to as "higher moments" are given below:

(a) Skewness:

$$\alpha_3 = \mu_3/\sigma^3, \text{where } \mu^3 = \int (x-\mu)^3 f(x)dx \quad . \tag{58}$$

A distribution has a positive skewness (is positively skewed) if its long tail is on the right and a negative skewness (is negatively skewed) if its long tail is on the left.

(b) Kurtosis:

$$\alpha_4 = \mu_4/\sigma^4, \text{where } \mu_4 = \int (x-\mu)^4 f(x)dx \quad . \tag{59}$$

Kurtosis measures the degree of "peakedness" of a distribution, usually in comparison with a normal distribution, which has the kurtosis value of 3.

(c) Moments of *k*th order:

Moments about the origin (raw moments):

$$\mu'_k = E(x^k) = \int x^k f(x)dx \quad . \tag{60}$$

Moments about the mean (central moments):

$$\mu_k = E(x-\mu)^k = \int (x-\mu)^k \, f(x)dx \quad . \tag{61}$$

F. Chebyshev's Theorem

If a probability distribution has the mean μ and the standard deviation σ, the probability of obtaining a value that deviates from the mean by *at least k* standard deviations is *at most* $1/k^2$ (see fig. 11). Expressed in mathematical form:

$$P(|X-\mu| \geq k\,\sigma) \geq 1/k^2 \quad \text{(upper bound)} \tag{62}$$

$$P(|X-\mu| \leq k\,\sigma) \geq 1-1/k^2 \quad \text{(lower bound)} \quad . \tag{63}$$

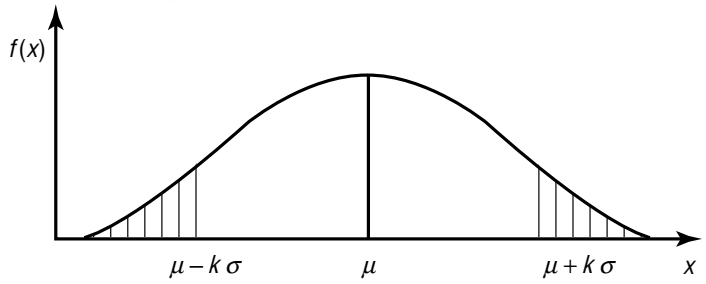


FIGURE 11.—Chebyshev's theorem.

Chebyshev's theorem is an example of a distribution-free method; i.e., it applies to any distribution with the only condition that its mean and standard deviation exist. This is its strength but also its weakness, for it turns out that the upper and lower bounds are too conservative to be of much practical value (see table 3 for a comparison with the k-factors derived from the normal distribution).

TABLE 3.—*Normal distribution compared with Chebyshev's theorem.*

Percentage	Normal *K*-factor	Chebyshev *K*-factor
90.0	1.65	3.16
95.0	1.96	4.47
99.0	2.58	10.0
99.9	3.29	31.6

The so-called Camp-Meidel correction for symmetrical distributions is only a slight improvement. It replaces the upper bound by $1/(2.25\times k^2)$.

G. Special Discrete Probability Functions

1. Binomial Distribution

Assumptions (Bernoulli trials) of binomial distribution are:

- There are only two possible outcomes for each trial ("success" and "failure")
- The probability p of a success is constant for each trial.
- There are n independent trials.

The random variable X denotes the number of successes in n trials. The probability function is then given by:

$$f(x)=P(X=x)=\binom{n}{x}p^x(1-p)^{n-x} \text{ for } x=0,1\ldots n \tag{64}$$

Mean $\mu=np$

Variance $\sigma^2=npq$ where $q=1-p$

Skewness $\alpha_3=\dfrac{q-p}{\sqrt{npq}}$

Kurtosis $\alpha_4=3+\dfrac{1-6pq}{\sqrt{npq}}$.

(Note that any of the subsequent BASIC programs can be readily edited to run on any computer.)

Binomial probabilities can be calculated using the following BASIC program, which is based on the recursion formula:

$$f(x)=\frac{p}{q}\times\frac{n-x}{x}\times f(x-1) \quad . \tag{65}$$

```
10:   "CUMBINOMIAL"
15:   CLEAR:INPUT "N="; N, "P="; P,"X="; X
20:   Q=1–P:F=Q^N:B=F:S=0
25:   IF X=0 GOTO 55
30:   FOR I=1 TO X
35:   E=P*(N–I+1)/Q/I
40:   F=F*E:S=S+F
45:   NEXT I
50:   CF=S+B
55:   PRINT CF:PRINT F
```

EXAMPLE: N=6, P=0.30, x=3

CF=0.9295299996, F=1.852199999E–01.

Simulation of Bernoulli trials can be performed using the following BASIC program:

```
100:  "BINOMIAL"
105:  INPUT "N=";N, "P=";P

110:  S=0
115:  FOR I=1 TO N
120:  U=RND .5
125:  IF U≤P LET X=1 GOTO 135
130:  X=0
135:  S=S+X
140:  NEXT I

145:  PRINT S: GOTO 110
```

2. Poisson Distribution

The Poisson distribution is expressed as:

$$f(x)=\frac{\mu^x e^{-\mu}}{x!} \quad \text{for } x=0, 1, 2, 3 \ldots\infty \tag{66}$$

Mean: $\mu=\mu$

Variance: $\sigma^2=\mu$

The Poisson distribution is a limiting form of the binomial distribution when $n\to\infty$, $p\to 0$, and $np=\mu$ remain constant. The Poisson distribution provides a good approximation to the binomial distribution when $n>20$ and $p<0.05$.

The Poisson distribution has many applications that have no direct connection with the binomial distribution. It is sometimes called the distribution of rare events; i.e., “number of yearly dog bites in New York City.” It might be of historical interest to know that its first application by Ladislaus von Bortkiewicz (*Das Gesetz der Kleinen Zahlen*, Teubner, Leipzig, 1898) concerned the number of Prussian cavalrymen killed by horse kicks.

The following BASIC program is based on the recursion formula:

$$f(x)=\frac{\mu}{x} f(x-1) \quad . \tag{67}$$

```
10:   "POISSON"
15:   INPUT "M=";M, "X=";X
20:   P=EXP-M:S=1:Q=1
```

```
25:   IF X=0 LET F=P GOTO 50
30:   FOR I=1 TO X
35:   Q=Q*M/I: S=S+Q
40:   NEXT I
45:   F=P*S:P=P*Q
50:   PRINT F:PRINT P
60:   END
```

EXAMPLE: μ=2.4 x=4

F=9.041314097×10^{-1}, P=$f(x)$=1.2540848986×10^{-1} .

3. Hypergeometric Distribution

This distribution is used to solve problems of sampling inspection without replacement, but it has many other applications (i.e., Florida State Lottery, quality control, etc.).

Its probability function is given by:

$$f(x)=\frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} \quad \text{for } x=0, 1, 2, 3\ldots n \tag{68}$$

where N=lot size, n=sample size, a=number of defects in lot, and x=number of defects (successes) in sample.

$$\text{Mean: } \mu=n\times\frac{a}{N} \qquad \text{Variance: } \sigma^2=\frac{n\times a\times(N-a)\times(N-n)}{N^2\times(N-1)} \ . \tag{69}$$

The following BASIC program calculates the hypergeometric probability function and its associated cumulative distribution by calculating the logarithms (to avoid possible overflow) of the three binomial coefficients of the distribution and subsequent multiplcation and division, which appears in line 350 as summation and difference of the logarithms.

```
300:  "HYPERGEOMETRIC"
305:  INPUT "N=";N, "N1, "A=";A
310:  INPUT "X=";X1

315:  NC=N:KC=N1:GOSUB 400
320:  CD=C

325:  CH=0: FOR X=0 TO X1
330:  NC=A:KC=X:GOSUB 400
335:  C1=C
340:  NC=N-A:KC=N1-X:GOSUB 400
345:  C2=C
350:  HX-EXP (C1+C2-CD)
355:  CH=CH+HX
360:  NEXT X:BEEP3
```

```
365:    PRINT CH: PRINT HX
370:    END

400:    C=0:;M=NC-KC
405:    IF KC=0 RETURN
410:    FOR I=1 TO KC
415:    C=C+LN((M+I)/I)
420:    NEXT I : RETURN
```

4. Negative Binomial (Pascal) Distribution

The negative binomial, or Pascal, distribution finds application in tests that are terminated after a predetermined number of successes has occurred. Therefore, it is also called the *binomial waiting time distribution*. Its application always requires a sequential-type situation. The random variable X in this case is the number of trials necessary for k successes to occur. (The last trial is always a success!)

The probability function is readily derived by observing that we have a binomial distribution up to the last trial, which is then followed by a success. As in the binomial distribution, the probability p of a success is again assumed to be constant for each trial.

Therefore:

$$f(x)=\binom{x-1}{k-1}p^k(1-p)^{x-k} \quad \text{for } x=k,\ k+1,\ k+2\ \ldots \tag{70}$$

where x=Number of trials
k=Number of successes
p=Probability of success.

Also:

$$\text{Mean: } \mu=\frac{k}{p} \qquad \text{Variance: } \sigma^2=\frac{k(1-p)}{p^2} \quad . \tag{71}$$

The special case k=1 gives the *geometric distribution*. The geometric distribution finds application in reliability analysis, for instance, if we investigate the performance of a switch that is repeatedly turned on and off, or a mattress to withstand repeated pounding in a "torture test."

PROBLEM: What is the probability of a mattress surviving 500 thumps if the probability of failure as the result of one thump is p=1/200?

SOLUTION: The probability of surviving 500 thumps is $P(X>500)$ and is called the reliability of the mattress. The reliability function $R(x)$ is related to the cumulative distribution as:

$$R(x)=1-F(x) \quad . \tag{72}$$

Since the geometric probability function is given as $f(x)=p(1-p)^{x-1}$, we obtain for the reliability of the mattress:

$$R(x)=\sum_{x=501}^{\infty} f(x)=(1-p)^{500}=(0.995)^{500}=0.0816 \quad . \tag{73}$$

It is sometimes said that the geometric distribution has no "memory" because of the relation:

$$P(X>x_0+x\,|\,X>x_0)=P(X>x) \quad . \tag{74}$$

In other words, the reliability of a product having a geometric (failure) distribution is independent of its past history, i.e., the product does not age or wear out. These failures are called *random failures*.

Sometimes the transformation $z=x-k$ is made where z is the number of failures. Then the probability function is given by:

$$f(z)=\binom{k+z-1}{z} p^k \times q^z \text{ for } z=0,1,2,3\ldots \tag{75}$$

and these are the successive terms in $p^k(1-q)^{-z}$, a binomial expression with a negative index.

When a program of the binomial distribution is available, the negative probabilities can be obtained by the simple identity:

$$f_p(x\,|\,k,p)=\frac{k}{x}\, f_b(k\,|\,x,p) \tag{76}$$

where the subscript p denotes the Pascal distribution and the subscript b the binomial distribution.

PROBLEM: Given a 70-percent probability that a job applicant is made an offer, what is the probability of needing *exactly* 12 sequential interviews to obtain eight new employees?

SOLUTION: $$x=12,\ k=8,\ p=0.7\ (\mu=11.42,\ \sigma=2.2) \tag{77}$$

$$f_p(12\,|\,8,\ 0.7=\binom{11}{7}(0.7)^8\ (0.3)^4=0.1541 \quad . \tag{78}$$

A more interesting complementary relationship exists between the cumulative Pascal and the cumulative binomial distribution. This is obtained as follows: The event that more than x sequential trials are required to have k successes is identical to the event that x trials resulted in less than k successes. Expressed in mathematical terms we have:

$$P_p(X>x\,|\,k,p)=P_b(K<k\,|\,x,p) \quad . \tag{79}$$

The cumulative Pascal distribution can then be obtained from the cumulative binomial by the relationship:

$$F_p(x\,|\,k,p)=1-F_b(k-1\,|\,x,p) \quad . \tag{80}$$

PROBLEM: Given a 70-percent probability that a job applicant is made an offer, what is the probability that, *at most*, 15 sequential interviews are required to obtain eight new employees?

SOLUTION: $$F_p(15\,|\,8,\ 0.7)=1-F_b(7\,|\,15,\ 0.7)=0.95=95 \text{ percent} \quad . \tag{81}$$

H. Special Continuous Distributions

1. Normal Distribution

The normal distribution is the most important continuous distribution for the following reasons:

- Many random variables that appear in connection with practical experiments and observations are normally distributed. This is a consequence of the so-called Central Limit Theorem or Normal Convergence Theorem to be discussed later.
- Other variables are approximately normally distributed.
- Sometimes a variable is not normally distributed, but can be transformed into a normally distributed variable.
- Certain, more complicated distributions can be approximated by the normal distribution.

The normal distribution was discovered by De Moivre (1733). It was known to Laplace no later than 1774, but is usually attributed to Carl F. Gauss. He published it in 1809 in connection with the theory of errors of physical measurements. In France it is sometimes called the Laplace distribution. Mathematicians believe the normal distribution to be a physical law, while physicists believe it to be a mathematical law.

The normal distribution is defined by the equation:

$$g(y)=\frac{1}{\sqrt{y}}e^{-2\sqrt{y}},0<y<\infty \quad . \quad f(x)=\frac{1}{\sigma\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty<x<\infty \tag{82}$$

Mean: $\mu=\mu$ Variance: $\sigma^2=\sigma^2$

Skewness: $\alpha_3=0$ Kurtosis: $\alpha_4=3$.

The cumulative distribution function of this density function is an integral that cannot be evaluated by elementary methods.

The existing tables are given for the so-called *standard* normal distribution by introducing the *standardized random variable*. This is obtained by the following transformation:

$$z=\frac{(x-u)}{\sigma} . \tag{83}$$

In educational testing, it is known as the standard score or the "z-score." It used to be called "normal deviate" until someone perceived that this is a contradiction in terms (oxymoron), in view of the fact that deviates are abnormal.

Every random variable can be standardized by the above transformation. The standardized random variable always has the mean $\mu=0$ and the standard deviation $\sigma=1$.

The cumulative distribution of every symmetrical distribution satisfies the identity:

$$F(-x)=1-F(x) \quad . \tag{84}$$

The *error* function is defined as:

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt \quad . \tag{85}$$

It is related to the normal cumulative distribution by:

$$F(x)=\frac{1}{2}\left[1+\Phi\left(\frac{x}{\sqrt{2}}\right)\right] \tag{86}$$

where

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-u^2/2}\,du \quad . \tag{87}$$

We are often interested in the probability of a normal random variable to fall below (above) or between *k* standard deviations from the mean. The corresponding limits are called one-sided or two-sided *tolerance limits*. Figures 12 and 13 are given as illustrations.

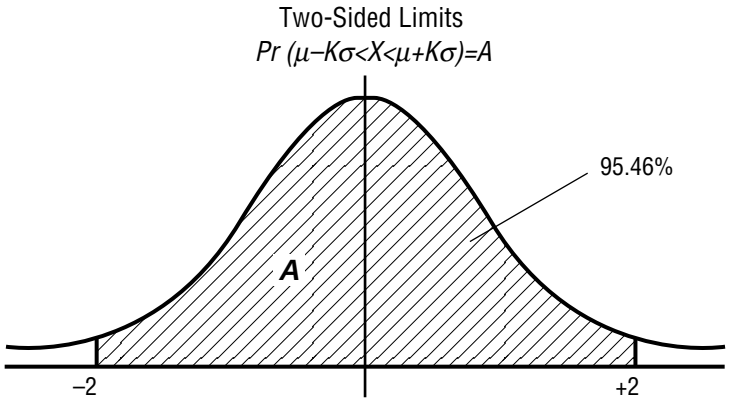


FIGURE 12.—Normal distribution areas: two-sided tolerance limits.

Stated, there is a 95.46-percent probability that a normal random variable will fall between the $\pm 2\ \sigma$ limits.

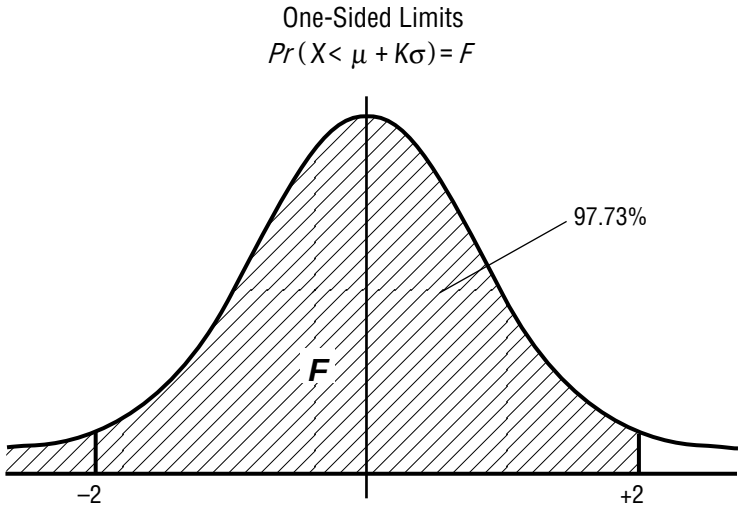


FIGURE 13.—Normal distribution areas: one-sided tolerance limits.

Stated, there is a 97.73-percent probability that a normal random variable will fall below (above) the $+2\sigma$ (-2σ) limit.

The relationship between the two areas A and F is:

$$A=2F-1 \quad \text{or} \quad F=(A+1)/2 \quad . \tag{88}$$

A normal random variable is sometimes denoted as:

$$X=N(\mu, \sigma^2)=\text{Gauss } (\mu, \sigma^2) \quad . \tag{89}$$

The standard normal variable or standard score is thus:

$$Z=N(0, 1)=\text{Gauss } (0, 1) \quad . \tag{90}$$

Table 4 gives the normal scores ("*K*-factors") associated with different levels of probability.

TABLE 4.—*Normal K-factors.*

One-Sided		Two-Sided	
Percent	K_1	Percent	K_2
99.90	3.0902	99.90	3.2905
99.87	3.0000	99.73	3.0000
99.00	2.3263	99.00	2.5758
97.73	2.0000	95.46	2.0000
95.00	1.6448	95.00	1.9600
90.00	1.2815	90.00	1.6449
85.00	1.0364	85.00	1.4395
84.13	1.0000	80.00	1.2816
80.00	0.8416	75.00	1.1503
75.00	0.6744	70.00	1.0364
70.00	0.5244	68.27	1.0000
65.00	0.3853	65.00	0.9346
60.00	0.2533	60.00	0.8416
55.00	0.1257	55.00	0.7554
50.00	0.0000	50.00	0.6745

As was already mentioned, the normal p.d.f. cannot be integrated in closed form to obtain the c.d.f.. One could, of course, use numerical integration, such as Simpson's Rule or the Gaussian quadrature method, but it is more expedient to calculate the c.d.f. by using power series expansions, continued fraction expansions, or some rational (Chebyshev) approximation. An excellent source of reference is *Handbook of Mathematical Functions* by Milton Abramowitz and Irene A. Stegun (eds.).

The following approximation for the cumulative normal distribution is due to C. Hastings, Jr.:

$$F(x)=1-\frac{e^{-x^2/2}}{\sqrt{2\pi}}\sum_{n=1}^{5}a_n y^n \tag{91}$$

$$y=\frac{1}{1+0.2316419x} \qquad 0\le x<\infty \tag{92}$$

where

a_1=0.3193815
a_2=–0.3565638
a_3=1.781478 Error=<7E–7
a_4=–1.821256
a_5=1.330274.

EXAMPLE: x=2.0 $F(x)$=0.97725

Sometimes it is required to work with the so-called inverse normal distribution, where the area F is given and the associated K-factor (normal score) has to be determined. A useful rational approximation for this case is given by equation (93) (equation 26.2.23 in the *Handbook of Mathematical Functions)* as follows:

We define $Q(k_p)=p$ where $Q=1-F$ and $0<p\le 0.5$.

Then

$$K_p=t-\frac{c_0+c_1t+c_2t^2}{1+d_1t+d_2t^2+d_3t^3}+\varepsilon(p), \quad t=\sqrt{-2\,\ell n\,p} \tag{93}$$

and $|\varepsilon(p)|<4.5\times10^{-4}$,

where

c_0=2.515517 d_1=1.432788
c_1=0.802853 d_2=0.189269
c_3=0.010328 d_3=0.001308.

A more accurate algorithm is given by the following BASIC program. It calculates the inverse normal distribution based on the continued fraction formulas also found in the *Handbook of Mathematical Functions*. Equation (94) in this reference (equation 26.2.14 in the *Handbook*) is used for $x>2$, equation (95) for $x<2$ (equation 26.2.15 in the *Handbook,* and the Newton-Raphson method using initial values obtained from equation (98).

$$Q(x)=Z(x)\left\{\frac{1}{x+}\frac{1}{x+}\frac{2}{x+}\frac{3}{x+}\frac{4}{x+}\ldots\right\},\ x>0 \tag{94}$$

$$Q(x)=\frac{1}{2}-Z(x)\left\{\frac{x}{1-}\frac{x}{3+}\frac{2x^2}{5-}\frac{3x^2}{7+}\frac{4x^2}{9-}\ldots\right\},\ x\geq 0 \tag{95}$$

$$x_p=t-\frac{a_0+a_1t}{1+b_1t+b_2t^2}+\varepsilon(t),\ t=\sqrt{1n\frac{1}{t^2}},\ and\ |\varepsilon(t)|<3\times10^{-3} \tag{96}$$

a_0=2.30753 b_1=0.99229
a_1=0.27061 b_2=0.04481.

```
'NORMIN (0.5<P<1)
DEFDBL A-Z
INPUT"P=";P:PI=3.141592653589793#

REM Equation 26.2.22

Q=1-P:T=SQR(-2*LOG(Q))
A0=2.30753:A1=.27061:B1=.99229:B2=.0481
NU=A0+A1*T:DE=1+B1*T+B2*T*T
X=T-NU/DE

L0: Z=1/SQR(2*PI)*EXP(-X*X/2)

IF X>2 GOTO L1

REM Equation 26.2.15
V=-25-13*X*X
FOR N=11 TO 0 STEP-1
U=(2*N+1)+(-1)^(N+1)*)N+1)*X*X/V
V=U:NEXT N
F=.5-Z*X/V
W=Q-F:GOTO L2

REM Equation 26.2.14
L1: V=X+30
FOR N=29 TO 1 STEP-1
U=X+N/V
V=U:NEXT N
F=Z/V:W=Q-F:GOTO L2

REM Newton-Raphson Method
L2: L=L+1
R=X:X=X-W/Z
E=ABS(R-X)
IF E>.0001 GOTO L0

PRINT USING "##.####";X
END
```

The normal distribution is often used for random variables that assume only positive values such as age, height, etc. This can be done as long as the probability of the random variable being smaller than zero, i.e., $P(X<0)$, is negligible. This is the case when the coefficient of variation is less than 0.3.

The normal and other continuous distributions are often used for discrete random variables. This is admissible if their numerical values are large enough that the associated histogram can be reasonably approximated by a continuous probability density function. Sometimes a so-called *continuity correction* is suggested which entails subtracting one-half from the lower limit of the cumulative distribution and adding one-half to its upper limit.

2. Uniform Distribution

The uniform distribution is defined as $f(x)=\frac{1}{b-a}$, $a\le x\le b$. The mean of the uniform distribution is $\mu=\frac{a+b}{2}$, and the variance $\sigma^2=\frac{(b-a)^2}{12}$. The uniform p.d.f. is depicted in figure 14.

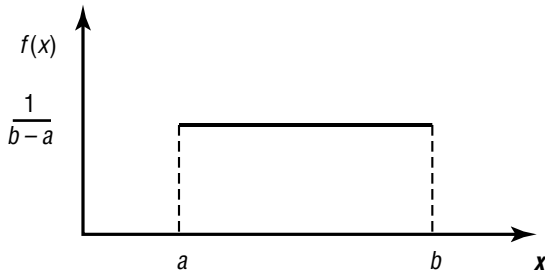


FIGURE 14.—Uniform p.d.f.

3. Log-Normal Distribution

Many statisticians believe that the log-normal distribution is as fundamental as the normal distribution itself. In fact, by the central limit theorem, it can be shown that the distribution of the *product* of *n* independent positive random variables approaches a log-normal distribution, just as the *sum* of *n* independent random variables approaches a normal distribution. It has been applied in a wide variety of fields including social sciences, economics, and especially in reliability engineering and life testing. The log-normal distribution is simply obtained by taking the natural logarithm of the original data and treating these transformed data as a normal distribution.

In short, $Y=\ell n\ X$ is normally distributed with *log mean* μ_Y and *log standard deviation* σ_Y. Since we are really concerned with the random variable X itself, we have to determine the probability density of X. By the methods shown later, it can be shown that it is:

$$f(x)=\frac{1}{x\sigma_Y\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{\ell n x-\mu_Y}{\sigma_Y}\right)^2},\quad x>0 \tag{97}$$

It can also be shown that $f(0)=f'(0)=0$.

The mode of the distribution occurs at:

$$x_{\text{Mode}} = e^{\mu_Y - \sigma_Y^2} \tag{98}$$

and the median at:

$$x_{\text{Median}} = e^{\mu_Y} \quad . \tag{99}$$

The mean is:

$$\mu_x = e^{\mu_Y + (1/2)\sigma_Y^2} \quad . \tag{100}$$

The variance is:

$$\sigma_x^2 = \left(e^{2\mu_Y + \sigma_Y^2} \right) \left(e^{\sigma_Y^2} - 1 \right) \quad . \tag{101}$$

The distribution has many different shapes for different parameters. It is positively skewed, the degree of skewness increasing with increasing σ_Y.

Some authors define the log-normal distribution in terms of the Briggsian (Henry Briggs, 1556–1630) or common logarithm rather than on the Naperian (John Napier, 1550–1616) or natural logarithm. The log mean μ_Y and the log standard deviation σ_Y are nondimensional pure numbers.

4. Beta Distribution

This distribution is a useful model for random variables that are limited to a finite interval. It is frequently used in Bayesian analysis and finds application in determining distribution-free tolerance limits.

The beta distribution is usually defined over the interval (0, 1) but can be readily generalized to cover an arbitrary interval (x_0, x_1). This leads to the probability density function:

$$f(x) = \frac{1}{(x_1 - x_0)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x - x_0}{x_1 - x_0} \right)^{(\alpha - 1)} \left(1 - \frac{x - x_0}{x_1 - x_0} \right)^{(\beta - 1)} \quad . \tag{102}$$

The standard form of the beta distribution can be obtained by transforming to a beta random variable over the interval (0, 1) using the standardized z-value:

$$z = \frac{x - x_0}{x_1 - x_0}, \text{ where } 0 \le z \le 1 \quad . \tag{103}$$

The beta distribution has found wide application in engineering and risk analysis because of its diverse range of possible shapes (see fig. 15 for different values of α and β.) The mean and variance of the standardized beta distribution is:

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \quad . \tag{104}$$

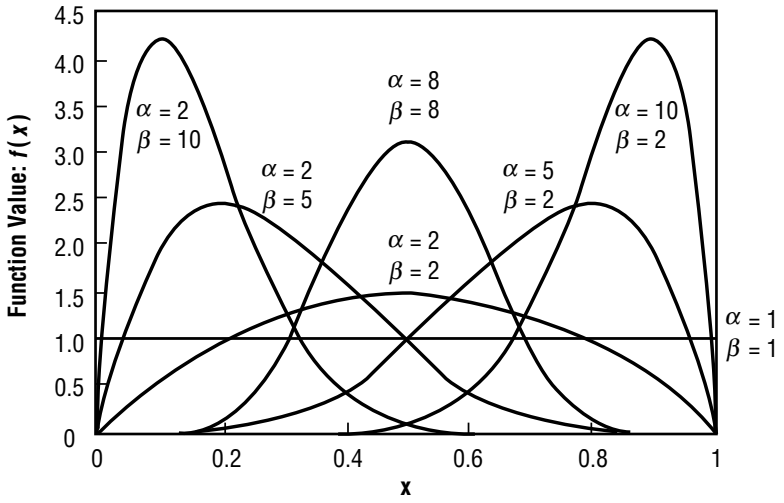


FIGURE 15.—Examples of standardized beta distribution.

The cumulative beta distribution is known as the incomplete beta function. An interesting and useful relationship exists between the binomial and beta distribution. If X is a binomial random variable with parameters p and n, then:

$$P(X \le x) = \frac{\Gamma(n+1)}{\Gamma(n-x)\Gamma(x+1)} \int_0^{1-p} t^{n-x-1}(1-t)^x \, dt \quad . \tag{105}$$

5. Gamma Distribution

The gamma distribution is another two-parameter distribution that is suitable for fitting a wide variety of statistical data. It has the following probability density function:

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \quad \text{for } x>0 \text{ and } \alpha, \beta>0 \quad . \tag{106}$$

A typical graph of the gamma p.d.f. is shown in figure 16.

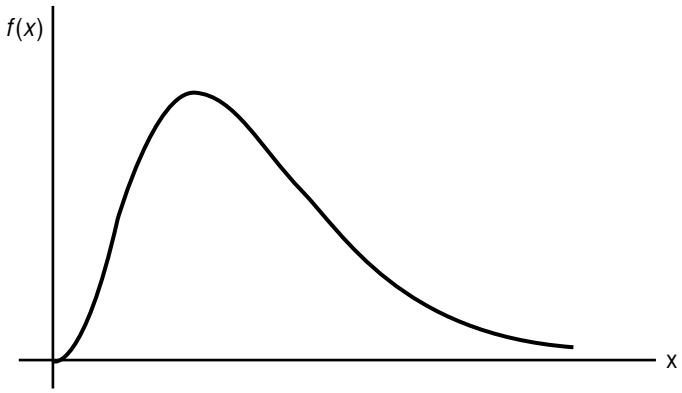


FIGURE 16.—Gamma distribution.

The parameter β is a scale parameter that only changes the horizontal scale. The parameter α is called the index or shape parameter because changes in α result in changes in the shape of the graph of the p.d.f. The quantity $\Gamma(\alpha)$ represents the gamma function defined by:

$$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\,dx \quad . \tag{107}$$

Integration by parts shows that $\Gamma(\alpha)=(\alpha-1)\,\Gamma(\alpha-1)$. If α is a positive integer then $\Gamma(\alpha)=(\alpha-1)!$ The gamma distribution has many applications in inventory control, queuing theory, and reliability studies. In reliability work it is often found to be in competition for application with the Weibull distribution, which is discussed in the following section.

The moment generating function of the gamma distribution is given by:

$$M(t)=E[e^{tx}]=(1-\beta t)^{-\alpha} \tag{108}$$

from which the first two central moments can be derived as:

$$\begin{aligned}&\text{Mean: } \mu=\alpha\beta\\ &\text{Variance: } \sigma^2=\alpha\beta^2 \ .\end{aligned} \tag{109}$$

This gamma distribution includes the exponential distribution as a special case when $\alpha=1$. The exponential distribution is commonly used in the study of lifetime, queuing, and reliability studies. Its single parameter β is the mean of the distribution and its inverse is called the failure rate or arrival rate. In reliability studies the mean is often referred to as the mean time between failures (MTBF).

Another special case of the gamma distribution is obtained by setting $\alpha=\nu/2$ and $\beta=2$. It is called the chi-square (χ^2) distribution, which has many statistical applications as discussed subsequently in more detail.

When the parameter α is restricted to integers, the gamma distribution is also called the Erlangian distribution. In this case, the cumulative distribution can be obtained by integration by parts. Otherwise, the gamma density cannot be integrated in closed form and the cumulative probability distribution is then referred to as the incomplete gamma function. The Erlangian represents the waiting time distribution for α independent events, each of which has an exponential distribution with mean β.

6. Weibull Distribution

This distribution was known to statisticians as the Fisher-Tippett Type III asymptotic distribution for minimum values or the third asymptotic distribution of smallest extreme values. It was used by the Swedish engineer Waloddi Weibull (1887–1979) in the analysis of the breaking strength of materials (1951). It has gained quite some popularity in reliability engineering due to its mathematical tractability and simple failure rate function. It is extensively used in life testing where it is competing with the gamma and log-normal distributions in search for the true nature of the random variable under investigation. The Weibull distribution can be easily memorized using its cumulative distribution function, which is:

$$F(x)=1-e^{-\alpha x^\beta}=1-e^{-(x/\eta)^\beta} \tag{110}$$

where η is called *characteristic life*.

It is easily seen that for $x=\eta$ the cumulative distribution is always $1-1/e=0.6321$ regardless of the value of the parameter β. The first form of the cumulative distribution is somewhat easier to manipulate mathematically. If the parameter α is known, the characteristic life can be calculated as $\eta=\alpha^{-1/\beta}$. The probability density is obtained by simple differentiation of the cumulative distribution.

There are two special cases that give rise to distributions of particular importance:

CASE 1: $\beta=1$

This is the *exponential* distribution, which is also a special case of the gamma distribution. It has many important applications in reliability and life testing.

CASE 2: $\beta=2$

This is the Raleigh distribution. It describes the distribution of the magnitude of winds if the north-south winds and the east-west winds are independent and have identical normal distributions with zero mean. It is also the distribution for the circular error probability (CEP). Another application arises in random vibration analysis where it can be shown that the envelope of a narrow-band random process has a Rayleigh distribution.

7. Extreme Value (Gumbel) Distribution

This distribution is seldom used in reliability engineering and in life and failure data analysis. It is, however, used for some types of "largest observations" such as flood heights, extreme wind velocities, runway roughness, etc. It is more precisely called the *largest extreme value distribution*. Sometimes it is also called the Gumbel distribution, after Emil J. Gumbel (1891–1967), who pioneered its use. It is briefly presented here for the sake of completeness.

Its cumulative distribution function is:

$$F(x)=\exp\{-\exp(-(x-\lambda)/\delta)\}\ , \qquad -\infty < X < \infty \tag{111}$$

$$\text{Mean: } \mu=\lambda+\gamma\delta, \text{ where } \gamma=0.57722 \text{ (Euler's constant)} \tag{112}$$

$$\text{Variance: } \sigma^2=\pi^2\delta^2/6\ . \tag{113}$$

I. Joint Distribution Functions

1. Introduction

In many engineering problems we simultaneously observe more than one random variable. For instance, in the cantilever beam shown in figure 17, we encounter five random variables: the length L, the force P, the Young's modulus E, the deflection δ, and the moment of inertia I_z. In general, all of them will, of course, have different distributions as indicated in figure 17.

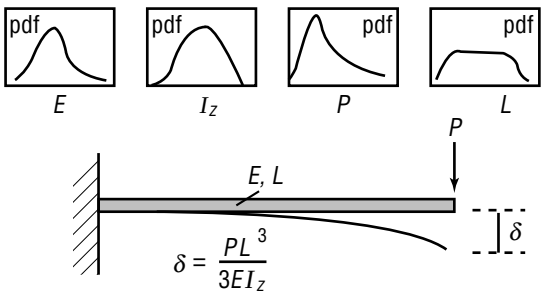


FIGURE 17.—Cantilever beam.

DEFINITION: A *k*-dimensional *random vector* $X=[X_1, X_2 \ldots X_k]$ is a set function that assigns to each outcome $E \in S$ real numbers $X_i(E)=x_i$ $(i=1, 2 \ldots k)$.

2. Discrete Case

The bivariate event is expressed as:

$$p(x_1, x_2)=P\{(X_1=x_1) \cap (X_2=x_2)) \quad . \tag{114}$$

This is called the (bivariate) *joint probability function*. Extensions to higher dimensional random vectors are self-evident. The bivariate probability function has the obvious properties:

$$p(x_1, x_2) \geq 0 \quad \text{and} \quad \sum_{x_1} \sum_{x_2} p(x_1, x_2)=1 \quad . \tag{115}$$

The cumulative bivariate distribution is defined as:

$$F(x_1, x_2)= \sum_{u_1 \leq x_1} \sum_{u_2 \leq x_2} P(u_1, u_2) \quad . \tag{116}$$

A natural question arises as to the distribution of the random variables X_1 and X_2 themselves. These are called *marginal probability functions* and are defined as:

$$p_1(x_1)= \sum_{x_2} P(x_1, x_2) \tag{117}$$

and

$$p_2(x_2)= \sum_{x_1} P(x_1, x_2) \quad . \tag{118}$$

NOTE: The terminology stems from the fact that if the probabilities associated with the different events are arranged in tabular form, the marginal probabilities appear in the margin of the table.

The *conditional probability function* is defined as:

$$p_1(x_1 \mid x_2)=\frac{p(x_1, x_2)}{p_2(x_2)} \quad . \tag{119}$$

3. Continuous Case

The extension to continuous random variable is easily achieved by replacing summation by integration (and usually by replacing p with f).

DEFINITION: The c.d.f. is defined as:

$$F(x_1,x_2)=P(X_1\le x_1, X_2\le x_2)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(u_1,u_2)\,du_1\,du_2 \tag{120}$$

and the bivariate probability density is obtained by partial differentiation as:

$$f(x_1,x_2)=\frac{\partial^2 F(x_1,x_2)}{\partial x_1\,\partial x_2} \quad . \tag{121}$$

Similarly, the conditional probability density is defined as:

$$f(x_1\,|\,x_2)=\frac{f(x_1,x_2)}{f_2(x_2)} \quad . \tag{122}$$

4. Independent Random Variables and Bayes' Rule

The notion of independence, which was introduced in connection with intersections of events, can be analogously defined for random vectors. Informally speaking, when the outcome of one random variable does not affect the outcome of the other, we state that the random variables are independent.

DEFINITION: If $f(x_1, x_2)$ is the joint probability density of the random variables X_1 and X_2, these random variables are independent if and only if:

$$f(x_1, x_2)=f_1(x_1)\,f_2(x_2) \quad . \tag{123}$$

Extensions to the multivariate case are obvious.

Sometimes the concept of independence is defined in terms of the cumulative distributions. In this case we have:

$$F(x_1, x_2)=F_1(x_1)\,F_2(x_2) \quad . \tag{124}$$

Bayes' Rule. In recent years there has been a growing interest in statistical inference, which looks upon parameters (mean, standard deviations, etc.) as random variables. This approach is known as Bayesian estimation. It provides a formal mechanism for incorporating our degree of personal judgment about an unknown parameter into the estimation theory before experimental data are available. As these data become available, the personal judgment is constantly updated to reflect the increase in knowledge about a certain situation.

DEFINITION 1: The probability distribution $h_o(p)$ of the parameter p, which expresses our belief about the possible value p before a sample is observed, is called the *prior distribution* of p. We reiterate that $h_o(p)$ makes use of the additional subjective information about the value p prior to taking a sample.

DEFINITION 2: The updated probability distribution of the parameter p, which expresses our increase in knowledge, is called *posterior distribution*. This distribution is calculated from the relationship that exists between the joint and conditional probability distributions as:

$$f(x,p)=g_1(x\,|\,p)\;h_o(p)=g_2(p\,|\,x)\;h_2(x)\quad . \tag{125}$$

The posterior distribution is then given by:

$$g_2(p\,|\,x)=\frac{g_1(x\,|\,p)h_o(p)}{h_2(x)} \tag{126}$$

where $h_2(x)$=marginal distribution. Here $g_1(x\,|\,p)$ is the sampling distribution of x given the parameter p.

It is readily recognized that the above formula is Bayes' theorem as applied to the continuous case. Once the posterior distribution is obtained, it can be used to establish confidence intervals. These confidence intervals are called Bayesian confidence intervals.

It should be mentioned that if the prior information is quite inconsistent with the information gained from the sample, the Bayesian estimation yields results that are inferior to other estimation techniques; e.g., the maximum likelihood estimation. The Bayesian estimation technique has become a favorite in many different statistical fields.

As an application, we discuss the estimation of the unknown parameter p of a *binomial distribution*:

$$g_1(x\,|\,p)=\binom{n}{x}p^x(1-p)^{n-x} \tag{127}$$

n=Number of trials
x=Number of successes
p=Probability of success.

We assume the *prior* distribution of p to be the beta distribution:

$$h_0(p)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,p^{(\alpha-1)}(1-p)^{(\beta-1)}\quad . \tag{128}$$

We first calculate the marginal distribution of x for the joint density $f(x, p)$ by integrating over p:

$$h_2(x)=\int_0^1 f(x,p)\,dp=\int_0^1 g_1(x\,|\,p)h_0(p)\,dp\quad . \tag{129}$$

This yields:

$$h_2(x)=\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{(x+\alpha-1)}(1-p)^{(n-x+\beta-1)}\,dp \tag{130}$$

or:

$$h_2(x)=\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \quad . \tag{131}$$

The *posterior distribution* is, therefore:

$$g_2(p\,|\,x)=\frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+x)\Gamma(N-x+\beta)}\,p^{(x+\alpha-1)}(1-p)^{(n-x+\beta-1)} \quad . \tag{132}$$

We have established the following theorem: If x is a binomial distribution and the *prior* distribution of p is a beta distribution with the parameters α and β, then the *posterior* distribution of $g_2(p\,|\,x)$ is also a beta distribution with the new parameters $x+\alpha$ and $n-x+\beta$.

The expected value (mean) of the posterior distribution is:

$$\bar{p}=\int_0^1 p\,g_2(p\,|\,x)\,dp=\frac{x+\alpha}{n+\alpha+b} \quad . \tag{133}$$

EXAMPLE 1: Uniform Prior Distribution: If the prior distribution is uniform, i.e., $h_0(p)=1$, then $\alpha=\beta=1$ and the mean of the posterior distribution is given by:

$$\bar{p}=\frac{x+1}{n+2} \quad . \tag{134}$$

This is known as the Laplace law of succession. It is an estimate of the probability of a future event given that x successes have been observed in n trials. Sometimes this law is stated as follows: The more often an event has been known to happen, the more probable it is that it will happen again.

EXAMPLE 2: Tests With No Failures: Assuming a uniform prior distribution, we calculate the posterior distribution for a situation where *n tests* were performed *without* a failure. We obtain the posterior distribution $g_2\,(p\,|\,n)=(n+1)\,p^n$ (see fig. 18).

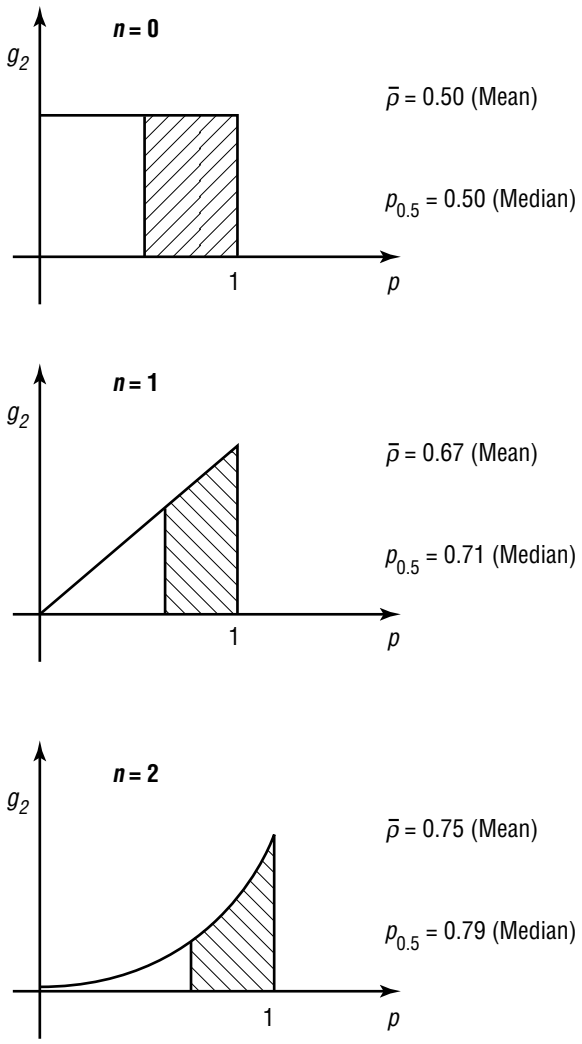


FIGURE 18.—Posterior distribution with no failures.

EXAMPLE 3: Two Tests and One Failure: The posterior distribution is $g_2(p \mid x)=6p\ (1-p)$ (see fig. 19).

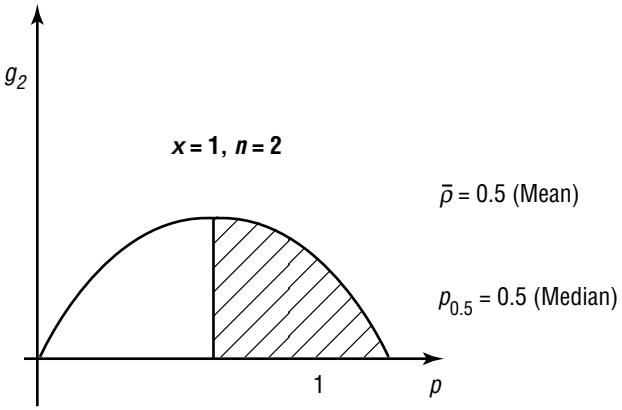


FIGURE 19.—Two tests and one failure.

EXAMPLE 4: Bayesian Confidence Limits: In reliability studies we are usually interested in one-sided confidence limits for the reliability p of a system. Given the posterior distribution $g_2(p \mid x)$ for the reliability p, we obtain the lower Bayesian $(1–a)$ confidence limit as:

$$\int_{p_L}^{1} g_2(p \mid x)\, dp = 1-\alpha \quad . \tag{135}$$

The lower confidence limit is obtained by solving this equation for the lower integration limit p_L (see fig. 20).

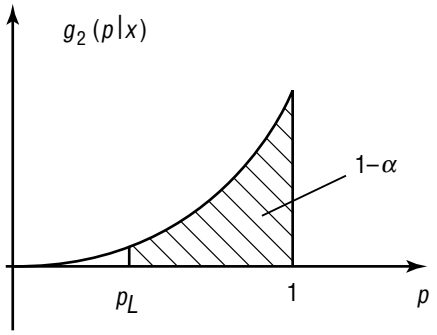


FIGURE 20.—Lower confidence limit.

The *upper* Bayesian $(1 – a)$ confidence limit is similarly obtained from the equation:

$$\int_{0}^{p_u} g_2(p \mid x)\, dp = 1-\alpha \tag{136}$$

If there are *no* test failures the posterior distribution is:

$$g_2(p \mid x) = (n+1)\, p^n \quad . \tag{137}$$

Inserting this posterior distribution into the above equation for the lower confidence limit for the reliability (probability of success), we can solve it for the $(1–\alpha)$ lower confidence limit in closed form and obtain:

$$p_L^{\,(n+1)} = \alpha \tag{138}$$

This lower confidence limit is often referred to as the *demonstrated* reliability when it is applied to a system which has undergone n tests without a failure.

It is sometimes necessary to find the number of tests required to demonstrate a desired reliability for a $(1 – \alpha)$ confidence level. This can be easily obtained by solving the above equation for n, which in this case is:

$$n=\frac{\ell n\,\alpha}{\ell n\,p_L}-1 \tag{139}$$

For instance, if it is required to demonstrate a 99-percent reliability at the 50-percent confidence level, it is necessary to run 68 tests without a failure.

If no confidence level is specified, the 50-percent confidence level is generally assumed. This is especially the case for high reliability systems.

The non-Bayesian confidence limit for the binomial parameter p for the above case is given by:

$$p_L{}^n=\alpha \tag{140}$$

It is somewhat more conservative than the Bayesian limit because it requires one more test to demonstrate the same level of reliability.

J. Mathematical Expectation

In many problems of statistics, we are interested not only in the expected value of a random variable X, but in the expected value of a function $Y=g(X)$.

DEFINITION: Let $\underline{X}=(X_1, X_2 \ldots X_n)$ be a continuous random vector having a joint probability density $f(\underline{x})=f(x_1, x_2 \ldots x_n)$ and let $u(\underline{X})$ be a function of $\underline{X}$. The *mathematical expectation* is then defined as:

$$E[u(\underline{X})]=\int_{-\infty}^{\infty} u(\underline{x})\, f(\underline{x})\, d\underline{x} \ . \tag{141}$$

For a discrete random vector the integral is replaced by a corresponding summation. Sometimes $E[.]$ is called the "expected value operator."

The vector notation introduced here simplifies the expression for the multiple integral appearing in the definition of the mathematical expectation. A further consideration of this general case is beyond the scope of the present text.

EXAMPLE: Chuck-A-Luck: The concept of a mathematical expectation arose in connection with the games of chance. It serves as an estimate of the average gain (or loss) one will have in a long series of trials. In this context the mathematical expectation is also called the *expected profit*. It is defined by the sum of all the products of the possible gains (losses) and the probabilities that these gains (losses) will occur. Expressed mathematically, it is given as:

$$E=\sum a_i p_i \ . \tag{142}$$

It is important to keep in mind that the coefficients a_i in the above equation are positive when they represent gains and negative when they represent losses.

In betting odds, the expected profit is set equal to zero (E=0). For example, if the odds of favoring an event are 3:1, the probability of the event happening is 3/4 and the expected profit is:

$$E=\$1\ (3/4)-\$3\ (1/4)=0 \ . \tag{143}$$

In *Scarne's New Complete Guide to Gambling*, John Scarne puts it this way: "When you make a bet at less than the correct odds, which you always do in any organized gambling operation, you are paying the operator the percentage charge for the privilege of making a bet. Your chances of winning has what is called a 'minus expectation.' When you use a system you make a series of bets, each with a minus expectation. There is no way of adding minuses to get a plus."

As an illustration for calculating the expected profit, we take the game of Chuck-A-Luck. In this game the player bets on the toss of three dice. He is paid whatever he bets for each die that shows his number.

The probability of any number showing on a die is p=1/6. Let the probability of one number showing on three dice be p_1, of two numbers p_2, and of three numbers p_3. Then:

$$p_1=3\,pq^2=3\ (1/6)\ (5/6)\ (5/6)=75/216 \tag{144}$$

$$p_2=3p^2q=3\ (1/6)\ (1/6)\ (5/6)=15/216 \tag{145}$$

$$p3=p^3=(1/6)\ (1/6)\ (1/6)=1/216\ . \tag{146}$$

The expected profit is, therefore:

$$E=(\$1)\ p_1+(\$2)\ p_2+(\$3)\ p_3-(\$1)\ p_{\mathrm{LOSS}} \tag{147}$$

where

$$p_{\mathrm{LOSS}}=1-(p_1+p_2+p_3)=125/216\ . \tag{148}$$

Thus,
$$E=\frac{1\times 75+2\times 15+3\times 1}{216}-1\times\frac{125}{216}=-\frac{17}{216}/\$ \tag{149}$$

or:
$$E=-7.87\ \text{cents/dollar}. \tag{150}$$

The house percentage is, therefore, 7.87 percent.

As may be noticed, we have already encountered a few special cases of mathematical expectation, such as the mean, variance, and higher moments of a random variable. Beyond these, the following simple functions are of particular interest.

1. Covariance

The covariance measures the degree of association between two random variables X_1 and X_2. It is defined as follows:

$$\mathrm{Cov}\ (X_1, X_2)=\sigma_{12}=E[(X_1-\mu_1)\ (X_2-\mu_2)]\ . \tag{151}$$

Performing the multiplication we obtain an alternate form of the covariance as:

$$\mathrm{Cov}\ (X_1, X_2)=E[X_1\times X_2]-E[X_1]\ E[X_2]=E[X_1\times X_2]-\mu_1\,\mu_2\ . \tag{152}$$

The covariance can be made nondimensional by dividing by the standard deviation of the two random variables. Thus, we get the so-called *correlation coefficient*:

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{V(X_1)}\sqrt{V(X_2)}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \quad . \tag{153}$$

It can be shown that if X_1 and X_2 are independent, then $\rho=0$. This is true for both discrete and continuous random variables. The converse of this statement is not necessarily true, i.e., we can have $\rho=0$ without the random variables being independent. In this case the random variables are called *uncorrelated*.

It can also be shown that the value of ρ will be on the interval (–1, +1), that is:

$$-1 \le \rho \le +1 \quad . \tag{154}$$

2. Linear Combinations

The linear combination of the elements of a random vector $\underline{X}=(X_1, X_2 \ldots X_n)$ is defined as

$$u(\underline{X}) = Y = a_0 + a_1 X + \ldots + a_n X_n \quad . \tag{155}$$

The mean and variance of this linear combination is:

$$E(Y) = \mu_Y = a_0 + \sum_{i=1}^{n} a_i E(X_i) = a_0 + \sum_{i=1}^{n} a_i \mu_i \quad , \tag{156}$$

and:

$$V(Y) = \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} a_i a_j \sigma_{ij} \quad . \tag{157}$$

If the variables are independent, the expression for the variance of Y is simplified because the covariance terms σ_{ij} are zero. In this case the variance reduces to:

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 \quad . \tag{158}$$

Notice that all these relationships are independent of the distribution of the random variables involved.

K. Functions of Random Variables

In many engineering applications we meet with situations in which we have to determine the probability density of a function of random variables. This is especially true in the theory of statistical inference, because all quantities used to estimate population parameters or to make decisions about a population are functions of the random observation appearing in a sample.

For example, suppose the circular cross section of a wire is of interest. The relationship between the cross section of the wire and its radius is given by $A=\pi R^2$. Considering R a random variable with a certain distribution, the area A is also a random variable. Our task is to find the probability density of A if the distribution of R is known.

DEFINITION: Let B be an event in the range space R_x, i.e., $B \in R_x$, and C be an event in range space R_y, i.e., $C \in R_y$ such that:

$$B=\{x \in R_X : h(x) \in R_y\} \quad . \tag{159}$$

Then B and C are *equivalent events*—that means they occur simultaneously. Figure 21 illustrates a function of a random variable.

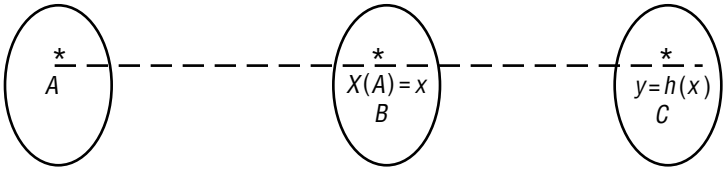


FIGURE 21.—A function of a random variable.

Equivalent events have equal probability:

$$P(A)=P(B)=P(C) \quad . \tag{160}$$

1. Discrete Random Variables

The p.d.f. of a function of a discrete r.v. is determined by the equivalent events method.

EXAMPLE: In the earlier coin-tossing problem, the random variable X can assume the four values x=0, 1, 2, 3 with respective probabilities 1/8, 3/8, 3/8, 1/8. Let the new variable be $Y=2X-1$. Then the equivalent events are given by $y=-1, 1, 3, 5$ with probabilities $p_Y(-1)=1/8$, $p_Y(1)=3/8$, $p_Y(3)=3/8$, $p_Y(5)=1/8$. Notice that equivalent events have equal probabilities.

2. Continuous Random Variables

The p.d.f. of a function of a continuous r.v. can be determined by any of the following methods:

Method I: Transformation of Variable Technique ("Jacobian" Method).

Method II: Cumulative Distribution Function Technique (Equivalent Event Method).

Method III: Moment Generating Function Technique.

APPLICATIONS:

Method I:

(a) Univariate case. Let $y=h(x)$ be either a decreasing or an increasing function. Then:

$$f(x)\,dx=g(y)\,dy, \tag{161}$$

or solving for g(y):

$$g(y)= f(x)\left|\frac{dx}{dy}\right| \ . \tag{162}$$

Note that the absolute value sign is needed because of the condition that $g(y) > 0$.

EXAMPLE 1: Let: $f(x)=e^{-x}$ for $0<x<\infty$. (163)

The new variable is given as $y=x^2/4$.

Thus:

$$\frac{dy}{dx}=x/2 \Rightarrow \frac{dx}{dy}=2/x \ . \tag{164}$$

Therefore:

$$g(y)=e^{-x}\left|\frac{2}{x}\right|=\frac{2}{x}e^{-x} \ . \tag{165}$$

In terms of the new variable y: $g(y)=\frac{1}{\sqrt{y}}e^{-2\sqrt{y}}, 0<y<\infty$. (166)

EXAMPLE 2: Random sine wave: $y=A \sin x$ (see figs. 22 and 23). The amplitude A is considered to be constant and the argument x is a random variable with a uniform distribution:

$$f(x)=\frac{1}{\pi} \text{ with} -\frac{\pi}{2}<x<\frac{\pi}{2} \tag{167}$$

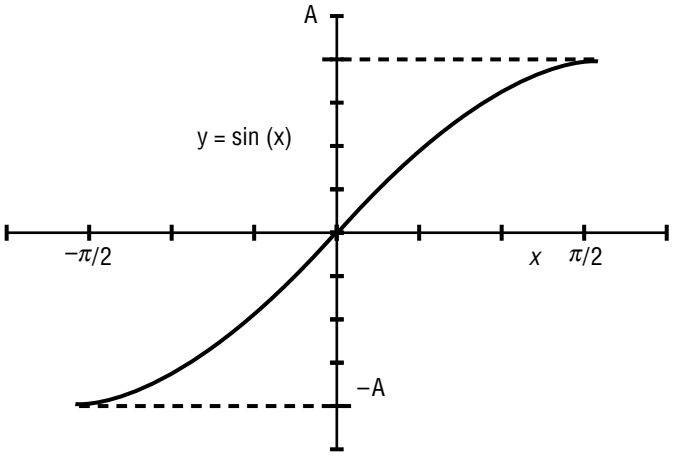


FIGURE 22.—Random sine wave.

$$\frac{dy}{dx} = A\cos x, \frac{dx}{dy} = \frac{1}{A\cos x} > 0, g(y) = \frac{1}{\pi}\frac{1}{\cos x} \quad . \tag{168}$$

Expressed in terms of the new variable *y*:

$$\cos(x) = \sqrt{1-\sin^2(x)} = \sqrt{1-(y/A)^2} \tag{169}$$

and finally:

$$g(y) = \frac{1}{\pi A\sqrt{1-(y/A)^2}} \quad . \tag{170}$$

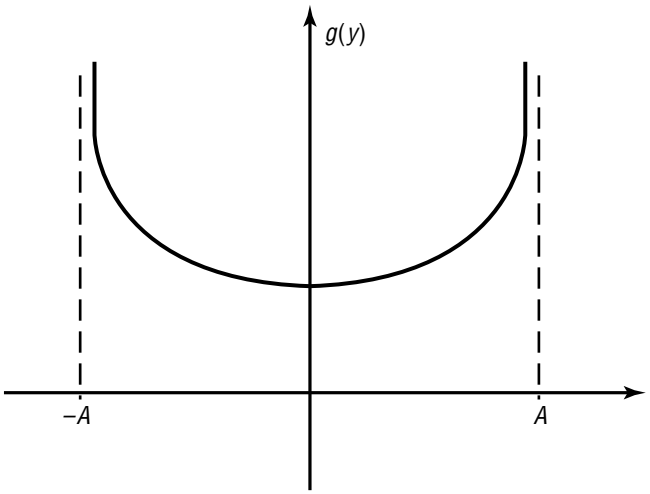


FIGURE 23.—Probability density of random sine wave.

EXAMPLE 3: Random number generator (probability integral transformation): A random variable *X* is given that has a uniform distribution over the interval (0, 1). Find a new random variable *Y* that has a desired probability distribution *g*(*y*).

According to method I, we have the relationship:

$$f(x)\ dx = g(y)\ dy \quad . \tag{171}$$

Since $f(x)=1$ for $0 \le x \le 1$ we obtain from this: $dx = g(y)\ dy$

and after integration,

$$x = \int_{-\infty}^{y} g(u)du = G(y) \tag{172}$$

where, of course, *G*(*y*) is the cumulative distribution of *y*. Note that the cumulative distribution *F*(*x*) has a uniform distribution over the interval (0, 1) independent of *f*(*x*).

The desired random variable Y is then obtained by solving the above equation for the inverse cumulative function, such that

$$y=G^{-1}(x), \text{ where } G^{-1}(x) \text{ is often referred to as the percentile function.} \tag{173}$$

Figure 24 shows the relationship between the two random variables in graphical form.

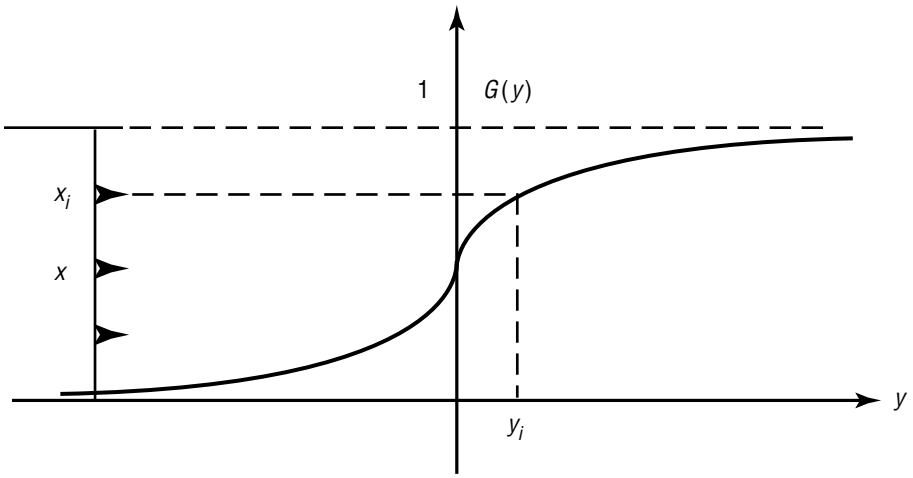


FIGURE 24.—Probability integral transformation.

EXAMPLE: Determine the transformation law that generates an exponential distribution, such that:

$$g(y)=\alpha e^{-\alpha y} \quad 0 < y < \infty \tag{174}$$

$$G(y)=\int_0^y \alpha e^{-\alpha u}\, du = 1-e^{-\alpha y}$$

$$x=G(y)=1-e^{-\alpha y} \tag{175}$$

$$y=-\frac{1}{\alpha}\ell n(1-X) \quad . \tag{176}$$

Random numbers can then be generated from:

$$y_i=-\frac{1}{\alpha}\ell n\, x_i \tag{177}$$

Bivariate case:

$$f(x, y)\, dx\, dy=g(u, v)\, du\, dv \quad . \tag{178}$$

As is known from advanced calculus, the differential elements *dxdy* and *dudv* are related through the Jacobian determinant as:

$$dxdy=\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} dudv=J(u,v)dudv \quad \text{where} \quad J(u,v)=\text{Jacobian determinant.} \tag{179}$$

The new joint probability density function is, then:

$$g(u,v)=f(x,y)\,|J(u,v)|\ . \tag{180}$$

EXAMPLE 1: If the random variables X_1 and X_2 have a standard normal distribution $N(0, 1)$, the ratio X_1/X_2 has a *Cauchy distribution*.

Let $y_1=x_1/x_2$ and $y_2=x_2 \Rightarrow x_1=y_1/y_2$ and $x_2=y_2$.

The partial derivatives are:

$$\frac{\partial x_1}{\partial y_1}=y_2,\ \frac{\partial x_1}{\partial y_2}=y_1,\ \frac{\partial x_2}{\partial y_1}=0,\ \frac{\partial x_2}{\partial y_2}=1 \tag{181}$$

and the Jacobian is then $J=y_2$.

The joint probability density function is:

$$g(y_1,y_2)=\frac{1}{2\pi}e^{-\frac{1}{2}x_1^2}\,e^{-\frac{1}{2}x_2^2}\left|y_2\right|=\frac{1}{2\pi}e^{-\frac{1}{2}y_2^2(1+y_1^2)}\left|y_2\right|\ . \tag{182}$$

The marginal distribution is:

$$g(y_1)=\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-\frac{1}{2}y_2^2(1+y_1^2)}\left|y_2\right|dy=\int_0^{\infty}\frac{1}{\pi}e^{-\frac{1}{2}y_2^2(1+y_1^2)}\,d\left(\frac{y_2^2}{2}\right) \tag{183}$$

which gives:

$$g(y_1)=\frac{1}{\pi(1+y_1^2)}\ . \tag{184}$$

EXAMPLE 2: Box-Muller method: An important use of the Jacobian method for finding the joint probability density of a function of two or more functions of a random vector is the Box-Muller algorithm. Since its introduction in 1958, it has become one of the most popular method for generating normal random variables. It results in high accuracy and compares favorably in speed with other methods. The method actually generates a pair of standard normal variables from a pair of uniform random variables.

Consider the two independent standard normal random variables x and y with densities:

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\ \text{ and }\ f(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2} \tag{185}$$

Let us introduce the following polar coordinates:

$$x=r\cos\phi \text{ and } y=r\sin\phi \ . \tag{186}$$

The joint density of the polar coordinates can be obtained from the joint probability density of the Cartesian coordinates using the Jacobian method as:

$$g(r,\phi)=f(x,y)\,|J| \tag{187}$$

where $|J|$ is the absolute value of the Jacobian determinant:

$$J=\begin{vmatrix}\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi}\\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi}\end{vmatrix}=\begin{vmatrix}\cos(\phi) & -r\sin(\phi)\\ \sin(\phi) & r\cos(\phi)\end{vmatrix}=r\cos^2(\phi)+r\sin^2(\phi)=r \ . \tag{188}$$

Because of the independence of the Cartesian normal random variables we have:

$$f(x,y)=f(x)\,f(y)=\frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}=\frac{1}{2\pi}e^{-\frac{r^2}{2}} \ . \tag{189}$$

and finally using the value of the Jacobian determinant:

$$g(r,\phi)=\frac{r}{2\pi}e^{-\frac{r^2}{2}} \ . \tag{190}$$

Since the new joint distribution is a product of a function of r alone and a constant, the two polar random variables r and ϕ are independent and have the following distribution:

$$g_1(r)=r\,e^{-\frac{r^2}{2}} \ , \ 0<\phi<\infty \tag{191}$$

and:

$$g_2(\phi)=\frac{1}{2\pi} \ , \ 0<\phi<2\pi \ . \tag{192}$$

The random variable r has a Rayleigh distribution that is a special case of the Weibull distribution with α=1/2 and β=2. It can be simulated using the probability integral transformation. The random variable ϕ is a uniform distribution over the interval $(0, 2\pi)$. Therefore, we can now generate the pair of independent normal random variables x and y using two uniform random numbers u_1 and u_2 by the following algorithm:

$$x=\sqrt{-2\,\ell n\,(u_1)}\cos(2\pi u_2) \text{ and } y=\sqrt{-2\,\ell n\,(u_1)}\sin(2\pi u_2) \tag{193}$$

Method II:

If $X_1, X_2 \ldots X_n$ are continuous random variables with a given joint probability density, the probability density of $y=h(x_1, x_2 \ldots x_n)$ is obtained by first determining the cumulative distribution:

$$G(y)=P(Y\leq y)=P\ [h(x_1,\ x_2,\ldots\ x_n)\leq y] \tag{194}$$

and then differentiating (if necessary) to obtain the probability density of y:

$$g(y)=\frac{dG(y)}{dy} \tag{195}$$

EXAMPLE 1: $Z=X+Y$. This is one of the most important examples of the function of two random variables X and Y. To determine the cumulative distribution $F(z)$ we observe the region of the xy plane in which $x+y\leq z$. This is the half plane left to the line defined by $x+y=z$ as shown in figure 25.

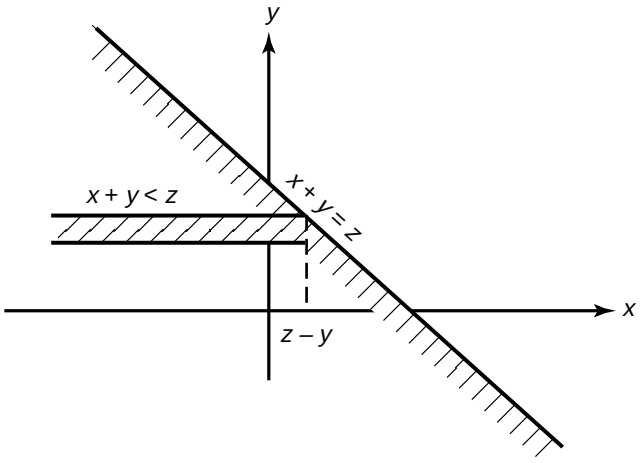


FIGURE 25.—Sum of two random variables.

We now integrate over suitable horizontal strips to get:

$$F(z)=\int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_{xy}(x,y)\,dxdy \quad . \tag{196}$$

Assuming independence of the random variables, we have $f_{xy}(x,\ y)=f_x(x)f_y(y)$. Introducing this in the above equation, we obtain:

$$F(z)=\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{z-y} f_x(x)dx\right\}f_y(y)dy=\int_{-\infty}^{\infty} F_x(z-y)\,f_y(y)dy \quad . \tag{197}$$

Differentiating with respect to z, we get the p.d.f. of z as:

$$f(z)=\int_{-\infty}^{\infty} f_x(z-y)\,f_y(y)dy=\int_{-\infty}^{\infty} f_x(x)\,f_y(z-x)dx \quad . \tag{198}$$

Note that the second integral on the right-hand side is obtained if we integrate over vertical strips in the left half plane.

RESULT: The p.d.f. of the sum of two independent random variables is the *convolution integral* of their respective probability densities.

EXAMPLE 2: $Z=X-Y$. Another important example is the probability of the difference between two random variables. Again we find the cumulative distribution $F(z)$ by integrating over the region in which $x-y<z$ using horizontal strips as indicated in figure 26. Following similar steps as in the previous example, we obtain the cumulative distribution of Z as:

$$F(z)=\int_{-\infty}^{\infty}\int_{-\infty}^{z+y} f_{xy}(x,y)\,dx\,dy \quad . \tag{199}$$

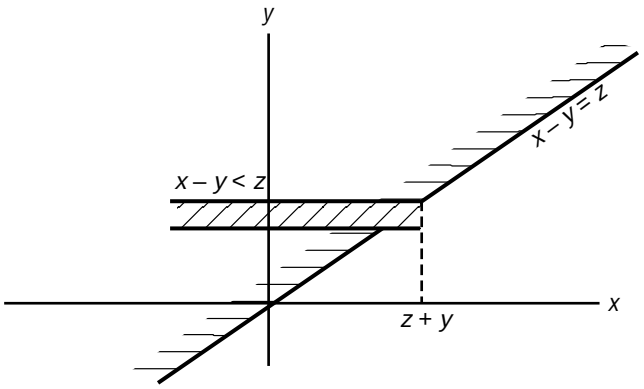


FIGURE 26.—Difference of two random variables.

Assuming independence of the random variables as above yields:

$$F(z)=\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{z+y} f_x(x)\,dx\right\} f_y(y)\,dy=\int_{-\infty}^{\infty} F_x(z+y)\,f_y(y)\,dy \quad . \tag{200}$$

Differentiating with respect to z, we get the p.d.f. of z as:

$$f(z)=\int_{-\infty}^{\infty} f_x(z+y)\,f_y(y)\,dy=\int_{-\infty}^{\infty} f_x(x)\,f_y(z+x)\,dx \quad . \tag{201}$$

RESULT: The p.d.f. of the difference between two independent random variables is the *correlation integral* of their respective probability densities.

APPLICATION: Probabilistic failure concept. One way a system can fail is when the requirements imposed upon the system exceed its capability. This situation can occur in mechanical, thermal, electrical, or any other kind of system. The probabilistic failure concept considers the requirement and the capability of the system as random variables with assumed probability distributions. Let C represent the capability (strength) having a distribution $f_C(x)$ and R the requirements (stress) of the system having a distribution $f_R(y)$. Then system failure occurs if $R>C$ or, in other words, if the difference $U=C-R$ is negative. The difference U is called the interference random variable (see fig. 27).

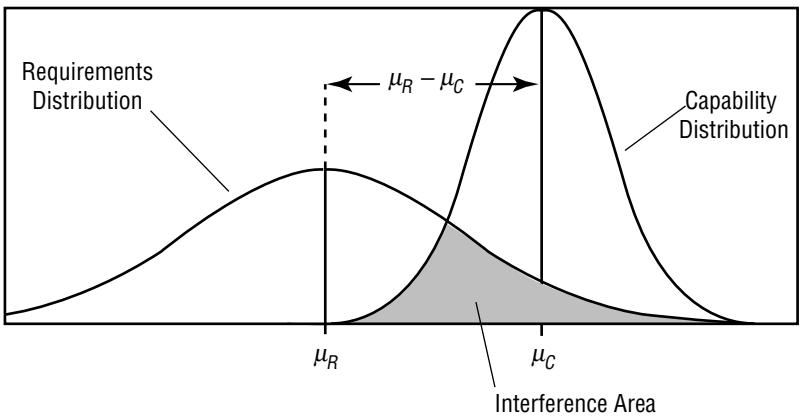


FIGURE 27.—Interference random variable.

The cumulative distribution of U is given as:

$$F(u) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{u+y} f_C(x)dx \right\} f_R(y)dy = \int_{-\infty}^{\infty} F_C(u+y)\, f_R(y)dy \quad . \tag{202}$$

Therefore, the probability of failure, also called the *unreliability* of the system, is:

$$F(0) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{y} f_C(x)dx \right\} f_R(y)dy = \int_{-\infty}^{\infty} F_C(y)\, f_R(y)dy \quad . \tag{203}$$

In general, the integration has to be done numerically. However, if the two random variables are normally distributed, we know that the difference of the two random variables is again normally distributed because of the reproductive property of the normal distribution.

Introducing the normal score of the difference u, we have:

$$z = \frac{u - (\mu_C - \mu_R)}{\sqrt{\sigma_C^2 + \sigma_R^2}} \quad . \tag{204}$$

The probability of failure is then given by setting u=0. By design, the probability of failure will be very small, such that the normal score will turn out to be negative. To use the normal score as a measure of reliability, it is therefore customary to introduce its negative value, which is called the *safety index* β. It is numerically identical with the one-sided K-factor and the above equation then becomes:

$$\beta = \frac{\mu_C - \mu_R}{\sqrt{\sigma_C^2 + \sigma_R^2}} \quad . \tag{205}$$

The corresponding reliability can then be obtained from the normal K-factor table or using a program that calculates the cumulative normal distribution. A good reference book on this subject is *Reliability in Engineering Design* by Kapur and Lamberson.

Method III:

This method is of particular importance for the case where we encounter linear combinations of independent random variables. The method uses the so-called moment generating function.

DEFINITION: Given a random variable X the moment generating function of its distribution is given by:

$$M(t)=\sum_{i=1}^{\infty} e^{tx_i}\, p(x_i) \qquad \text{for discrete } X \tag{206}$$

and

$$M(t)=\int_{-\infty}^{\infty} e^{tx} f(x)dx \qquad \text{for continuous } X \ . \tag{207}$$

The moment generating function derives its name from the following moment generating property:

$$\frac{d^n M(t)}{dt^n}=\sum_{i=1}^{\infty} x_i^n e^{tx_i}\, p(x_i) \qquad \text{for discrete } X \tag{208}$$

and

$$\frac{d^n M(t)}{dt^n}=\int_{-\infty}^{\infty} x^n e^{tx} f(x)dx \qquad \text{for continuous } X \ . \tag{209}$$

Taking the derivatives and setting t=0 gives:

$$\frac{d^n M(0)}{dt^n}=\sum_{i=1}^{\infty} x_i^n\, p(x_i)=\mu_n \qquad \text{for discrete } X \tag{210}$$

and

$$\frac{d^n M(0)}{dt^n}=\int_{-\infty}^{\infty} x^n f(x)dx \qquad \text{for continuous } X \ . \tag{211}$$

Readers who are familiar with the Laplace transform will quickly notice that the moment generating function for a continuous random variable X is identical with the definition of the (two-sided or bilateral) Laplace transform if one substitutes t=–s. In fact, if one defines the moment generating function as the bilateral Laplace transformation of the probability density function, it is possible to use existing Laplace transform tables to find the moment generating function for a particular p.d.f..

In some advanced textbooks the variable t in the moment generating function is replaced by "$i\,t$," where i=$\sqrt{-1}$. It is then called the *characteristic function* of the distribution. This measure is sometimes necessary because certain distributions do not have a moment generating function.

Method III makes use of the following theorem:

If $X_1, X_2 \ldots X_n$ are independent random variables and $Y=X_1+X_2+\ldots X_n$, then

$$M_Y(t)=\prod_{i=1}^{n} M_{x_i}(t) \quad . \tag{212}$$

EXAMPLE: The normal distribution has the moment generating function:

$$M_n(t)=e^{\mu t+\frac{1}{2}\sigma^2 t^2} \quad . \tag{213}$$

It is readily seen that the sum of k normal random variables is again a normal distribution with mean $\mu_Y=\sum\mu_i$ and variance $\sigma_Y^2=\sum\sigma_i^2$. This is called the *reproductive property* of the normal distribution. Only a few distributions have this property. The Poisson distribution is another one that has this property.

L. Central Limit Theorem (Normal Convergence Theorem)

If $X_1, X_2,\ldots X_n$ are *independent* random variables with *arbitrary* distributions such that

mean: $$E(x_i)=\mu_i \ , \tag{214}$$

and variance: $$V(x_i)=\sigma_i^2 \ , \tag{215}$$

then: $$Z=\frac{\sum x_i-\sum\mu_i}{\sqrt{\sum\sigma_i^2}} \tag{216}$$

has an approximate standard normal distribution $N(0, 1)$ for large n. This is the basic reason for the importance and ubiquity of the normal distribution.

APPLICATION: Normal random number generator:

$$\text{Gauss}(0,1)=N(0,1)=\sum_{i=1}^{12} U_i(0,1)-6 \quad . \tag{217}$$

Summing up 12 random numbers, which are uniform over the interval (0,1), gives an easy method and, for practical purposes, accurate standard normal random variable. It is truncated at $\pm 6\ \sigma$. The probability of exceeding these limits is less than 2×10^{-9}.

M. Simulation (Monte Carlo Methods)

Simulation techniques are used under the following circumstances:

- The system cannot be analyzed by using direct and formal analytical methods.
- Analytical methods are complex, time-consuming, and costly.
- Direct experimentation cannot be performed.
- Analytical solutions are beyond the mathematical capability and background of the investigator.

EXAMPLE: Buffon's needle: This famous problem consists of throwing fine needles of length L on a board with parallel grid lines separated by a constant distance a. To avoid intersections of a needle with more than one line, we let $a>L$ (see fig. 28).

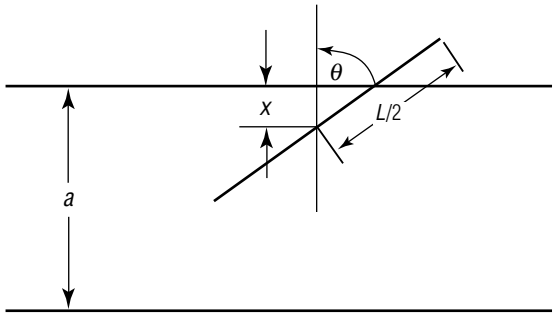


FIGURE 28.—Buffon's needle.

We define the position of a needle in terms of the distance x of its midpoint to the *nearest* line, and its direction with respect to the grid line by the angle θ.

For an intersection to occur, we have the condition:

$$x<L/2 \cos \theta \tag{218}$$

where the range of the angle is $0<\theta<\pi/2$ and the range of the distance $0<x<a/2$.

Both random variables x and θ are assumed to be uniform over their respective interval. Geometrically, we can observe that this condition is given by the shaded area under the cosine curve in figure 29.

The rectangle represents all possible pairs of (x,θ) and the shaded area all pairs (x,θ) for which an intersection occurs. Therefore, probability of an intersection is given by the ratio of these two areas. The area under the cosine curve is $A_1=L/2$ and the area of the rectangle is $A_2=a\pi/4$. Taking this ratio, we obtain the probability as:

$$P=A_1/A_2=\frac{2L}{a\pi} \quad . \tag{219}$$

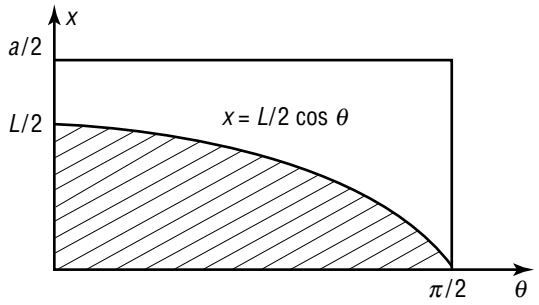


FIGURE 29.—Area ratio of Buffon's needle.

The following BASIC program simulates Buffon's needle experiment using equal needle length and grid distance (a=L=1). For this condition the probability of an intersection is obtained from the above for*m*ula as p=0.6362.

Using a run size of 10,000 for the simulation, we can calculate the maximum error of estimate, which is given by:

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \quad . \tag{220}$$

Using an 80-percent confidence level for which $z_{\alpha/2}$=1.28, we have a maximum error of approximately 0.006. The corresponding confidence interval is then:

$$0.6302 < p < 0.6422 \quad . \tag{221}$$

```
BUFFON'S NEEDLE

PI=4*ATN(1)
RANDOMIZE 123

L1:S=0: FOR I=1 TO 10000
X=RND
Y=COS(PI/2*RND)
IF X<Y THEN X=1:GOTO L2
X=0
L2:S=S+X:NEXT I

BEEP:BEEP
PRINT S:GOTO L1
END
```

The computer simulation gave the following 12 results: p=0.6315, 0.6325, 0.6355, 0.6352, 0.6382, 0.6382, 0.6353, **0.6230**, 0.6364, 0.6357, 0.6364, **0.6453**, 0.6362. We observe that 2 out of 12 runs fall outside the 80 percent confidence interval, which is what one would expect.

An actual experiment was supposed to have been carried out by an Italian mathematician named Lazzarini in 1901, who made 3,408 needle tosses and counted 2,169 intersections. This gave a value for π=3.142461964, which is only wrong in the third decimal place. One can easily recognize that it is highly unlikely to achieve such an accuracy with only 3,408 tosses. It is more likely that Lazzarini knew that π is approximately 22/7. If we substitute this number in the above expression for the probability, we get:

$$P = \frac{2\times 7}{22} = \frac{14\times 155}{22\times 155} = \frac{2170}{3410} \quad . \tag{222}$$

Of course, this might have been too obvious, so the numbers were slightly altered. What would have been the result if he had thrown one more needle?

III. STATISTICS

A. Estimation Theory

The concepts of probability theory presented in the preceding chapters begin with certain axioms (a self-evident fact that itself is not proved, but which is the basis of all other proofs) and derive probability laws for compound events. This is a *deductive* process. In statistics we are dealing with the inverse process, which is an *inductive* one: given certain observations, we want to draw conclusions concerning the underlying population. The study of such inferences is called *statistical inference*. *Probability* looks at the population and predicts a sample. *Statistics* looks at the sample and predicts the population.

1. Random Sample

In order that conclusions drawn from statistical inference are valid, it is necessary to choose samples that are representative of the population. The study of sampling methods and concomitant statistical analysis is called the *design of experiments.*

One can easily see that it is often difficult to get a random sample. For example, if we are after a random sample of people, it is not good enough to stand on a street corner and select every 10th person who passes. Many textbooks present the following classical example of a faulty sampling method. In 1936, a magazine called *Literary Digest* sent its readers and all telephone subscribers—approximately 10 million people—a questionnaire asking how they intended to vote at the forthcoming presidential election.

There were 2,300,000 replies, from which it was confidently predicted that the Kansas Republican Governor Alfred M. Landon would be elected. As it turned out, he lost every state but Maine and Vermont to the incumbent president Franklin D. Roosevelt.

This erroneous prediction happened because readers of this literary magazine and people who had telephones did not, in 1936, represent a fair, i.e., random sample of the American voters. It was especially risky to overlook the enormous proportion of nonreplies.

DEFINITION: If $X_1, X_2 \ldots X_n$ are independent and identically distributed (iid) random variables, they are said to constitute a random sample of size n. The joint distribution of this set of random variables is:

$$g(x_1,x_2 \ldots x_n)=f(x_1)\, f(x_2) \ldots f(x_n) \tag{223}$$

where f(x) is the *population* distribution.

The term population is used to describe the sample space (universe) from which the sample is drawn. Some statisticians use the term *parent population.*

2. Statistic

A statistic is a random variable that is a function of the random variables that constitute the random sample. The probability distribution of a statistic is called *sampling* distribution.

3. Estimator

An estimator is a statistic that is used to make a statistical inference concerning a population parameter θ:

$$\hat{\Theta}=\hat{\Theta}(X_1,X_2\ldots X_n) \quad . \tag{224}$$

B. Point Estimation

The value of an estimator provides a point estimate of a population paramater θ.

1. Properties of Estimators

There are often many possible functions capable of providing estimators. In the past the choice of an estimator was sometimes governed by mathematical expediency. With the arrival of high-speed computational means, this aspect is no longer as dominant as it used to be. For example, in on-line quality control the range was commonly used to estimate the dispersion of data rather than the standard deviation.

We list several desirable characteristics for good estimators as follows:

(a) Unbiasedness:

$$E(\hat{\Theta})=\theta \quad . \tag{225}$$

(b) Consistency (asymptotically unbiased):

$$\lim_{n\to\infty} P\left\{\left|\hat{\Theta}-\theta\right|<\varepsilon\right\}=1 \text{ or } \lim_{n\to\infty} E(\hat{\Theta})=\theta \quad . \tag{226}$$

(c) Efficiency: If

$$Var(\hat{\Theta}_1)<Var(\hat{\Theta}_2) \quad , \tag{227}$$

$\hat{\Theta}_1$ is more efficient than $\hat{\Theta}_2$ if *both are unbiased.*

(d) Sufficiency. An estimator is sufficient if it extracts all the information in a random sample relevant to the estimation of the population parameter θ. Mathematically expressed as the Fisher-Newman factorization theorem, $\hat{\Theta}=\hat{\Theta}$ $(x_1, x_2\ldots x_n)$ is sufficient if the joint p.d.f. of $X_1, X_2\ldots X_n$ can be factorized as:

$$f(x_1, x_2\ldots x_n \mid \theta)=g[\hat{\Theta} x_1, x_2 \ldots x_n \mid \theta]h(x_1, x_2 \ldots x_n) \tag{228}$$

where g depends on $x_1, x_2\ldots x_n$ only through $\hat{\Theta}$ and h is entirely independent of θ.

EXAMPLE 1: Let $X_1, X_2\ldots X_n$ be normally distributed. By setting $\mu=\theta_1$ and $\sigma^2=\theta_2$ and $\underline{\theta}=(\theta_1, \theta_2)$, we have:

$$f(x_1,x_2\ldots x_n\,|\,\theta)=\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n exp\left[-\frac{1}{2\theta_2}\sum_{j=1}^{n}(x_j-\theta_1)^2\right] \tag{229}$$

But
$$\sum_{j=1}^{n}(x_j-\theta_1)^2=\sum_{j=1}^{n}(x_j-\bar{x})^2+n(\bar{x}-\theta_1)^2 \tag{230}$$

so that

$$f(x_1,x_2\ldots x_n\,|\,\underline{\theta})=\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n exp\left[-\frac{1}{2\theta_2}\sum_{j=1}^{n}(x_j-\bar{x})^2-\frac{n}{2\theta_2}(\bar{x}-\theta_1)^2\right]\ . \tag{231}$$

It follows that $\left(\bar{X},\ \sum_{j=1}^{n}(X_j-\bar{X})^2\right)$ is sufficient for $\theta=(\theta_1,\ \theta_2)$.

EXAMPLE 2: Consider the uniform p.d.f.:

$$\begin{aligned} f(x)&=\frac{1}{\theta}\ \text{for}\ 0<x<\theta \\ &=0\ \text{otherwise} \end{aligned} \tag{232}$$

where θ is to be estimated on the basis of n independent observations. We want to prove that the statistic $\hat{\Theta}=\max(x_1, x_2\ldots x_n)$ is a sufficient estimator of the parameter θ.

SOLUTION: It can be shown that the cumulative distribution function of $\hat{\Theta}$ is

$$F_{\hat{\Theta}}(x)=\left(\frac{x}{\theta}\right)^n\ \text{ for } 0<x<\theta \tag{233}$$

and, therefore, the p.d.f. of $\hat{\Theta}$ is $f_{\hat{\Theta}}(x)=\dfrac{nx^{n-1}}{\theta^n}$ for $0<x<\theta$. Therefore, since the joint p.d.f. of X_1, $X_2\ldots X_n$ is $\left(\frac{1}{\theta}\right)^n$, which may be factored as

$$\left(\frac{1}{\theta}\right)^n=\left(\frac{nx^{n-1}}{\theta^n}\right)\frac{1}{nx^{n-1}}\ ,\ \text{or}\ f(x_1,\ x_2\ \ldots\ x_n\,|\,\theta)=g[\,\hat{\Theta}(x_1,x_2\ldots x_n\,|\,\theta)]\ h\ (x_1,\ x_2\ldots x_n) \tag{234}$$

as was to be shown. It is seen that the right-hand side of this expression has the desired product form such that the first term is a p.d.f. of the statistic $\hat{\Theta}$ and the second term does not depend on the parameter θ.

(e) Mean square error (MSE). The MSE of an estimator is defined as the expected value of the square of the deviation of the estimator from the parameter Θ being estimated. It can be shown that the MSE is equal to the variance of the estimator plus the square of its bias:

$$MSE \equiv E(\hat{\Theta}-\theta)^2 = E[\hat{\Theta}-E(\hat{\Theta})]^2 + [E(\hat{\Theta})-\theta]^2 \quad . \tag{235}$$

A biased estimator is often preferred over an unbiased estimator if its MSE is smaller. It is often possible to trade off bias for smaller MSE.

PROBLEM: Suppose $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two different estimators of the population parameter Θ. In addition, we assume that $E(\hat{\Theta}_1)=\Theta$, $E(\hat{\Theta}_2)=0.9\,\Theta$, Var $(\hat{\Theta}_1)=3$ and Var $(\hat{\Theta}_2)=2$. Which estimator has the smaller MSE?

SOLUTION:
$$\text{MSE}\,(\hat{\Theta}_1)=\text{Var}\,(\hat{\Theta}_1)+(\text{Bias})^2=3+0=3 \tag{236}$$

$$\text{MSE}\,(\hat{\Theta}_2)=\text{Var}\,(\hat{\Theta}_2)+(\text{Bias})^2=2+0.01\,\hat{\Theta}^2 \quad . \tag{237}$$

We notice that as long as $|\Theta| < 10$ the estimator $\hat{\Theta}_2$ has a smaller MSE and is, therefore, "better" than the estimator $\hat{\Theta}_1$, in spite of the fact that it is a biased estimator (see fig. 30).

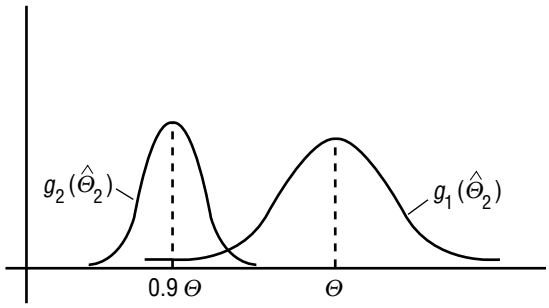


FIGURE 30.—Sampling distribution of biased and unbiased estimator.

EXAMPLE 1: Variance of normal distribution. A commonly used estimator for the variance of a distribution is the statistic:

$$S^2 = \frac{\sum(X_i-\overline{X})^2}{n-1} \quad . \tag{238}$$

The popularity of this estimator stems from the fact that it is an unbiased estimator for *any* distribution. However, it is important to realize that the corresponding sample standard deviation s, which is the positive square root of the variance s^2, is a *biased* estimator of the population standard deviation σ.

Another frequently used estimator of the variance of a population is the statistic:

$$S_B^2 = \frac{\sum(X_i-\overline{X})^2}{n} \quad . \tag{239}$$

If the sample is taken from a normal distribution, the above estimator is, in fact, the maximum likelihood estimator, but it is *not* unbiased.

The MSE of the two estimators, when calculated for a normal distribution with variance σ^2, is:

$$\text{MSE}(s^2)=\frac{2\sigma^4}{n-1} \text{ and } \text{MSE}(s_B^2)=\frac{(2n-1)\sigma^4}{n^2} \quad . \tag{240}$$

It can be readily verified that the MSE of the biased estimator is smaller than that of the unbiased estimator for all sample sizes *n*. For a sample size of *n*=5, the MSE of the unbiased estimator is 39 percent higher than that of the biased one.

EXAMPLE 2: Binomial parameter *p*. As a particular interesting example we compare two estimators available for the parameter *p* of the binomial distribution. The unbiased estimator of this parameter is:

$$\hat{P}u=\frac{x}{N} \tag{241}$$

with *x* being the number of successes and *N* the number of trials. This is sometimes called the "natural" estimator and is, coincidentally, also the maximum likelihood estimator. The other biased estimator is the Bayesian estimator obtained by using a uniform prior distribution and is given by:

$$\hat{p}_b=\frac{x+1}{N+2} \quad . \tag{242}$$

It can be shown that the variance of the unbiased estimator is $\text{Var}(\hat{p}_u)=\frac{pq}{N}$ and the variance of the Bayesian estimator $\text{Var}(\hat{p}_b)=\frac{Npq}{(N+2)^2}$. Clearly the variance of the biased estimator is smaller than the unbiased one. To obtain the MSE we need to determine the bias of the Bayesian estimator. Its value is given by:

$$\text{Bias}=\frac{1-2p}{N+2} \quad . \tag{243}$$

Figure 31 shows the absolute value of the bias of the estimator as a function of the parameter *p*.

It is seen that the bias is zero for *p*=0.5 and is a maximum for *p*=±1. Theoretically, the MSE of the biased estimator is smaller than that of the unbiased one for:

$$pq>\frac{N}{4(2N+1)} \quad . \tag{244}$$

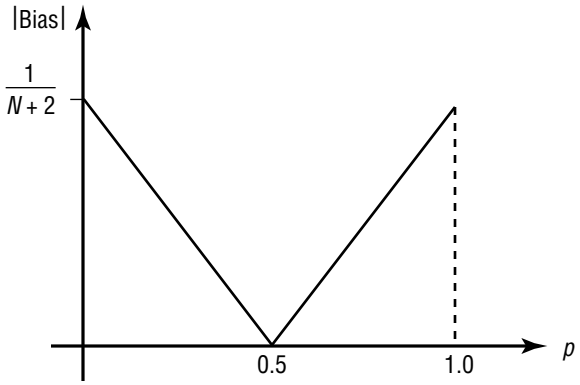


FIGURE 31.—Estimator bias as a function of parameter.

For $N \rightarrow \infty$, the critical value for p=0.854, implying that for greater values the unbiased estimator should be preferred because of its smaller MSE.

However, as the following BASIC Monte Carlo program demonstrates, the Bayesian estimator will *always* be better for the practical case regardless of the value of p. The discrepancy between the theoretical case and the real case stems from the fact that the theoretical variance of both estimators becomes very small for high values of the parameter p. As a consequence, the bias term becomes dominant and tilts the inequality in favor of the unbiased estimator.

```
'BINOMIAL ESTIMATOR (4/2/92)

BEGIN:
INPUT"N=",N
INPUT"P=",P
RANDOMIZE 1 2 3

'HERE STARTS THE MONTE CARLO RUN WITH N=100
L2:EB=0:EC=0: FOR R=1 TO 100

'THE NEXT 5 LINES ARE THE BERNOULLI TRIALS
S=0:FOR B=1 TO N
U=RND
IF U<=P THEN X=1:GOTO L1
X=0
L1:S=S+X:NEXT B

'CALCULATION OF MSE
PC=X/N:PB=(X+1)/(N+2)
EC=EC+(PC-P)^2:EB=EB+(PB-P)^2

NEXT R: BEEP:BEEP

PRINT EC,EB
GOTO L2
```

Sample run: N=5; p=0.90

MSE $(\hat{Pu}) = 53.16, 52.20, 53.16, 54.12, 50.28$

MSE $(\hat{Pb}) = 40.28, 39.69, 40.28, 40.87, 38.52$

It is seen that the MSE of the Bayesian binomial estimator is always smaller than that of the classical estimator.

2. Methods of Point Estimation

There are three methods of point estimation that will be discussed:

- Method of (matching) Moments (Karl Pearson, 1894)
- Method of Maximum Likelihood (R.A. Fisher, 1922)
- Method of Least Squares (C.F. Gauss, 1810).

(a) Method of Moments. The k^{th} *sample moment* is defined by:

$$m_k' = \frac{\sum x_i^k}{n} \text{ , where } n\text{=sample size .} \tag{245}$$

The k^{th} *population moment* is defined by:

$$\mu_k' = \int x^k f(x)\,dx \quad . \tag{246}$$

The method of moments consists of equating the sample moment and the population moment, which means we set $m_k' = \mu_k'$. In general, this leads to k simultaneous (nonlinear) algebraic equations in the k unknown population parameters.

EXAMPLE: Normal distribution:

$$\mu_1' = \mu \text{ and } \mu_2' = \sigma^2 + \mu^2 \quad . \tag{247}$$

Therefore:

$$\mu = \frac{\sum x_i}{n} \text{ and } \sigma^2 + \mu^2 = \frac{\sum x_i^2}{n} \quad . \tag{248}$$

From this follows:

$$\begin{aligned} \hat{\mu} &= \bar{x} = \frac{1}{n}\sum x_i \\ \hat{\sigma}^2 &= s^2 = \frac{1}{n}\left(\sum x_i^2 - n\bar{x}^2\right) = \frac{1}{n}\sum (x_i - \bar{x})^2 \ . \end{aligned} \tag{249}$$

(b) Method of Maximum Likelihood. The likelihood function of a random sample is defined by:

$$L(\theta)=f(x_1, x_2 \ldots x_n;\theta) \quad . \tag{250}$$

The method of maximum likelihood maximizes the likelihood function with respect to the parameter θ. Statistically speaking, this means that the maximum likelihood estimator maximizes the probability of obtaining the observed data.

EXAMPLE: Normal distribution:

$$L(\mu,\sigma^2)=\prod_{i=1}^{n} N(x_i;\mu,\sigma^2)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \quad . \tag{251}$$

Partial differentiation of the "log-likelihood" function $\ell=\ell n\ L$ with respect to μ and σ^2 yields:

$$\frac{\partial \ell}{\partial \mu}=\frac{1}{\sigma^2}\sum(x_i-\mu)=0 \tag{252}$$

$$\frac{\partial \ell}{\partial \sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}(x_i-\mu)^2=0 \tag{253}$$

Solving for the mean μ and the variance σ^2 we get:

$$\hat{\mu}=\frac{1}{n}\sum x_i=\bar{x} \text{ and } \hat{\sigma}^2=\frac{1}{n}\sum(x_i-\bar{x})^2=s^2 \quad . \tag{254}$$

It is interesting to observe that the method of moments and the method of maximum likelihood give identical estimators for a normally distributed random variable. This is, of course, not so in general.

Note that:

- The maximum likelihood estimators are often identical with those obtained by the method of moments.
- The set of simultaneous equations obtained by the method of likelihood is often more difficult to solve than the one for the method of moments.
- Maximum likelihood estimators, in general, have better statistical properties than the estimators obtained by the method of moments.

Invariance Property

If $\hat{\Theta}$ is a maximum likelihood estimator of the parameter θ, then $g(\hat{\Theta})$ is also a maximum likelihood estimator of $g(\theta)$. For example, if $\hat{\sigma}^2$ is a maximum likelihood estimator of the variance σ^2, then its square root $\hat{\sigma}$ is a maximum likelihood estimator of the standard deviation σ; i.e.,

$$\hat{\sigma}=\sqrt{\frac{1}{n}\sum\left(x_i-\bar{x}\right)^2} \quad . \tag{255}$$

(c) Method of Least Squares. The method of least squares is used in multivariate analyses like regression analysis, analysis of variance (ANOVA), response surface methodology, etc. The least-square method can be used to estimate the parameters of a two-parameter distribution by transforming the cumulative distribution to a straight line, using a probability scale and performing a least-square-curve fit to this line. In general, this is a rather unwieldly process unless the cumulative distribution can be obtained in closed form such as the Weibull distribution.

3. Robust Estimation

Over the last 2 decades statisticians have developed so-called *robust estimation* techniques. The term "robust" was coined by the statistician G.E.P. Box in 1963. In general, it refers to estimators that are relatively insensitive to departures from idealized assumption, such as normality, or in case of a small number of data points that are far away from the bulk of the data. Such data points are also referred to as "outliers." Original work in this area dealt with the location of a distribution.

An example of such a robust estimator is the so-called trimmed mean. In this case, the trimming portions are 10 or 20 percent from each end of the sample. By trimming of this kind, one removes the influence of extreme observations so that the estimator becomes more robust to outliers. (The reader might be familiar with the practice of disregarding the highest and the lowest data points for computing athletic scores.) A less severe method consists of allocating smaller weights to extreme observations in order to reduce the influence of any outliers. This leads to what is called a "Winsorized" mean. Further information on this subject matter can be found in *Data Analysis and Regression*, by Mosteller et al., and in *Robust Statistics: The Approach Based on Influence Functions* by Hampel et al.

Outliers. At some time or another, every engineer engaged in statistical data analysis will be confronted with the problem of outliers. Upon examination of the data, it appears that one or more observations are too extreme (high, low, or both) to be consistent with the assumption that all data come from the same population. There is no problem when it is possible to trace the anomalous behavior to some assignable cause, such as an experimental or clerical error. Unfortunately this is the exception rather than the rule. In the majority of the cases one is forced to make a judgment about the outliers, whether or not to include them, or whether to make some allowance on some compromise basis. Clearly it is wrong to reject an outlier simply because it seems highly unlikely or to conclude that some error in reading an instrument must have happened. In fact, in the history of science, attention given to anomalous observations has often led to startling new discoveries. For example, the discovery of nuclear fission involved enormous energy levels which were originally considered to be outliers. This temptation must be strongly resisted. Once one starts handpicking the data, the sample becomes a biased one. The best rule, therefore, is to never reject an outlier unless an assignable cause has been positively identified.

What actually is needed is some objective statistical method to decide whether or not an outlier does in fact exist. However, since the outlier problem is by its very nature concerned with the tails of a distribution, any statistical method will be very sensitive to departures from the assumed distribution. Due to the mathematical difficulties involved, the majority of statistical research has so far dealt with the normal distribution.

One of the earliest methods for dealing with outliers is known as Chauvenet's criterion (1863). This rather strange proposal is stated as follows: If in a sample of size n the probability of the deviation of the outlier from the mean is smaller than $1/2n$, the outlier is to be rejected.

Chauvenet's criterion is based on the concept of the *characteristic extreme* (largest or smallest value), which was introduced by Richard V. Mises. The *characteristic largest value* u_n is defined as:

$$nR(u_n)=n[1-F(u_n)]=1 \tag{256}$$

where n=sample size and

$Pr(u \le u_n)=F(u_n)$ is the cumulative distribution function.

$$Pr(u>u_n)=R(u_n)=1-F(u_n) \quad . \tag{257}$$

Accordingly, we expect that on the average, one observation in a sample of size n will exceed the value u_n.

Similarly, the *characteristic smallest value* u_1 is defined as:

$$nF(u_1)=1 \quad . \tag{258}$$

The number of observations n corresponding to the characteristic largest or smallest value is also called the *return period* $T(u_n)$ or $T(u_1)$, respectively. This means we expect the maximum or minimum value to occur once in a return period. The concept of the return period is sometimes used in engineering design.

Note that if an event has a probability of p, then the mathematical expectation for this event to happen is defined by $E=n\,p$. Therefore, if we set E=1, we can make the statement that it takes $1/p$ trials on the average for the event to happen once.

It is important to notice that the characteristic extreme value is a population parameter and not a random variable. The characteristic largest value is found by solving the above equation for $F(u_n)$:

$$F(u_n)=1-1/n \quad . \tag{259}$$

According to Chauvenet's criterion, data points will be rejected as outliers if they fall above the critical value defined by:

$$nR(u_n{}^c)=n\left[1-F\left(u_n{}^c\right)\right]=\frac{1}{2} \quad . \tag{260}$$

In other words, we expect that in a sample of size n, on the average "one-half" observation will exceed the critical value $u_n{}^c$. In the case where data points are rejected, the parameters of the distribution are recalculated without the rejected values. If we know the distribution of the population from which the sample is drawn, we can determine this critical value from the relation

$$F(u_n{}^c)=1-\,{}^{1}\!\!\diagup\!\!{}_{2n} \quad . \tag{261}$$

For a normal distribution $N(0, 1)$ we obtain:

$u_{50}^c = 1 - {}^{1}\!\!\diagup\!\!{}_{1000}$

$$n=50 \qquad Fu_{50}^c = 1-{}^{1}\!\!\diagup\!\!{}_{100} \qquad u_{50}^c = K_{0.99} = 2.33 \tag{262}$$

$$n=500 \qquad Fu_{500}^c = 1-{}^{1}\!\!\diagup\!\!{}_{100} \qquad u_{500}^c = 1-{}^{1}\!\!\diagup\!\!{}_{1000}\,. \tag{263}$$

It is seen that when applied to a normal distribution, Chauvenet's criterion is absolutely meaningless. At most, one could use this criterion to flag suspicious data points in a sample.

Return periods are often chosen to be equal to *twice* the service life of a structure when maximum loads or winds are fitted to an extreme value distribution (Gumbel distribution). It is seen that the basis for the factor two is actually Chauvenet's criterion.

C. Sampling Distributions

The probability distribution of a statistic *S* is called its *sampling distribution*. If, for example, the particular statistic is used to estimate the mean of a distribution, then it is called the sampling distribution of the mean. Similarly we have sampling distributions of the variance, the standard deviation, the median, the range, etc. Naturally, each sampling distribution has its own population parameters such as the mean, standard deviation, skewness, kurtosis, etc. The standard deviation of a sampling distribution of the statistic *S* is often called its *standard error* σ_S.

1. Sample Mean

If the random variables $X_1, X_2 \ldots X_n$ constitute a random sample of size *n*, then the sample mean is defined as:

$$\overline{X} = \frac{\sum X_i}{n} \quad . \tag{264}$$

Note that it is common practice to also apply the term statistic to values of the corresponding random variable. For instance, to calculate the mean of a set of observed data, we substitute into the formula:

$$\overline{x} = \frac{\sum x_i}{n} \tag{265}$$

where the x_i are the observed values of the corresponding random variables.

First, let us state some general results that hold for *arbitrary* distributions with mean μ and variance σ^2:

- Infinite population:

$$E(\overline{X}) = \mu_{\overline{x}} = \mu \text{ and } Var(\overline{X}) = \sigma_{\overline{x}}^2 = \frac{\sigma^2}{n} \quad . \tag{266}$$

- Finite population of size *N*:

$$E(\overline{X}) = \mu_{\overline{x}} = \mu \text{ and } Var(\overline{X}) = \sigma_{\overline{x}}^2 = \frac{\sigma^2}{n} \times \frac{N-n}{N-1} \tag{267}$$

Note that the information embodied in the sample is minimally affected by the proportion that the sample bears relative to its population and essentially determined by the sample size itself. In fact, the finite population correction factor is usually ignored if the sample size does not exceed 5 percent of the population.

- Large sample size ($N > 30$): As a consequence of the Central Limit Theorem, the sampling distribution of the mean approaches that of the normal distribution with mean μ and variance σ^2/n. The sampling distribution of the mean (fig. 32) is said to be asymptotically normal.

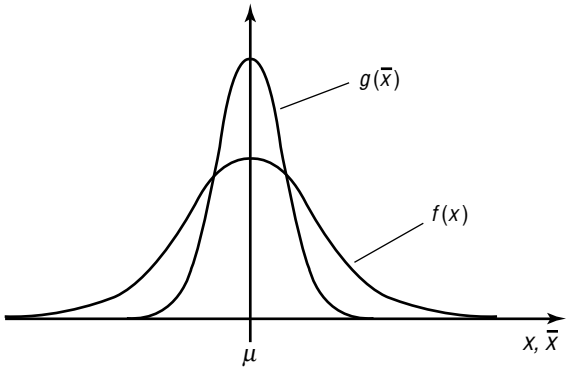


FIGURE 32.—Population and sampling distribution.

- Normal population (σ known): If the sample mean is taken from a normal distribution, then its sampling distribution is also normal regardless of the sample size. This is due to the reproductive property of the normal distribution.

- Normal population (σ unknown): If the variance of the normal distribution is not known, it is required to find a suitable statistic to estimate it. A commonly used estimate is the (unbiased) sample variance:

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} \quad . \tag{268}$$

It can be shown that the statistic:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{269}$$

has the Student-*t*, or *t*-distribution with $(n-1)$ degrees of freedom. This distribution was first obtained by William S. Gosset, a 32-year-old research chemist employed by the famous Irish brewer Arthur Guinness, who published it in 1908 under the pseudonym "Student." The *t*-statistic is also called a *t*-score.

Note that because of its smaller MSE, some statisticians prefer to use the (biased) sample variance:

$$S_B^2 = \frac{\sum (X_i - \overline{X})^2}{n} \quad . \tag{270}$$

In this case the *t*-statistic becomes:

$$T = \frac{\overline{X} - \mu}{S_B/\sqrt{n-1}} \quad . \tag{271}$$

The Student-*t* distribution is given by:

$$f(t)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad -\infty<t<+\infty \tag{272}$$

where the parameter $\nu=n-1$ is called the "degrees of freedom."

The variance of the *t*-distribution for $\nu>2$ is $\nu/(\nu-1)$ and the kurtosis for $\nu>4$ is $3+6/(\nu-4)$. It is seen that for large values of ν, the variance becomes 1 and the kurtosis becomes 3. In fact it can be shown that for $\nu\to\infty$, the *t*-distribution is asymptotically normal (see fig. 33).

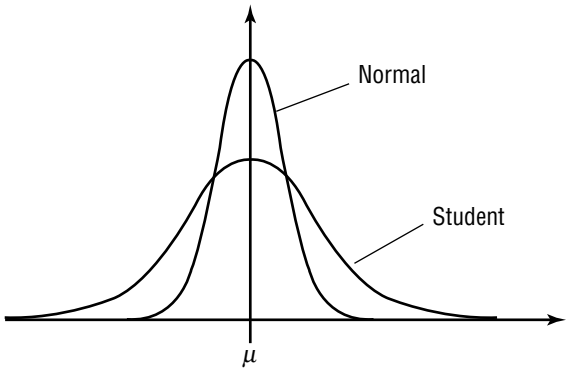


FIGURE 33.—Student versus normal distribution.

The *t*-distribution can be integrated in closed form, but not very easily. Tables are usually given for the inverse cumulative distribution; i.e., they present *t*-values for various percentile levels.

NOTE OF CAUTION: There exist two categories of *t*-distribution tables. One type shows percentile values t_p such that the area under the distribution to the left is equal to p; the other shows values of t_α such that the area under the distribution to the right; i.e., the tail area, is equal to α. A person who is somewhat familiar with the *t*-distribution should have no difficulty identifying which table to use.

2. Sample Variance

Theorem 1. If s^2 is the (unbiased) sample variance of a random sample of size n taken from a *normal* distribution with variance σ^2, then the statistic:

$$X^2=\frac{(n-1)S^2}{\sigma^2} \tag{273}$$

has a χ^2 distribution with $\nu=n-1$ degrees of freedom whose probability density is:

$$f\left(x^2\right)=\frac{1}{\Gamma\left(\frac{\nu}{2}\right)2^{\frac{\nu}{2}}}\left(x^2\right)^{\frac{\nu}{2}-1}e^{-\frac{x^2}{2}} \qquad 0<x^2<\infty \tag{274}$$

with mean $\mu=\nu$ and variance $\sigma^2=2\nu$.

Note that the χ^2 distribution (shown in fig. 34) is a special case of the gamma distribution with the parameters $\alpha=\nu/2$ and $\beta=2$.

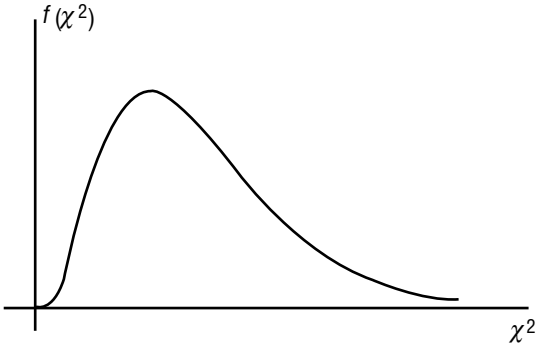


FIGURE 34.—χ^2 distribution.

The distribution of the sample variance s^2 itself is given by:

$$f(s^2)=\frac{\left(\frac{n-1}{2}\right)^{\frac{n-1}{2}}(s^2)^{\frac{n-3}{2}}e^{-\frac{(n-1)s^2}{2\sigma^2}}}{\sigma^{n-1}\Gamma\left(\frac{n-1}{2}\right)} \quad . \tag{275}$$

The cumulative χ^2 distribution is:

$$F\left(\chi^2\right)=\int_0^{\chi^2}f(x)\,dx \quad . \tag{276}$$

Theorem 2. If s_1^2 and s_2^2 are two sample variances of size n_1 and n_2, respectively, taken from two *normal* distributions having the *same* variance, then the statistic:

$$F=\frac{S_1^2}{S_2^2} \tag{277}$$

has a Fisher's *F*-distribution with:

$$\nu_1=n_1-1 \text{ and } \nu_2=n_2-1 \text{ degrees of freedom.} \tag{278}$$

The *F*-distribution is also known as the variance-ratio distribution:

$$g(F)=\frac{\Gamma\left[\frac{\nu_1+\nu_2}{2}\right]\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} F^{\frac{(\nu_1-2)}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\nu_1\frac{F}{\nu_2}\right)^{\frac{(\nu_1+\nu_2)}{2}}} \quad 0<F<\infty \tag{279}$$

with:

$$\text{mean } \mu=\frac{\nu_2}{\nu_2-2} \text{ for } \nu_2>2 \tag{280}$$

$$\text{variance } \sigma^2=\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ for } \nu_2>4 \quad . \tag{281}$$

REMARK: The *F*-distribution is related to the beta function as follows: If *X* has a beta distribution with parameters $\alpha=\frac{\nu_1}{2}$ and $\beta=\frac{\nu_2}{2}$, then the statistic:

$$F=\frac{\nu_2 X}{\nu_1(1-X)} \tag{282}$$

has an *F*-distribution with ν_1 and ν_2 degrees of freedom.

3. Sample Standard Deviation

For a sample size $n>30$ taken from an *arbitrary* population, the sampling distribution of the standard deviation *S* can be approximated by a normal distribution with:

$$\mu_s=\sigma \text{ and } \sigma_s^2=\frac{\sigma^2}{2(n-1)} \quad . \tag{283}$$

The standard normal random variable is then:

$$Z=\frac{S-\sigma}{\sigma/\sqrt{2(n-1)}} \quad . \tag{284}$$

D. Interval Estimation

Interval estimates are used to assess the accuracy or precision of estimating a population parameter θ. They are defined by the confidence (fiducial) intervals within which we expect a population parameter θ to be located.

- Two-sided confidence interval:

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha \ . \tag{285}$$

- One-sided confidence interval:
  - Lower confidence interval:

$$P(\hat{\Theta}_L < \theta) = 1 - \alpha \tag{286}$$

  - Upper confidence limit:

$$P(\theta < \hat{\Theta}_U) = 1 - \alpha \ . \tag{287}$$

The confidence interval is a random variable because the confidence limits (upper and lower endpoints) are random variables.

The probability $(1-\alpha)$ that the confidence interval contains the true population parameter θ is called the degree of confidence or the confidence (level). The corresponding interval is called a $(1-\alpha)$ 100-percent confidence interval. The quantity α is called the significance coefficient.

An analogy given by N.L. Johnson and F.C. Leone in *Statistics and Experimental Design* states that: "A confidence interval and statements concerning it are somewhat like the game of horseshoe tossing. The stake is the parameter in question. (It never moves regardless of some sportsmen's misconceptions.) The horseshoe is the confidence interval. If out of 100 tosses of the horseshoe one rings the stake 90 times on the average, he has 90-percent assurance (confidence) of ringing the stake. The parameter, just like the stake, is the constant. At any one toss (or interval estimation) the stake (or parameter) is either enclosed or not. We make a probability statement about the variable quantities represented by the positions of the arms of the horseshoe."

1. Mean

To obtain a two-sided confidence interval for the mean, we use the *t*-statistic, which is given by:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \quad \text{with} \quad \nu = n - 1 \tag{288}$$

and establish the following double inequality

$$P\left[-t_{\alpha/2} < T < t_{\alpha/2}\right] = 1 - \alpha \tag{289}$$

or

$$P\left[-t_{\alpha/2}<\frac{\overline{X}-\mu}{S/\sqrt{n}}<t_{\alpha/2}\right]=1-\alpha \quad . \tag{290}$$

Now we solve the inequality for the desired mean μ and obtain:

$$P\left[\overline{X}-t_{\alpha/2}\frac{S}{\sqrt{n}}<\mu<\overline{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right]=1-\alpha \quad \text{with } \nu=n-1 \quad . \tag{291}$$

This leads to the following $(1-\alpha)$ 100-percent confidence interval for the mean μ (shown in fig. 35):

$$\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}}<\mu<\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{with} \quad \nu=n-1 \quad . \tag{292}$$

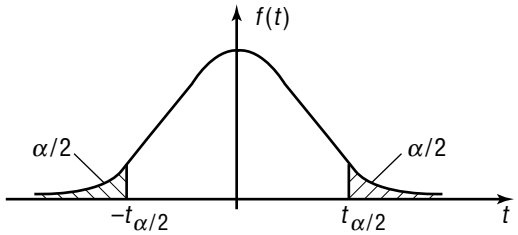


FIGURE 35.—Confidence interval for mean.

Following a similar approach as above, we can also establish one-sided upper or lower confidence intervals for the mean. For example, the one-sided upper $(1-\alpha)$ 100-percent confidence limit for the mean is:

$$\mu<\bar{x}+t_{\alpha}\frac{s}{\sqrt{n}} \quad \text{with } \nu=n-1 \quad . \tag{293}$$

For large sample size (n>30) we simply replace the t-statistic by the z-statistic. Because of the central limit theorem, the confidence interval for the mean can also be used for non-normal distributions provided n is sufficiently large.

The most widely used confidence levels $1-\alpha$ are the 95- and 99-percent levels. For large sample sizes, the corresponding two-sided standard scores (K factors) are $z_{0.025}$=1.96 and $z_{0.005}$=2.576.

If no confidence level is specified, the "one-sigma" interval (K factor is 1) is generally assumed. For a two-sided confidence interval, this represents a 68.27-percent confidence level.

Error of Estimate. The *error of estimate* E is defined as the absolute value of the difference between the estimator $\hat{\Theta}$ and the population parameter θ, i.e., $E=\left|\hat{\Theta}-\theta\right|$. The maximum error of estimate is equal to the half-width of the two-sided confidence interval. The maximum error of the mean is, therefore:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} \quad . \tag{294}$$

Note that the half-width of the two-sided, 50-percent confidence interval for the mean is known as the *probable error*. For large sample size *n*, the corresponding standard score (*K* factor) is $z_{0.25}$=0.6745.

The above formula can also be used to determine the sample size that is needed to obtain a desired accuracy of an estimate of the mean. We solve for the sample size *n* and get:

$$n = \left[t_{\alpha/2}\frac{s}{E}\right]^2 \quad . \tag{295}$$

2. Variance

Using the χ^2 statistic we can establish a confidence interval for the variance of a normal distribution by duplicating the steps we used for the mean (see fig. 36).

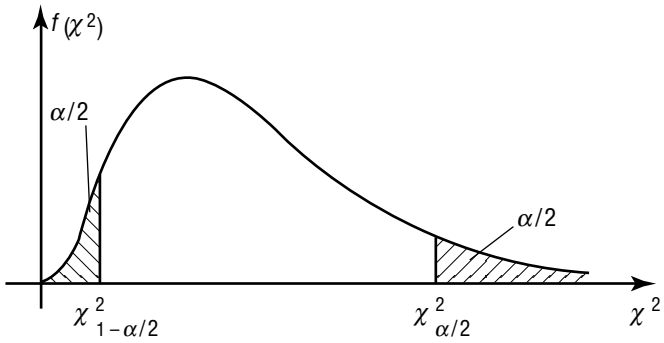


FIGURE 36.—Confidence interval for variance.

We obtain:

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right\} = 1-\alpha \quad . \tag{296}$$

This is an "equal tails" confidence interval. Actually it does not give the shortest confidence interval for σ^2 because of the asymmetry of the χ^2 distribution.

In general, it is desirable that the width of the confidence interval be as small as possible. However, the decrease in the interval, which is achieved by searching for the smallest interval, is usually not worth the labor involved.

The confidence of the standard deviation is obtained by simply taking the square root of the above formula. It is, of course, also possible to establish upper and lower confidence intervals just as in the case of the mean.

3. Standard Deviation

If the sample size is large (n>30), the sampling distribution of the standard deviation S of an *arbitrary* distribution is nearly normal, and we can use its sampling distribution to establish a (1–α) 100-percent confidence interval as:

$$\frac{s}{1+z_{\alpha/2}/\sqrt{2(n-1)}}<\sigma<\frac{s}{1-z_{\alpha/2}/\sqrt{2(n-1)}} \quad . \tag{297}$$

4. Proportion

(a) Approximate Confidence Limits. When the sample size n is large, we can construct an approximate confidence interval for the binomial parameter p by using the normal approximation to the binomial distribution. To this end we observe that the estimate $\hat{P}=\frac{X}{n}$ of the population parameter p has an asymptotically normal distribution with mean $E(\hat{P})=p$ and variance $\mathrm{Var}(\hat{P})=\frac{pq}{n}$. The equivalent z-score is:

$$Z=\frac{\hat{P}-E(\hat{P})}{\sqrt{Var(\hat{P})}}=\frac{\hat{P}-p}{\sqrt{\frac{pq}{n}}} \quad \text{where } q=1-p \quad . \tag{298}$$

The two-sided large sample (1–α) 100-percent confidence interval is then given by:

$$-z_{\alpha/2}<\frac{P-p}{\sqrt{\frac{p(1-p)}{n}}}<z_{\alpha/2} \quad . \tag{299}$$

This is a quadratic equation in the parameter p and could be solved to obtain the *approximate* confidence interval for p. But for the sake of simplicity, we make the further approximation of substituting $\hat{p}$ in the expression for the variance of $\hat{p}$ appearing in the denominator of the above equation. The *approximate* confidence interval is then determined by:

$$P\left\{\hat{P}-z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{Q}}{n}}<p<\hat{P}+z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{Q}}{n}}\right\}=1-\alpha \quad . \tag{300}$$

It is again of interest to find the sample size n that is necessary to attain a desired degree of accuracy for the estimate of the binomial parameter p. It is:

$$n=p(1-p)\left[\frac{z_{\alpha/2}}{E}\right]^2 \quad . \tag{301}$$

Note that if no information is available for the value of p, we set p=1/2 to obtain the maximum sample size.

EXAMPLE: Suppose we want to establish a 95-percent confidence that the error does not exceed 3 percent. Then we must have a sample size of:

$$n = \frac{1}{4}\left(\frac{1.96}{0.03}\right)^2 = 1{,}067 \quad . \tag{302}$$

(b) Exact Confidence Limits. In reliability work the concern is usually with one-sided confidence limits. For a large sample size n one can, of course, establish one-sided limits by the same method that was used above for the one-sided confidence limits of the mean. However, it is not very difficult to find the exact limits by resorting to available numerical algorithms. These are based upon the following mathematical relationships.

(1) The Upper Confidence Limit. P_U is determined from the solution to the following equation:

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k \left(1 - p_U\right)^{n-k} = \alpha \quad . \tag{303}$$

This equation can be solved for the upper limit p_U by using the equivalency of the binomial sum and the beta integral and then equating the beta integral to the F-distribution. The result is:

$$p_U = \frac{(x+1)\, F_\alpha\,(\nu_1, \nu_2)}{(n-x) + (x+1)\, F_\alpha\,(\nu_1, \nu_2)} \text{ with } \nu_1 = 2(x+1) \text{ and } \nu_2 = 2(n-x) \quad . \tag{304}$$

(2) The Lower Confidence Limit. P_L is determined from the solution to the following equation:

$$\sum_{k=x}^{n} \binom{n}{k} p_L^k (1 - p_L)^{n-k} = \alpha \tag{305}$$

and solved for the unknown lower limit p_U in terms of the F-distribution:

$$p_L = \frac{x}{x + (n-x+1) F_\alpha (\nu_1, \nu_2)} \quad \text{with } \nu_1 = 2(n-x+1) \quad \text{and} \quad \nu_2 = 2x \quad . \tag{306}$$

NOTE: Comparing these confidence limits for the binomial parameter p with the analogous Bayesian confidence limits, it is noticed that the upper limit corresponds to a prior beta distribution with α=1 and β=0, whereas the lower limit is obtained by setting these prior parameters to α=0 and β=1.

(3) Two-Sided Confidence Limits. The two-sided confidence limits are obtained by simply replacing the significance coefficient α by $\alpha/2$ for the corresponding upper and lower limits.

E. Tolerance Limits

Tolerance limits establish confidence intervals for a desired percentile point ξ_p where the subscript p indicates the proportion to the left of the percentile point of a distribution. They are used in design specifications ("design allowables") and deal with the prediction of a future single observation. The

following discussion is restricted to a *normal* distribution with sample mean $\bar{x}$ and sample standard deviation *S*.

1. One-Sided Tolerance Limits

Here we determine a confidence interval which guarantees a confidence of $(1-\alpha)$ that at least 100 *p* percent of the population lies below (or above) a certain limit. In contrast to the percentile point ξ_p, which is a population parameter, the tolerance limit itself is a random variable.

The one-sided *upper* tolerance limit (see fig. 37) is defined as:

$$\hat{F}_U = \overline{X} + K_1 S \quad . \tag{307}$$

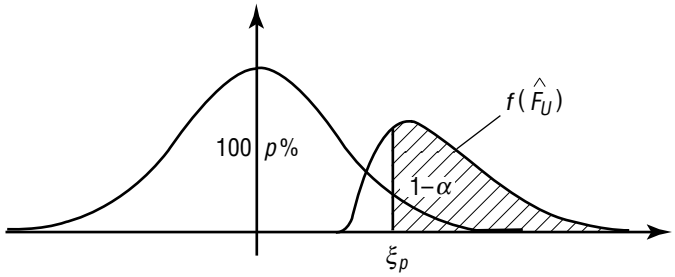


FIGURE 37.—One-sided upper tolerance limit.

Symbolically the above tolerance limit definition can be written as:

$$Pr\left[\Phi(\hat{F}_U) \geq p\right] = 1-\alpha \tag{308}$$

where Φ is the cumulative normal distribution with mean μ and standard deviation σ. This can be expressed somewhat simpler as:

$$Pr\left[\hat{F}_U \geq \xi_p\right] = 1-\alpha \quad . \tag{309}$$

Being a random variable, the tolerance limit has its own probability distribution $f(\hat{F}_U)$. For a normal distribution it can be shown that it is a noncentral *t*-distribution.

Theoretical work for other distributions is not well developed because of the associated mathematical and numerical difficulties. Some approximate tolerance limits have been determined for the Weibull distribution. References are listed in the Military Standardization Handbook MIL–HDBK–5F, Vol. 2, 1 November 1990. Another reference on tolerance limits is: *Statistical Tolerance Regions: Classical and Bayesian*, by I. Guttman.

The one-sided *lower* tolerance limit is defined as:

$$\hat{F}_L = \overline{X} - K_1 S \tag{310}$$

with an analogous interpretation for the left-hand tail of the normal distribution.

Symbolically, we can write:

$$Pr\left[\hat{F}_L \le \xi_1 - p\right] = 1-\alpha \quad . \tag{311}$$

Notice that for the normal distribution, we have the relation $\xi_1 - p = -\xi_p$.

2. Two-Sided Tolerance Limits

Here we select an interval such that the probability is $(1-\alpha)$ that at least 100 percent of the population is contained between the upper limit $\hat{F}_U = \overline{X} + K_2 S$ and the lower limit $\hat{F}_L = \overline{X} - K_2 S$.

Symbolically expressed, we have:

$$Pr\left[\Phi(\hat{F}_U) - \Phi(\hat{F}_L) \ge p\right] = 1-\alpha \quad . \tag{312}$$

The mathematical treatment of the two-sided case is surprisingly difficult. The distribution of the two-sided tolerance limits (see fig. 38) cannot be expressed in closed form and the numerical effort to calculate the exact tolerance limit is quite substantial. For this reason several approximations have been worked out and are given in tables. For example, see table 14 in *Probability and Statistics for Engineers* by Miller, Freund, and Johnson, listed in the bibliography.

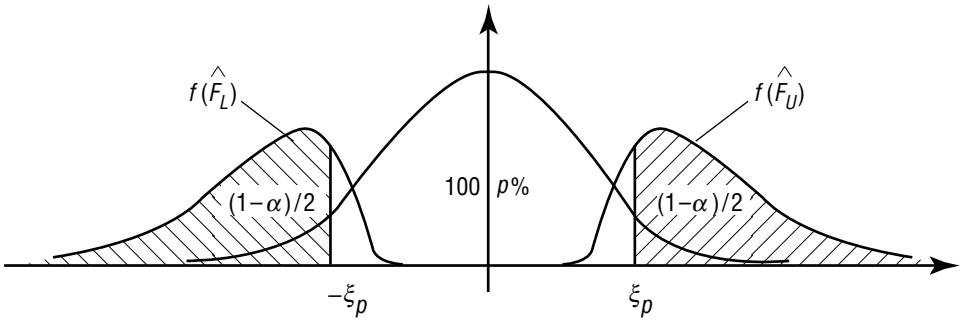


FIGURE 38.—Two-sided tolerance limits.

NOTE: Two particular tolerance limits have been widely used. Both are using a 95-percent confidence level. One is the so-called *A-Basic* (*A*-Allowables), which contains 99-percent of the population, and the other is the B-Basis (*B*-Allowables), which contains 90 percent of it.

F. Hypothesis/Significance Testing

Here we discuss mainly the Neyman-Pearson theory (1933), also called the classical theory of hypothesis testing. The main distinction between hypothesis testing and estimation theory is that in the latter we choose a value of a population parameter from a possible set of alternatives. In hypothesis testing, a predetermined value or set of values of a population parameter is either accepted or rejected.

DEFINITION: A statistical hypothesis is an assumption about a distribution. If the hypothesis completely specifies the distribution, it is called a *simple* hypothesis; otherwise, it is called a *composite* hypothesis.

EXAMPLE: Simple H:μ=10 (313)

Composite H:μ>10. (314)

1. Elements of Hypothesis Testing

There are four elements of hypothesis testing:

- Null hypothesis H_0: Baseline or "status quo" claim (design specifications)
- Alternative hypothesis H_1: Research claim, "burden of proof" claim
- Test statistic $\hat{\Theta}$
- Rejection (critical) region.

In subjecting the null hypothesis to a test of whether to accept or reject it, we are liable to make two types of error:

Type I error: Reject the null hypothesis H_0 when it is true.

Type II error: Accept the null hypothesis H_0 when it is false.

In the context of quality control of manufactured goods from a certain vendor, the Type I error is called the *producer's risk*, because it represents the probability of a good product; i.e., one which performs according to specifications, being rejected. On the other hand, the Type II error is called the *consumer's risk*, because it is the probability of a bad product being accepted.

Symbolically, we have the following *conditional* probabilities:

(R=reject H_0, A=accept H_0)

$$\alpha=Pr\,(R\,|\,H_0) \text{ and } 1-\alpha=Pr\,(A\,|\,H_0) \tag{315}$$

$$\beta=Pr\,(A\,|\,H_1) \text{ and } 1-\beta=Pr\,(R\,|\,H_1)\ . \tag{316}$$

The first column represents wrong decisions and the second column, right decisions. The probabilities associated with the right decisions are called the *confidence* $(1-\alpha)$ of the test and the *power* $(1-\beta)$ of the test.

In summary,

- Type I error: Reject good product (α)
- Type II error: Accept bad product (β)
- Confidence: Accept good product $(1-\alpha)$
- Power: Reject bad product $(1-\beta)$.

NOTE: The probability of a Type I error is also called the significance level. The following two levels are most common:

α=0.05, called "probably significant"

α=0.01, called "highly significant."

We will, subsequently, notice that the two types of errors have opposite trends; i.e., if one is made smaller, the other necessarily increases. *Both* errors can only be made smaller *simultaneously* by increasing the sample size, thereby increasing the amount of information available.

To illustrate the general concepts involved in hypothesis testing, we take the example of selecting a job applicant for a typing position. As our test statistic $\hat{\Theta}$, we take the average number $\bar{x}$ of typing errors per page. We define the null hypothesis H_0 by the mean number μ_0 of typing errors per page, which represents our requirement for the job. The alternative hypothesis H_1 is given by the number μ_1 of errors per page, which we consider unacceptable for the job.

The next and most important step is to select a criterion based on an actual typing test of the applicant that allows us to make a decision whether to accept or reject the applicant. We must categorically insist to select this criterion *before* the experiment is performed. The basic desire is to find an economical balance between the two types of errors. This is often not a simple matter because (for a given sample size) an attempt to decrease one type of error is accompanied by an increase in the other type of error. In practice the "losses" attached to each of the two types of errors have to be carefully considered. One type of error may also be more serious than the other.

In our example we will select a critical value μ_c somewhere between μ_0 and μ_1 to define the dividing line between the acceptance region and the rejection region. Accordingly, the applicant will be accepted if in an actual typing test he or she will have an average typing error $\bar{x}$ which is below the critical value μ_c. Figure 39 illustrates the given example.

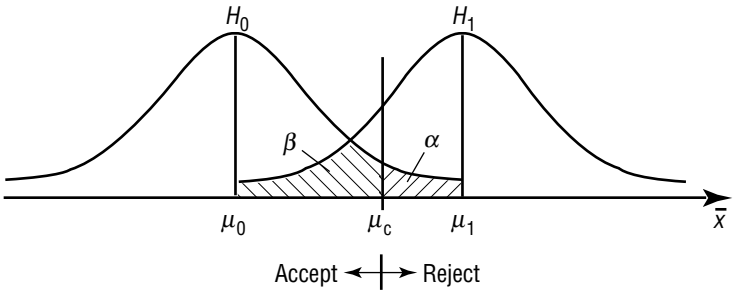


FIGURE 39.—Hypothesis test (H_0: $\mu=\mu_0$, H_1: $\mu=\mu_1$).

The different type of errors α and β and the associated consequences of making a wrong decision should be considered. For example, consider the difference between a civil service job and one in private industry in consideration of the prevailing policies pertaining the hiring and firing of an employee and the availability of applicants for the job to be filled.

NOTE: It is common practice to choose the null hypothesis so that the Type I error becomes more serious than the Type II error; in other words, we put the burden of proof on the alternative hypothesis H_1.

Which value to choose for the significance level depends on the risks and consequences associated with committing a Type I error.

In the following, we present a numerical example of hypothesis testing.

PROBLEM: A coin is tossed five times. Consider the hypotheses H_0: p_0=0.5 and H_1: p_1=0.7 in which p is the probability of obtaining a head. The null hypothesis H_0 is rejected if five heads come up. Determine the Type I and Type II errors.

SOLUTION:

$$H_0: p_0=0.5 \qquad H_1: p_1=0.7 \tag{317}$$

$$\alpha = \mathrm{B}(5 \mid 5, 0.5) = p_0^5 = 0.5^5 = 0.031=3.1 \text{ percent}, \tag{318}$$

where B represents the Binomial distribution.

Under the alternative hypothesis,

$$\beta = \sum_{x=0}^{4} \mathrm{B}(x|5, 0.7) = 1 - p_1^5 = 1 - 0.168=0.832=83.2 \text{ percent}. \tag{319}$$

Textbooks often cite the analogy that exists between hypothesis testing and the American judicial system. The null hypothesis in a criminal court is:

H_0: "The accused is innocent."

The system must decide whether the null hypothesis is to be rejected (finding of "conviction") or accepted (finding of "acquittal"). A Type I error would be to convict an innocent defendant, while a Type II error would be to acquit a guilty defendant. Clearly, the system could avoid Type I errors by acquitting all defendants. Also Type II errors could be avoided by convicting everyone who is accused. We must choose a strategy that steers a course between these two extreme cases. In quantitative terms we must ask what kind of "losses" are attached to each kind of error.

Our judicial system considers a Type I error much more serious: "It is better that 99 percent guilty persons should go free than that one innocent person should be convicted" (Benjamin Franklin, 1785). The very premise of the system (the accused is innocent until proven guilty beyond a reasonable doubt) reflects this concern of avoiding Type I errors, even at the expense of producing Type II errors.

In the American court of law, the judge often advises the jury to consider the following levels of confidence to arrive at a verdict:

- Preponderance of evidence (> 50 percent)
- Probable reason (90 percent)
- Clear and convincing evidence (95 percent)
- Beyond reasonable doubt (99 or 99.9 percent).

NOTE: Many problems people have stem from making Type II errors:

"It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so." Will Rogers (1879–1935).

"It is better to be ignorant, than to know what ain't so."

2. Operating Characteristic (OC) Curve

The operating characteristic function (see fig. 40) is the probability of accepting a bad product (Type II error) as a function of the population parameter θ:

$$L(\theta)=\beta \qquad L=\text{"Loss."} \tag{320}$$

Notice that when the parameter $\mu=\mu_0$ the null hypothesis and the alternative hypothesis coincide. In this case, the probability of the confidence and the consumer's risk are the same, which is to say $\beta=1-\alpha$.

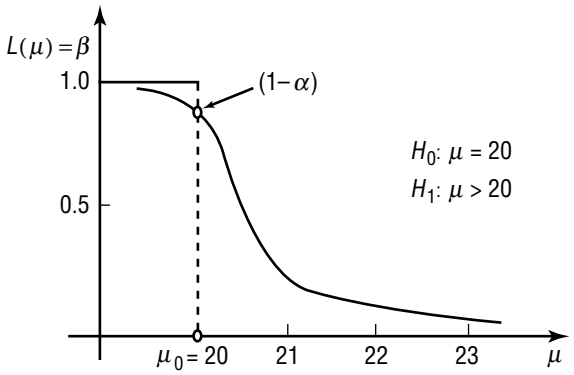


FIGURE 40.—Operating characteristic curve.

Sometimes it is more convenient to work with the power of the test. Remember that the power is the probability of rejecting the null hypothesis when the alternative is true. The so-called *power function* $\pi(\theta)$ of the test is related to the operating characteristic function by:

$$\pi(\theta)=1-L(\theta) \ . \tag{321}$$

Any test becomes more powerful as the sample size *n* increases.

3. Significance Testing

It is often difficult to specify a particularly meaningful alternative to the null hypothesis. It might also not be feasible in terms of resources and economy to determine the Type II error. In this case the alternative hypothesis H_1 is composite (*i.e.*, $H_1: \mu > \mu_0$). These tests are also called *null hypothesis* tests.

If the test statistic $\hat{\Theta}$ falls in the acceptance region, we are then reluctant to accept the null hypothesis because the protection against a Type II error (consumer's risk) is not known. Therefore, one usually prefers to say that the null hypothesis "cannot be rejected" or we have "failed to reject" the null hypothesis.

There are two different classes of hypothesis tests, depending on the type of rejection criterion that is used.

(a) One-Sided Criterion (Test). The null hypothesis is rejected if the value of the test statistic $\hat{\Theta}$ lies in one (upper or lower) tail of the sampling distribution associated with the null hypothesis H_0 (see fig. 41). This is also called a *one-tailed* test.

In case of a significance test, the alternative hypothesis H_1 is called *one-sided*:

$$H_0 : \theta = \theta_0, H_1 : \theta > \theta_0, \text{or } H_1 : \theta < \theta_0 \ . \tag{322}$$

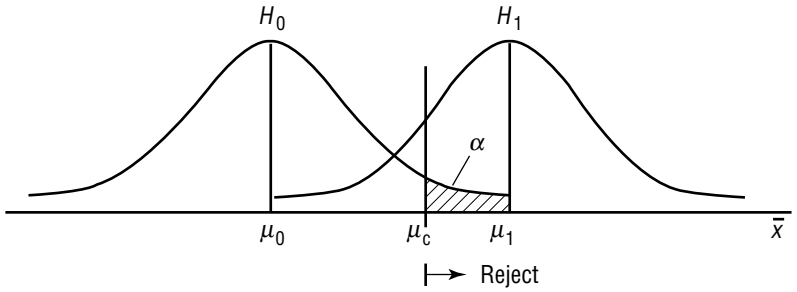


FIGURE 41.—One-sided hypothesis test.

(b) Two-Sided Criterion (Test). Here the null hypothesis is rejected if the value of the test statistic $\hat{\Theta}$ falls into either tail of the sampling distribution of the associated null hypothesis H_0 (see fig. 42).

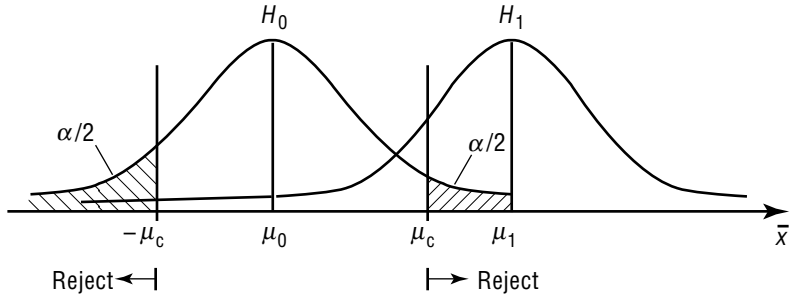


FIGURE 42.—Two-sided hypothesis test.

In case of a significance test, the alternative hypothesis H_1 is called *two-sided*:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0 \ . \tag{323}$$

Figure 43 illustrates the decision steps taken in a significance test.

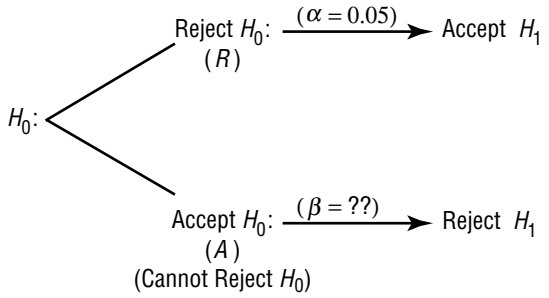


FIGURE 43.—Significance test (α=0.05)

EXAMPLE: Significance testing concerning one mean:

Test statistic: $$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\ ,\ \ \nu=n-1 \tag{324}$$

$$H_0 : \mu=\mu_0 \tag{325}$$

H_0	Reject H_0
$\mu < \mu_0$	$t < -t_\alpha$
$\mu > \mu_0$	$t > t_\alpha$
$\mu \neq \mu_0$	$t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$.

REMARK: Tests concerning other statistics can be found in most statistics reference books.

G. Curve Fitting, Regression, and Correlation

1. Regression Analysis

Regression assumes the existence of a functional relationship between *one* dependent *random* variable *y* and *one or more* independent *nonrandom* variables, as well as several unknown parameters:

$$Y = f(x_1, x_2 \ldots x_n; \theta_1, \theta_2 \ldots \theta_m) + E \tag{326}$$

Y = Response (criterion) variable
x_i = Prediction (controllable) variable
θ_j = Regression parameters
E = Random error.

The regression analysis consists of estimating the unknown regression parameters θ_j for the purpose of *predicting* the response *Y*. Estimation of the (population) regression parameters is done by the method of least squares, which is almost universally chosen for "curve fitting." Accordingly, we minimize:

$$S=\sum\left[y_i - f(x_1,x_2\ldots x_n;\theta_1,\theta_2\ldots\theta_m)\right]^2 \tag{327}$$

by setting the partial derivatives with respect to the *m* unknown parameters θ_j equal to zero. The resulting *m* simultaneous equations are called *normal equations*. They are, in general, nonlinear and difficult to solve unless the parameters enter linearly. Sometimes a suitable transformation can be found which "linearizes" the problem. Another approach of linearization is to find good initial estimates for the parameter, so that the nonlinear functions can be approximated by the linear terms of the series expansion.

2. Linear Regression

Linear regression is expressed as: $Y=\alpha + \beta x + E$.

We estimate the regression parameters α and β by the method of least squares (see fig. 44).

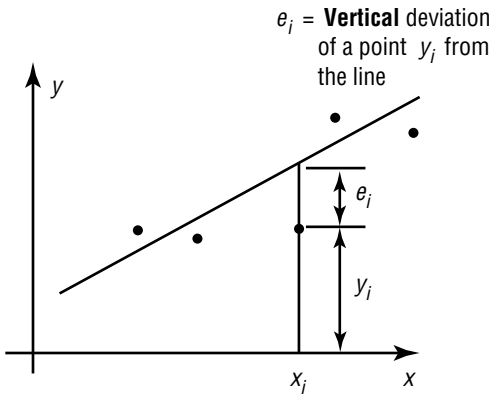


FIGURE 44.—Linear regression line.

Setting $S=\sum e_i^2=\sum\left[y_1-\left(\hat{\alpha}+\hat{B}x_i\right)\right]^2$, we solve $\frac{\partial S}{\partial\hat{\alpha}}=0$ and $\frac{\partial S}{\partial\beta}=0$. (328)

(a) The Normal Equations are given as:

$$\begin{aligned}\hat{\alpha}n+\hat{\beta}\sum x_i&=\sum y_i\\ \hat{\alpha}\sum x_i+\hat{\beta}\sum x_i^2&=\sum x_i y_i\ .\end{aligned} \tag{329}$$

(b) Abbreviations. We next introduce the following standard notation:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \left(\sum x_i\right)^2 / n$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 / n \tag{330}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n \, .$$

(c) Least-Square Estimators. The least-square estimators are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \text{ and } \hat{\beta} = \frac{S_{xy}}{X_{xx}} \text{ with } E(\hat{\alpha}) = \alpha \text{ and } E(\hat{\beta}) = \beta. \tag{331}$$

3. Gauss Markov Theorem

The Gauss Markov theorem states that:

- The estimators $\hat{\alpha}$ and $\hat{\beta}$ are unbiased.
- The estimators $\hat{\alpha}$ and $\hat{\beta}$ have the smallest variance, i.e., they are the most efficient ones.

This theorem is *independent* of the probability distribution of the random variable *y*. The variance σ^2 of the random error *E* is usually estimated by:

$$\hat{\sigma}^2 = s_e^2 = \frac{1}{n-2} \sum \left[y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right]^2 \, . \tag{332}$$

The division $n-2$ is used to make the estimator for σ^2 unbiased. The term s_e is called the *standard error of the estimate*.

The *computational form of* $\hat{\sigma}^2$ is:

$$\hat{\sigma}^2 = \frac{S_{yy} - \hat{\beta} S_{xy}}{n-2} \, . \tag{333}$$

4. Normal Regression Model

The normal regression model is expressed as:

$$f(y_i \mid x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{1}{2} \left[\frac{y_i - (\alpha + \beta x_i)}{\sigma} \right]^2 \text{ for } -\infty < y_i < \infty \, . \tag{334}$$

The maximum likelihood estimators for α, β, and σ^2 are identical to the least square estimators except that $\hat{\sigma}^2$ has the divisor *n* instead of $(n-2)$.

The following random variables are used to establish confidence intervals and to perform significance tests concerning the estimated regression parameters $\hat{\alpha}$ and $\hat{\beta}$.

Intercept $\hat{\alpha}$: $$t=\frac{\hat{\alpha}-\alpha}{S_e}\sqrt{\frac{nS_{xx}}{S_{xx}+\bar{x}^2}} \qquad \nu=n-2 \tag{335}$$

Slope $\hat{\beta}$: $$t=\frac{\hat{\beta}-\beta}{S_e}\sqrt{S_{xx}} \qquad \nu=n-2 \tag{336}$$

EXAMPLE: Slope β: H_0: $\beta=\beta_0$ (337)

Alternative hypothesis H_1	Reject H_0 if
$\beta < \beta_0$	$t < -t_\alpha$
$\beta > \beta_0$	$t > t_\alpha$
$\beta \neq \beta_0$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Note that if the independent variable is time, the regression line is also called a *trend line*.

Confidence interval about the regression line:

$$P\left[\hat{y}-t_{\alpha/2}\,s_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}<\bar{y}<\hat{y}+t_{\alpha/2}\,s_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right]=1-\alpha \quad (\nu=n-2)\ . \tag{338}$$

Prediction interval about the regression line:

$$P\left[\hat{y}-t_{\alpha/2}\,s_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}<\bar{y}<\hat{y}+t_{\alpha/2}\,s_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right]=1-\alpha \ \ (\nu=n-2)\ . \tag{339}$$

The minimum width occurs at $x=\bar{x}$ (see fig. 45).

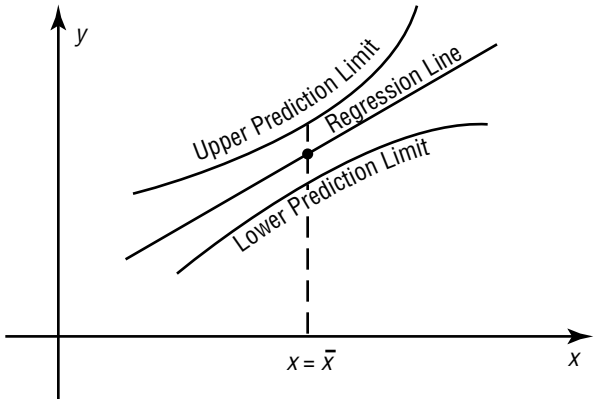


FIGURE 45.—Prediction limits of linear regression.

Note that the width of the prediction interval does not approach zero as n approaches infinity, expressing the inherent uncertainty of a future observation. Also, the intervals become increasingly wide for x—values outside the range of data.

5. Nonintercept Linear Mode

Sometimes both theoretical considerations or empirical data analysis indicate that the linear regression line goes through zero; i.e., we impose the condition that $\alpha=0$. The regression line (see fig. 46) is then simply given by:

$$y_i = b\, x_i \quad . \tag{340}$$

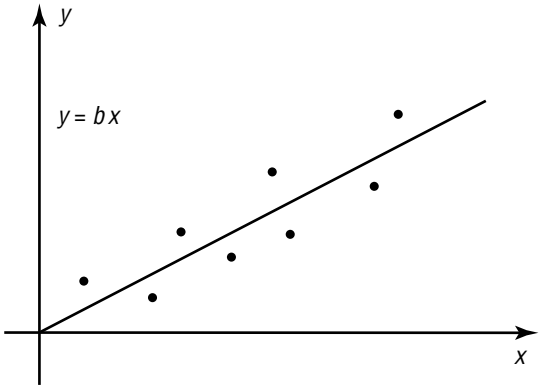


FIGURE 46.—Nonintercept linear regression model.

The regression parameter b is again determined by minimizing the error sum of squares:

$$\text{Minimizing } S=\sum(y_i - bx_i)^2 \quad \text{we set} \quad \frac{\partial S}{\partial b}=0 \quad \text{to obtain} \quad b=\frac{\sum x_i y_i}{\sum x_i^2}\,. \tag{341}$$

The parameter b is again unbiased, such that $E(b)=\beta$ and its variance is $\sigma_b^2=\dfrac{\sigma^2}{\sum x_i^2}$.

The *confidence* interval for the regression parameter β can then be established from the fact that the following statistic follows a *t*-distribution:

$$t=\frac{b-\beta}{s_e}\sqrt{\sum x_i^2} \quad \text{with } \nu=n-1 \text{ degrees of freedom} \tag{342}$$

and

$$s_e^2=\frac{1}{n-1}\sum\left(y_i-bx_i\right)^2 . \tag{343}$$

The *two-sided prediction* interval can be shown to be:

$$y=bx\pm t_{\alpha/2}\, s_e\sqrt{\left(1+x^2/\sum x_i^2\right)} \quad \text{with } \nu=n-1 \ . \tag{344}$$

The *one-sided prediction* interval is obtained by replacing $t_{\alpha/2}$ by t_α and choosing the proper sign for the upper or lower limit.

REMARK: If it is difficult to decide on physical grounds whether the model should contain an intercept term, it is common practice to initially fit an intercept model and perform a nullity test for the intercept term (i.e., test the null hypothesis H_0:α=0). If this term is not significant, then the model is refitted without the intercept term.

6. Correlation Analysis

It is assumed that the data points (x_i, y_i) are the values of a pair of random variables with joint probability density:

$$f(x, y)=g(y \mid x)\ h_1(x) \tag{345}$$

where:

$$g(y \mid x)=N(\alpha+\beta x,\sigma^2)$$

$$h_1(x)=N\left(\mu_1,\sigma_1^2\right) \tag{346}$$

$$h_2(y)=N\left(\mu_2,\sigma_2^2\right)$$

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma}exp-\left\{\frac{[y-(\alpha+\beta x)]^2}{2\sigma^2}+\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}. \tag{347}$$

It can be shown that: $\mu_2=\alpha+\beta\mu_1$ and $\sigma_2^2=\sigma^2+\beta^2\sigma_1^2$. (348)

Define: $\rho=1-\frac{\sigma^2}{\sigma_2^2}$ or $\rho=\sqrt{1-\frac{\sigma^2}{\sigma_2^2}}$. (349)

Also: $$\rho=\frac{\sigma_1}{\sigma_2}\beta \quad \text{and} \quad \sigma^2=\sigma_2^2(1-\rho^2) \ . \tag{350}$$

Then $$f(x,y)=\frac{exp-\left\{\frac{Q(x,y)}{2(1-\rho^2)}\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$ which is the Bivariate Normal Distribution, (351)

where

$$Q(x)=\left\{\left[\frac{x-\mu_1}{\sigma_1}\right]^2-2\rho\left[\frac{x-\mu_1}{\sigma_1}\right]\left[\frac{y-\mu_2}{\sigma_2}\right]+\left[\frac{y-\mu_2}{\sigma_2}\right]^2\right\}. \tag{352}$$

7. Other Regression Models

Three other regression models are polynomial, multiple, and exponential.

Polynomial regression:

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\ldots B_p x^p+\varepsilon \ . \tag{353}$$

Multiple regression:

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots B_r x_r+\varepsilon \ . \tag{354}$$

These two models are still linear regression models because they are linear in the unknown regression parameters.

Using matrix notation, we can write:

$$Y=X\,\underline{\beta}+\underline{\varepsilon} \tag{355}$$

$$S=\underline{\varepsilon}^{\mathrm{T}}\,\underline{\varepsilon}=(\underline{Y}-X\underline{\beta})^T(\underline{Y}-X\underline{\beta}) \ . \tag{356}$$

Minimizing S we obtain

$$\frac{\partial S}{\partial \underline{\beta}} \ = \ 2X^T\left(\underline{Y}-X\underline{\beta}\right)=0 \tag{357}$$

and

$$\underline{\hat{\beta}}=\left(X^T X\right)^{-1}X^T\underline{Y} \ . \tag{358}$$

The matrix $C=(X^T X)^{-1}X^T$ is called the pseudo-inverse of X.

For example, if we consider the exponential model:

$$y = \alpha\, e^{\beta x} \quad , \tag{359}$$

We "linearize" it by taking logarithims of both sides, to obtain:

$$\ell n y = \ell n \alpha + \beta x. \tag{360}$$

8. Sample Correlation Coefficient

Sample correlation coefficient scattergrams are shown in figure 47. The correlation coefficient allows a *quantitative* measure of how well a curve describes the functional relationship between the dependent and independent variables.

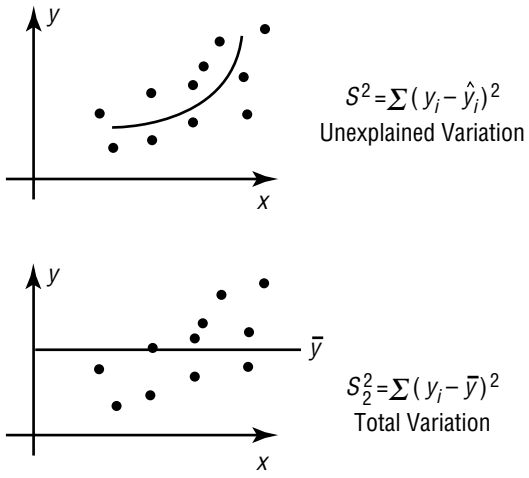


FIGURE 47.—Sample correlation coefficient (scattergrams).

Decomposition of the Total Sum of Squares. The derivation of the correlation coefficient is based on the fact that the total sum of squares of the deviation of the individual observations y_1 (i=1, 2 . . . n) from the mean $\bar{y}$ can be decomposed in two parts if the regression model is a polynomial of the form:

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots b_k x^k \, . \tag{361}$$

The error sum of squares is then given as:

$$\begin{aligned} S &= \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \\ &= \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x + b_2 x^2 + \ldots b_k x^k) \right]^2 \, . \end{aligned} \tag{362}$$

Partial differentiation with respect to the regression parameters β_i yields the normal equations:

$$\frac{\partial S}{\partial b_0} = -2\sum\left[y_i - \left(b_0 + b_1 x + \ldots b_k x^k\right)\right] = 0 \Rightarrow \sum \varepsilon_i = 0$$

$$\frac{\partial S}{\partial b_1} = -2\sum\left[y_i - \left(b_0 + b_1 x + \ldots b_k x^k\right)\right] x_i = 0 \Rightarrow \sum \varepsilon_i x_i = 0 \tag{363}$$

$$\frac{\partial S}{\partial b_k} = -2\sum\left[y_i - \left(b_0 + b_1 x + \ldots b_k x^k\right)\right] x_k = 0 \Rightarrow \sum \varepsilon_i x_i^k = 0 \,.$$

Therefore:

$$\sum \varepsilon_i \hat{y}_i = \sum \varepsilon_i \left(b_0 + b_1 x + \ldots b_k x^k\right) = \sum \varepsilon_i + \sum \varepsilon_i x_i + \ldots \sum \varepsilon_i x_i^k = 0 \,. \tag{364}$$

This is a significant result because it reveals that the correlation between the error ε and the estimate $\hat{y}$ is zero, or one can also say that all the information contained in the data set *y* has been removed.

The total sum of squares denoted by SST can then be written as:

$$\begin{aligned} \text{SST}=S_{yy} &= \sum\left(y_i - \bar{y}\right)^2 = \sum\left(y_i - \bar{y} + \hat{y}_i - \hat{y}_i\right)^2 \\ &= \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 + 2\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \,. \end{aligned} \tag{365}$$

The last term on the right-hand side can be seen to be zero as follows:

$$\sum\left(\hat{y}_i - \bar{y}\right)\left(y_i - \hat{y}_i\right) = \sum\left(\hat{y}_i - \bar{y}\right)\varepsilon_i = \sum \hat{y}_i \varepsilon_i - \bar{y}\sum \varepsilon_i = 0 \,. \tag{366}$$

Therefore, we have the final result that the total variation can be decomposed as follows:

$$\text{SST} = \sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 = \text{SSR} + \text{SSE} \tag{367}$$

where SSR is the regression ("explained") sum of squares and SSE the error ("unexplained") sum of squares.

The *coefficient of determination* is the ratio of the explained variation to the total variation. Since it is always positive, we denote it by r^2:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \hat{y})^2} = 1 - \frac{S^2}{S_2^2} \,. \tag{368}$$

Note that r^2 does not depend on the units employed because of its nondimensional nature.

The *positive* square root is called the (nonlinear) *correlation coefficient*:

$$r = \sqrt{1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}} \quad . \tag{369}$$

The correlation coefficient measures the goodness of fit of the curve. For instance, if $y_i \approx \hat{y}_i$, then r ≈ 1. When r ≈ 0, the relationship between the variables is poor with respect to the assumed regression model.

NOTE: The correlation coefficient as defined in equation (3) above can become imaginary for nonlinear regression when the assumed model does not fit very well. However, the upper limit is always 1.

9. Linear Correlation Coefficient

Here it is assumed that the relationship between the two variables is *linear*. Therefore:

$$\begin{aligned} S^2 &= \sum (y_i - \hat{y})^2 \text{ where } \hat{y}_i = a + bx_i \\ S^2 &= \sum y_i^2 - 2\sum (a + bx_i) y_i + \sum (a + bx_i)^2 \ . \end{aligned} \tag{370}$$

Remember:

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ b &= \frac{S_{xy}}{S_{xx}} \end{aligned} \tag{371}$$

$$S^2 = \sum y_i^2 - 2a\sum y_i - 2b\sum x_i y_i + na^2 + 2ab\sum x_i + b^2 \sum x_i^2 \tag{372}$$

$$S^2 = \sum y_i^2 - 2(\bar{y} - b\bar{x})\sum y_i - 2b\sum x_i y_i + n(\bar{y} - b\bar{x})^2 + 2(\bar{y} - b\bar{x})b\sum x_i + b^2 \sum x_i^2 \tag{373}$$

$$\begin{aligned} S^2 = \sum y_i^2 &- 2\bar{y}\sum y_i + 2b\bar{x}\sum y_i - 2b\sum x_i y_i + n\bar{y}^2 - 2nb\bar{y}\bar{x} + nb^2 \bar{x}^2 \\ &+ 2b\bar{y}\sum x_i - 2b^2 \bar{x}\sum x_i + b^2 \sum x_i^2 \end{aligned} \tag{374}$$

$$S^2 = \left[\sum y_i^2 - n\bar{y}^2\right] - 2b\left[\sum x_i y_i - n\bar{x}\bar{y}\right] + b^2\left[\sum x_i^2 - n\bar{x}^2\right] \tag{375}$$

$$S^2 = S_{yy} - 2bS_{xy} + b^2 S_{xy} = S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} \ . \tag{376}$$

Recall $S_2^2 = S_{xx}$ so that

$$r = \sqrt{1 - \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}S_{yy}}} = \sqrt{\frac{S_{xy}^2}{S_{xx}S_{yy}}} \tag{377}$$

$$r = \sqrt{\frac{S_{xy}^2}{S_{xx}S_{yy}}} \quad .$$

This is called the linear correlation coefficient. The linear correlation coefficient can also be written as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \, . \tag{378}$$

Usually the term correlation coefficient is used to mean *linear* correlation coefficient.

The sample correlation coefficient $r = \hat{\rho}$ is a biased estimator of ρ in the bivariate normal distribution.

NOTE: The method of maximum likelihood applied to the parameter σ_1, σ_2, and ρ of the bivariate normal distribution yields the same estimator for the correlation coefficient as above.

The correlation coefficient can assume positive and negative values ($-1 < r < 1$), depending on the slope of the linear regression line (see fig. 48).

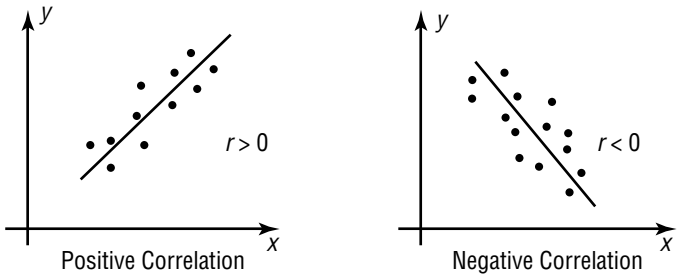


FIGURE 48.—Positive versus negative correlations.

From equation (368) we can write:

$$r^2 = 1 - \frac{s^2}{s_2^2} \tag{379}$$

or

$$s^2 = s_2^2 \left(1 - r^2\right) . \tag{380}$$

The variance s^2 is, so to speak, the "conditional" variance of y given x.

For $r = \pm 1$, the conditional variance $s^2 = 0$; i.e., the data points fall on a straight line (perfect correlation).

Sometimes equation (379) is written in the form:

$$r^2 = \frac{S_2^2 - S^2}{S_2^2} \times 100 \,\text{percent} \,. \tag{381}$$

Thus, 100 r^2 is the percentage of the total variation of the dependent variable of the dependent variable y which is attributed to the relationship with x. This is also true for the *nonlinear* case.

EXAMPLE: If r=0.5, then 25 percent of the variation of y is due to the functional relationship with x. Or we might say that a correlation of r=0.6 is "nine times as strong" as a correlation of r=0.2.

NOTE: There are several serious pitfalls in the interpretation of the correlation coefficient. It has often been said that it is the most abused statistical quantity.

PITFALL 1: The *linear* correlation coefficient is an estimate of the strength of the *linear* association between the random variables. See figure 49 for a quadratic relationship with zero correlation.

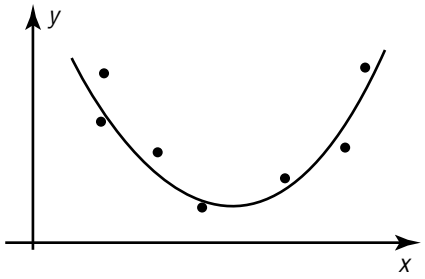


FIGURE 49.—Quadratic relationship with zero correlation.

PITFALL 2: A significant correlation does not necessarily imply a *causal* relationship between the two random variables. For example, there may be a high correlation between the number of books sold each year and the crime rate in the United States, but crime is not caused by reading books (spurious correlation).

Often two variables are having a mutual relationship with a third variable (e.g., population size) which produces the correlation. These cases can sometimes be treated by "partial" correlation.

10. Sampling Distributions of *r*

To establish confidence intervals and to perform significance tests, one would need to know the probability distribution of r for random samples from a bivariate normal population. This distribution is rather complicated. However, R.A. Fisher (1921) found a remarkable transformation of r which approximates the normal distribution.

11. Fisher *Z*-Transformation

The Fisher Z-transformation is given as:

$$z = \frac{1}{2} \ell n \frac{1+r}{1-r} = tanh^{-1} r \tag{382}$$

with

$$\mu_z = \frac{1}{2}\ell n\frac{1+\rho}{1-\rho} \quad \text{and} \quad \sigma_2^2 = \frac{1}{n-3} \quad \text{for} \quad n > 30 \ . \tag{383}$$

EXAMPLE:

$$r=0.70 \qquad n=30 \tag{384}$$

$$Z=0.867 \qquad \alpha=0.05 \tag{385}$$

CONFIDENCE INTERVAL:

$$Z - \frac{Z_{\alpha/2}}{\sqrt{n-3}} < \mu_z < Z + \frac{Z_{\alpha/2}}{\sqrt{n-3}} \tag{386}$$

where $Z_{\alpha/2}=1.96$ for 95-percent confidence:

$$0.867 - \frac{1.96}{\sqrt{27}} < \mu_z < 0.867 + \frac{1.96}{\sqrt{27}} \tag{387}$$

$$0.490 < \mu_z < 1.244 \tag{388}$$

or for the correlation coefficient:

$$0.45 < \rho < 0.85 \ . \tag{389}$$

Estimates of correlation coefficients based on small sample sizes are *not very reliable*. For instance, if the sample size $n=20$, the calculated correlation coefficient has to exceed the critical value $r_c=0.377$ to be significant at $\alpha=0.05$.

H. Goodness-of-Fit Tests

Goodness-of-fit tests examine how well a sample of data agrees with a proposed probability distribution. Using the language of hypothesis testing, the null hypothesis H_0 is that the given data set follows a specified distribution $F(x)$. In most applications the alternative hypothesis H_1 is very vague and simply declares that H_0 is wrong. In other words, we are dealing with significance tests in which the Type II error or the power of the test is not known. Moreover, in contrast to the usual significance tests where we usually look for the rejection of the null hypothesis in order to prove a research claim, goodness-of-fit tests actually are performed in the hope of accepting the null hypothesis H_0.

To make up for the lack of existence of a well-defined alternative hypothesis, many different concepts have been proposed and investigated to compare the power of different tests. The result of this quite extensive research indicates that uniquely “best” tests do not exist. There are essentially two different approaches to the problem of selecting a proper statistical model for a data set—probability plotting (graphical analysis) and statistical tests.

1. Graphical Analysis

Graphical techniques can be easily implemented and most statistical software packages furnish specific routines for probability plotting. They are very valuable in exploratory data analysis and in combination with formal statistical tests. In the former, the objective is to discover particular features of the underlying distribution, such as outliers, skewness, or kurtosis (i.e., thickness of the tails of the distribution). The old saying that a picture is worth a thousand words is especially appropriate for this type of analysis.

Of particular importance is the so-called *empirical distribution function* (E.D.F.) or sample cumulative distribution. It is a step function that can be generally defined by:

$$F_n(x_i)=\frac{i-c}{n-2c+1} \text{ for } o\le c\le 1 \tag{390}$$

with the observed ordered observations $x_1 \le x_2 \ldots \le x_n$. Several values for the constant c are in vogue for the plotting (rank) position of the E.D.F. The "midpoint" plotting position (c=0.5) has been found to be acceptable for a wide variety of distributions and sample sizes. The "mean" plotting position (c=0) is also often used. Another one is the so-called "median" rank position which is well approximated by c=0.3 (Benard and Bos-Levenbach, 1953). In practice, it does not make much difference which plotting position is used, considering the statistical fluctuation of the data. However, one should consistently use one particular plotting position when comparing different samples and goodness-of-fit tests.

A major problem in deciding visually whether an E.D.F. is conforming to the hypothesized distribution is caused by the curvature of the ogive. Therefore, it is common practice to "straighten out" the plot by transforming the vertical scale of the E.D.F. plot such that it will produce a straight line for the distribution under consideration. A probability plot is a plot of:

$$z_i=G^{-1}\left(F_n\left(x_i\right)\right) \tag{391}$$

where $G^{-1}(.)$ is the inverse cumulative distribution. Sometimes the z-score is placed on the horizontal axis and the observations x_i on the vertical axis. By proper labeling of the transformed scale, one obtains the corresponding *probability graph paper* which is commercially available or can be generated with the computer graphics package.

2. χ^2 Test

This test is the oldest and most commonly used procedure for examining whether a set of observations follows a specified distribution. The theoretical work underlying the χ^2 test was done in 1875 by the German physicist Friedrich Helmert. The English statistician Karl Pearson (1857–1936) demonstrated its application as a goodness-of-fit test in 1900. The major advantage of the test is its versatility. It can be applied for both discrete and continuous distributions without having to know the population parameters. The major drawback is that the data have to be grouped into a frequency distribution (histogram), which requires a fairly large number of observations. It is also usually less powerful than tests based on the EDF or other special purpose goodness-of-fit tests. The test statistic uses the observed class frequencies O_i of the histogram and the expected theoretical class frequencies E_i, which are calculated from the distribution under consideration, and is defined by:

$$\chi^2=\sum_{i=1}^{k}\frac{\left(O_i-E_i\right)^2}{E_i}=\sum_{i=1}^{k}\frac{O_i^2}{E_i}-n \text{ with } E_i>5 \tag{392}$$

where k is the number of class intervals of the histogram and n the total number of observations. The following constraint exists between the observed and the expected frequencies:

$$\sum O_i = \sum E_i = n \,. \tag{393}$$

The sampling distribution of this statistic is approximately the χ^2 distribution with degrees of freedom $\nu = k - 1 - m$ where m is the number of population parameters that have to be estimated from the data. For this approximation to be valid, the number of expected frequencies E_i should be greater than five. If this number is smaller, then adjacent class intervals can be combined ("pooled"). Class intervals do not have to be of equal size.

The χ^2 test statistic is intuitively appealing in a sense that it becomes larger with increasing discrepancy between the observed and the expected frequencies. Obviously, if the discrepancy exceeds some critical value we have to reject the null hypothesis H_0 which asserts that the data follow a hypothesized distribution. Since the alternative hypothesis is not specified, this procedure is also called the χ^2 test of significance. Typical significance levels are α=0.10, 0.05, and 0.01. When selecting a significance level it is important to keep in mind that a low level of significance is associated with a high Type II error, i.e., the probability of assuming that the data follow a suggested distribution when they are really not. Therefore the higher we choose the significance level, the more severe is the test. This aspect of the goodness-of-fit test is, at first sight, counterintuitive, because, as was mentioned above, we are usually interested in rejecting the null hypothesis rather than in accepting it.

Table 5 illustrates the procedure of applying the χ^2 test. It shows a table of the observed and expected frequencies of tossing a die 120 times.

TABLE 5.—*Procedure of applying the χ^2 test.*

Die face	1	2	3	4	5	6
Observed frequency	25	17	15	23	24	16
Expected frequency	20	20	20	20	20	20

The test statistic yields χ^2=5.00. Since the number of class intervals is 6 and no population parameter is estimated, the degrees of freedom are $\nu = 6 - 1 = 5$. If we choose a significance level of α=0.05 the critical value is $\chi^2_{0.05}$=11.1. Therefore, we accept the null hypothesis, which means we assume that the die is fair.

We must also look with skepticism upon a χ^2 value that is unreasonably close to zero. Because of the natural statistical fluctuation of the data, we should not expect the agreement between the observed and the expected frequencies to be too good.

The problem of a small χ^2 value is illustrated by the strange case of monk Gregor Mendel's peas. Writing in 1936, the famous English statistician R.A. Fisher wondered if, given the statistical fluctuations of experimental data in the field of genetics (Mendel, 1822–1884), Mendel's results were too good. In effect, Fisher tested the left-hand side of the χ^2 distribution for a too low χ^2 value. Examining Mendel's data conducted over an 8-year interval, he found the χ^2 value to be χ^2=41.606 with 84 degrees of freedom. This corresponds to a level of significance of $\alpha = 2.86 \times 10^{-5}$. Therefore one would expect to

find such a low χ^2 value only three times in 100,000 experiments. Perhaps Mendel was deceived by an overly loyal assistant who knew what results were desired.

It has been said by some critics of the χ^2 test, that it will always reject the null hypothesis if the sample size is large enough. This must not necessarily be the case.

3. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (see fig. 50) is one of many alternatives to the χ^2 test. The test is called an E.D.F. test because it compares the empirical distribution function with the theoretical cumulative distribution. The E.D.F. for this test is defined as:

$$F_n(x_i)=i/n \tag{394}$$

where the x_i's are the *ordered n* observations. With each increase of an x value, the step function takes a step of height $1/n$. The test statistic is:

$$D=\max|F_n(x_i)-F(x)| \quad . \tag{395}$$

If D exceeds a critical value D_α, the null hypothesis is rejected.

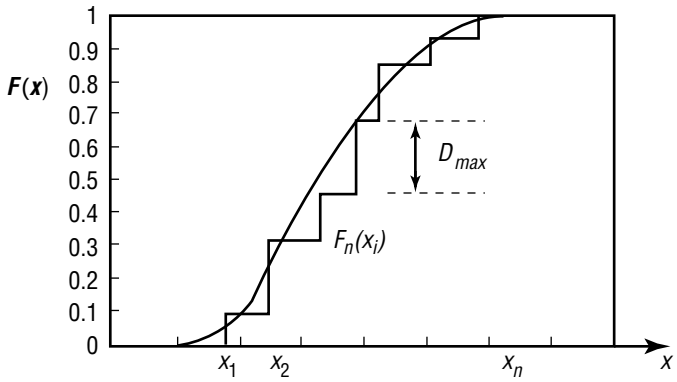


FIGURE 50.—Kolmogorov-Smirnov test.

The Kolmogorov-Smirnov test can also be used for discrete distributions. However, if one uses the tables which assume that the distribution is continuous, the test turns out to be conservative in the sense that if H_0 is rejected, we can have greater confidence in that decision. In cases where the parameters must be estimated, the Kolmogorov-Smirnov test is not applicable. Special tables exist for some particular distributions such as the exponential and the normal distribution. For a normal distribution for which the mean and variance have to be estimated, the following asymptotic critical D values (table 6) can be used if the sample size $n > 20$.

TABLE 6.—*Normal distribution.*

α	0.01	0.05	0.10
D_c	$\frac{1.031}{\sqrt{n}}$	$\frac{0.886}{\sqrt{n}}$	$\frac{0.805}{\sqrt{n}}$

Other E.D.F. statistics that measure the difference between $F_n(x)$ and $F(x)$ are found in the literature. A widely used one is the quadratic statistic:

$$Q = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 \psi(x) dF(x) \tag{396}$$

where $\psi(x)$ is a suitable weighting function for the squared differences under the integral. When $\psi(x)=1$, the statistic is the Cramer von Mises statistic, now usually called W^2, and when $\psi(x)=[F(x)\{1-F(x)]^{-1}$ the statistic is the Anderson-Darling (1954) statistic, called A^2. The E.D.F. test associated with the latter statistic is recommended in Volume 2 of MIL–HDBK–5F for testing the normality of data.

I. Quality Control

In recent years engineers have witnessed a dramatic revival of interest in the area of statistical quality control (SQC). One of its new features is a "process orientation" according to which emphasis is shifted towards the improvement of a product *during* the engineering design and manufacturing phases rather than attempting to control quality by inspecting a product *after* it has been manufactured. Time-honored maxims are being quoted and rediscovered. Some of them are: "It is more efficient to do it right the first time than to inspect it later." "One should fix processes and not products." "All or no inspection is optimal." The oldest and most widely known is, of course, that "*one cannot inspect quality into a product*."

In addition, the engineering design phase is being modernized and improved by recognizing the importance of applying the concepts of experimental design, which had been miserably neglected or overlooked in the field of engineering in the past. This is now often referred to as off-line quality control. Someone has recently remarked that if engineers would have a better background in engineering probability and statistics, it would not have been possible for Professor Genichi Taguchi to dazzle them with what are essentially elementary and simple concepts in the design of experiments that had been known for a long time by practicing statisticians. And last, but not least, a new element has been introduced called "the voice of the customer" (VOC), which is the attempt to get some feedback from the consumer about the product.

SQC centers primarily around three special techniques: (1) determining tolerance limits in the design phase, (2) setting up control charts for on-line quality control or statistical process control (SPC), and (3) devising acceptance sampling plans after the product has been manufactured.

1. Acceptance Sampling

Acceptance sampling is designed to draw a statistical inference about the quality of a lot based on the testing of a few randomly selected parts. While acceptance sampling is usually considered part of quality control, it is very important to understand that it hardly exercises direct leverage over process control. Because many contracts still contain requirements for submitting acceptance sampling plans, it is desirable for an engineer to know the basic underlying statistical concepts.

Statistical methods of acceptance sampling are well developed for a variety of sampling procedures such as single, multiple, and sequential sampling. A single sampling plan consists of drawing a random sample and develops criteria for accepting or rejecting the lot. In a multiple sampling plan, the sampling occurs in several stages. At each stage a decision is made whether to accept or reject the lot or whether to continue taking another sample. The concept of multiple sampling can be generalized to sequential

sampling, in which a decision is made to accept, reject, or continue sampling after each observation. It has turned out, against original expectations, that sequential sampling can significantly reduce the amount of inspection.

Acceptance sampling is further classified, depending on whether it applies to attribute scales or measurement (variable) scales. In general, inspection by attributes is less expensive than by measurements. However, inspection by variables provides more information than attributes and can be better used to improve quality control.

To facilitate the design and use of acceptance sampling plans, several standard plans have been published. Among the most widely used are the Military Standard 105D Tables (1963) for attribute sampling and the Military Standard 414 Tables (1957) for measurement sampling.

The subsequent discussion will be limited to single sampling plans.

- n=Sample size
- N=Lot size
- x=Number of defective items in sample
- c=Acceptance number (accept lot when $x \le c$)
- $H_0 : p_0$=Acceptable quality level (AQL) denoting a "good" lot
- $H_1 : p_1$=Lot tolerance percent defective (LTPD) denoting a "bad" lot
- α=Pr (Reject $H_0 \mid p_0$)=producer's risk=probability of rejecting a good lot
- β=Pr (Accept $H_0 \mid p_1$)=consumer's risk=probability of accepting a bad lot.

A given sampling plan is best described by its OC curve, which is the probability of the consumer's risk for each lot proportion defective p. This probability can be calculated using the hypergeometric distribution as follows:

$$h(x)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, \tag{397}$$

where $D=N_p$=defective items in the lot.

The cumulative hypergeometric distribution yields the probability of accepting a lot containing the proportion of defectives p:

$$L(p)=\sum_{x=0}^{c} h(x \mid n,N,D) . \tag{398}$$

For large lot sizes it is acceptable to approximate the hypergeometric distribution by the binomial distribution. A sketch of the OC curve is given in figure 51. The OC curve always passes through the two points (p_0, α) and (p_1, β), which are reached by agreement between the consumer and producer.

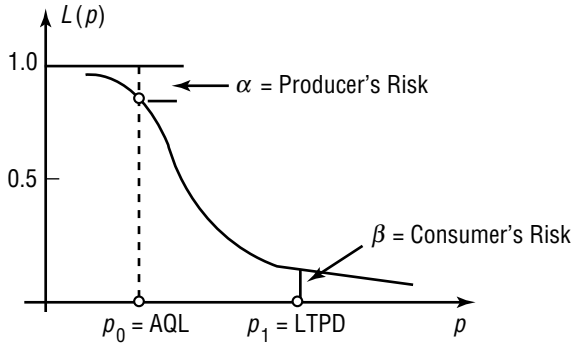


FIGURE 51.—OC curve for a single sampling plan.

Sometimes rejected lots are screened (100-percent inspection) and all defective items are replaced by good ones. This procedure is called *rectifying* inspection. In this case the sampling plan is described by its *average outgoing quality* or AOQ curve. The formula for it is derived under the above assumption that all defectives in a rejected lot are replaced by good items before their final acceptance. Accordingly we have:

$$\text{AOQ}=p\,L(p) + 0 \times (1 - L(p)) \tag{399}$$

or

$$\text{AOQ}=p\,L(p) \ . \tag{400}$$

In general, there will be a maximum average outgoing quality as a function of the incoming lot quality p. This maximum is called the *average outgoing quality limit* (AOQL). Thus, no matter how bad the incoming quality becomes, the average outgoing quality will never be worse than the AOQL.

2. Control Charts

While studying process data in the 1920's, Dr. Walter Shewhart of Bell Laboratories first developed the concept of a control chart. Control charts can be divided into control charts for measurements (variables) and for attributes. Measurements are usually continuous random variables of a product characteristic such as temperature, weight, length, width, etc. Attributes are simply judgments as to whether a part is good or bad.

All control charts have the following two primary functions:

- To determine whether the manufacturing process is operating in a state of *statistical control* in which statistical variations are strictly random.
- To detect the presence of serious deviations from the intrinsic statistical fluctuations, called *assignable variables* (*causes*), so that corrective action can be initiated to bring the process back in control.

Control charts are defined by the *central line*, which designates the expected quality of the process, and the *upper and lower control limits* whose exceedance indicates that the process is out of statistical control. However, even when the sample point falls within the control limits a trained operator is constantly monitoring the process for unnatural patterns such as trends, cycles, stratification, etc. This aspect is called control chart pattern analysis.

The following lists the three types of control charts for measurements:

$\bar{X}$-Chart:

Given: μ, σ Central Line: μ

UCL: $\mu + A\,\sigma$ where A=3/$\sqrt{n}$

LCL: $\mu - A\,\sigma$

Given: $\bar{\bar{x}}$, s Central Line: $\bar{\bar{x}}$

UCL: $\bar{\bar{x}} + A_1$ s

$$\text{with } \bar{\bar{x}} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i \tag{401}$$

LCL: $\bar{\bar{x}} - A_1$ s

In on-line quality control, the biased sample standard *s* is used:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}\ . \tag{402}$$

Typical sample sizes *n* are 4, 5, 6, or 7. These relatively small sample sizes often come from the selection of subgroups. The control factor A_1 is obtained from the expected value of the standard deviation *s*, which is $\mu_s = c_2\,\sigma$ where:

$$c_2 = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sqrt{\frac{2}{n}}, \text{ where } \Gamma(x) \text{ is the gamma function.} \tag{403}$$

Therefore, the control factor is $A_1 = A/c_2 = 3/\left(c_2\sqrt{n}\right)$.

This and subsequent control factors are calculated under the assumption that the measurements come from a normal population.

Given: $\bar{\bar{x}}$, $\bar{R}$ Central Line: $\bar{\bar{x}}$

UCL: $\bar{\bar{x}} + A_2\ \bar{R}$ with $\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R_i$. (404)

LCL: $\bar{\bar{x}} - A_2\ \bar{R}$

The range R is widely used in quality control because it can be easily obtained if the sample size is small. However, since R depends only on two observations, namely the maximum and the minimum value of the sample, it contains less information than the standard deviation. This loss of information is acceptable for a small sample size. A practical rule is to use the standard deviation s when the sample size is larger than 12.

In general, the probability distribution of the range R cannot be expressed in simple explicit form. For the normal case the first four moments of the distribution of the range R have been extensively tabulated as a function of the sample size n. For the control charts we need only the mean and standard deviation of the range R. They are denoted by $\mu_R=d_2\,\sigma$ and $\sigma_R=d_3\,\sigma$. The above control factor is then given by $A_2=A/d_2$.

R-Chart: ($n < 12$)

Given: σ Central Line: $\mu_R=d_2\,\sigma$ (405)

UCL: $D_2\,\sigma$

LCL: $D_1\,\sigma$

where $D_2=d_2+3d_3$ and $D_1=d_2-3d_3$. (406)

Given: $\overline{R}$ Central Line: $\overline{R}$

UCL: $D_4\,\overline{R}$

LCL: $D_3\,\overline{R}$

where $D_4=D_2/d_2=(d_2+3d_3)/d_2$ and $D_3=D_1/d_2=(d_2-3d_3)/d_2$. (407)

s-Chart: ($n > 12$)

Given: σ Central Line: $\mu_s=c_2\,\sigma$ (411) (408)

UCL: $B_2\,\sigma$

LCL: $B_1\,\sigma$.

The control factors for this chart are obtained using the mean and standard deviation of the sample distribution of s which are $\mu_s=c_2\,\sigma$ and $\sigma_s=c_3\,\sigma$ with:

$$c_3=\sigma\sqrt{\frac{n-1}{n}-c_2^2}\ . \tag{409}$$

We have, then, $B_2=c_2+3c_3$ and $B_1=c_2-3c_3$. (410)

Given: $\bar{s}$ Central Line: $\bar{s}$

UCL: $B_4\ \bar{s}$

LCL: $B_3\ \bar{s}$

where $B_4=B_2/c_2=(c_2+3c_3)/c_2$ and $B_3=B_1/c_2=(c_2-3c_3)/c_2$. (411)

J. Reliability and Life Testing

In *life testing*, the random variable under consideration is the time-to-failure of a component. In *fatigue studies*, the time-to-failure is measured in the number of cycles to failure and is, therefore, a discrete random variable.

The probability density $f(t)$ associated with the time T to failure is called the *failure-time density* or *life distribution*. The probability that the component will fail in the time interval $(0, t)$ is given by the cumulative distribution:

$$P(T\le t)=F(t)=\int_0^t f(x)\,dx \tag{412}$$

which, in this context, is called the *unreliability function*. The complementary cumulative distribution defines the probability that the component functions longer than time t or survives at least to time t. It is given by:

$$P(T\ge t)=R(t)=1-F(t)\ . \tag{413}$$

This is called the *reliability function* by engineers and *survivorship function* by actuaries.

NOTE: From a mathematical viewpoint, reliability is by definition a probability, therefore a number between 0 and 1.

By its very nature reliability theory has its focus in aerospace engineering on the tail area of a distribution, which defines low risk or high reliability systems. Thus the difference among different reliability models becomes significant only in the tails of the distribution where actual observations are sparse because of limited experimental data. In order to be able to discriminate between competing life distribution models, reliability engineers resort to a concept which attempts to differentiate among distributions because of physical failure mechanisms, experience, and/or intuition. Such a concept is the *failure rate function* which originated in actuary theory and mortality statistics where it is called the *force of mortality*. It is also indispensable in the analysis of extreme values where it is known as the *intensity function*.

To define this function, we first determine the conditional probability that a component of "age" t will fail in the subsequent interval $(t, t+\Delta t)$ by defining the following events:

A=lifetime $t \le T \le t+\Delta t$ (414)

B=lifetime $T \ge t$. (415)

Notice that the event B is the condition that no failure has occurred up to time t; i.e., the component has age t. In terms of these events the conditional probability is, therefore:

$$P[A\,|\,B]=\frac{P[A\cap B]}{P[B]}=P[t\le T\le t+\Delta t\,|\,T\ge t]=\frac{P[(t\le T\le t+\Delta t)\cap(T\ge t)]}{P[T\ge t]}\equiv\Delta G(t)\,. \qquad (416)$$

The numerator in the above equation is the probability that T lies between t and $t+\Delta t$ and the denominator is the reliability function $R(t)$. The equation can, therefore, be rewritten as:

$$\Delta G(t)=\frac{F(t+\Delta t)-F(t)}{R(t)}\,. \qquad (417)$$

The failure rate is defined as the probability of failure in the interval *per unit time* given that no failure has occurred before time t. Accordingly, we divide the above equation by Δt and then take the limit $\Delta t\to 0$ to arrive at the *failure rate function*:

$$Z(t)=F'(t)/R(t)=f(t)/R(t)=f(t)/\{1-F(t)\}\,. \qquad (418)$$

This function is also called *hazard function*, *hazard rate*, or simply *hazard*. It is, in a sense, the propensity of failure as a function of age. The failure rate function contains the same information as the failure time distribution but it is better suited to formulate reliability models for given sets of experimental data.

The difference between the failure time density $f(t)$ and the failure rate function $z(t)$ can be illustrated by comparing it to human mortality. There $f(t)\,dt$ is the probability that a newborn person will die between age t and $t+\Delta t$, whereas $z(t)\,dt$ is the probability that a person of age t will die in the subsequent time interval Δt.

Some statisticians have been investigating the reciprocal 1/z(t) of the failure rate function, which is called Mill's Ratio. It has no application in reliability studies.

Two important relationships exist between the failure rate function and the reliability function. They can be obtained by the following steps. First using $F(t)=1-R(t)$ we obtain by differentiation that $F'(t)=-R'(t)$.

Therefore:

$$z(t)=\frac{R'(t)}{R(t)}=-\frac{d[\ell n\,R(t)]}{dt}\,. \qquad (419)$$

Solving this differential equation for $R(t)$ and subsequently using the relationship $f(t)=z(t)\,R(t)$ yields:

$$R(t)=e^{-\int_0^t z(x)\,dx} \quad\text{and}\quad f(t)=z(t)\,e^{-\int_0^t z(x)\,dx}\,. \qquad (420)$$

The property that $R(\infty)=0$ requires that the area under the failure rate curve be *infinite*. This is in distinction to the failure time density for which the area under its curve is one. Therefore the failure rate function is not a probability density. Whereas $f(t)$ is the time rate of change of the unconditional failure probability, $z(t)$ is the time rate of change of the conditional failure probability given that the component has survived up to time t.

EXAMPLE: The function $f(t)=c\,e^{-at}$ does not quality as a failure rate function because the area under its curve is finite. Besides this requirement, the failure rate function has also to be *nonnegative*. Therefore we have the two properties of the hazard function:

$$z(t) \geq 0 \quad \text{for all } t \tag{421}$$

and

$$\int_0^\infty z(t)dt = \infty \tag{422}$$

1. Life Testing

In life testing, a random sample of n components is selected from a lot and tested under specified operational conditions. Usually the life test is terminated before all units have failed. There are essentially two types of life tests. The first type is a test that is terminated after the first r failures have occurred ($r \leq n$) and the test time is random. Data obtained from such a test are called *failure censored* or type I censored data. The other more common type test is terminated after a predetermined test time. In this case, the number of failures is random. Such data are called *time censored* or type II censored data. Failure censoring is more commonly dealt with in the statistical literature because it is easier to treat mathematically. If the failed units are replaced by new ones, the life test is called a *replacement* test; otherwise, it is called a *nonreplacement* test. The unfailed units are variously called survivors, run-outs, removals, suspensions, or censored units.

In recent years the Weibull distribution has become one of the most popular lifetime distributions. Because it can assume a wide variety of shapes, it is quite versatile. Although there exists a quick and highly visual graphical estimation technique, the maximum likelihood method is more accurate and lends itself easily to accommodate censored data.

In the following, we present the maximum likelihood method for estimating the two parameters of the Weibull distribution and a BASIC computer program for solving the maximum likelihood equations.

Let the life data consist of n units of which R units have failed at times t_i and S units have survived up to time t_S. Assuming a Weibull distribution, the likelihood function is given by:

$$L = \prod_{i=1}^{R} \alpha\beta t_i^{\beta-1} e^{-\alpha t_i^\beta} \prod_{k=1}^{S} e^{-\alpha t_k^\beta} \; . \tag{423}$$

Taking the logarithm of the likelihood function called the *log-likelihood function* yields:

$$\ell n\, L = R\,\ell n\,\alpha + R\,\ell n\,\beta + (\beta-1)\sum_{i=1}^{R} \ell n\, t_i - \alpha \sum_{i=1}^{R} t_i^\beta - \alpha \sum_{k=1}^{S} t_k^\beta \; . \tag{424}$$

The locations of the extreme values of the logarithm of a function are identical with those of the function itself. This can be shown by using the chain rule as follows:

$$\frac{d\,\ell n\,F(x)}{dx}=\frac{d\,\ell n\,F(x)}{dF}\times\frac{d\,F(x)}{dx}=\frac{1}{F(x)}\times\frac{d\,F(x)}{dx}=0\,. \tag{425}$$

Therefore, if the function $F(x)$ remains finite in the interval in which the local extreme values are located, the derivative of the logarithm of the function is zero at the same values of x for which the derivative of the function $F(x)$ itself is zero.

The derivatives with respect to the Weibull parameters α and β are:

$$\frac{\partial\,\ell n\,L}{\partial\alpha}=\frac{R}{\alpha}-\left(\sum_{i=1}^{R}t_i^{\beta}+\sum_{K=1}^{S}t_k^{\beta}\right)=0 \tag{426}$$

and

$$\frac{\partial\,\ell n\,L}{\partial\beta}=\frac{R}{\beta}+\sum_{i=1}^{R}\ell n\,t_i-\alpha\left(\sum_{i=1}^{R}t_i^{\beta}\,\ell n\,t_i+\sum_{k=1}^{S}t_k^{\beta}\,\ell n\,t_k\right)=0\ \ . \tag{427}$$

We can eliminate α from the first equation and insert it in the second equation to obtain a single nonlinear equation in β. Thus, we have:

$$\alpha=\frac{R}{\sum_{i=1}^{R}t_i^{\beta}+\sum_{k=1}^{S}t_k^{\beta}} \tag{428}$$

and

$$\frac{R}{\beta}+\sum_{i=1}^{R}\ell n t_i-\frac{R\left(\sum_{i=1}^{R}t_i^{\beta}\,\ell n t_i+\sum_{k=1}^{S}t_k^{\beta}\,\ell n t_k\right)}{\sum_{i=1}^{R}t_i^{\beta}+\sum_{k=1}^{S}t_k^{\beta}}=0\,. \tag{429}$$

The second equation can be solved for β by an iterative method and then be used to solve for α.

It is noteworthy that if there are no suspensions, it takes at least two failure times to be able to estimate the two Weibull parameters. However, if there are suspensions, only *one* failure time is sufficient to do this. On the other hand, one cannot obtain the parameters if there are only suspensions and no failures. That is why one often hears the statement that suspensions are a weak source of information.

The following is a BASIC program using Newton's method to calculate the maximum likelihood estimators for the Weibull parameters in an iterative manner. This method requires an initial guess for the shape parameter β of the Weibull distribution. However, the initial values do not have to be too close to the actual parameter for the algorithm to converge.

```
WEIMLE
DEFDBL A-Z
INPUT"R=",R
DIM STATIC R(R)
FOR I=1 TO R:INPUT"TR=", R(I):NEXT I

130 :INPUT "S=",S
IF S=0 GOTO 140
INPUT"TS=",TS
'DIM STATIC S(S)
'FOR I=1 TO S:INPUT"TS=", S(I):NEXT I

140 :INPUT "B=",B:J=0
150 : H=B*.0001:GOSUB 500
D=B:Y=F
B=B+H:GOSUB 500
B=D-H*Y/(F-Y):J=J+1
IF ABS((D-B)/D)>.000001 GOTO 150
BEEP:BEEP

PRINT"B=";B:PRINT"J=";J
A=R/(R1+S1):E=A^(-1/B)
PRINT"A=";A:PRINT"E="'E
END

500 : R1=0:R2=0:R3=0
FOR I=1 TO R
U=R(I)^B:V=LOG(R(I)):W=U*V
R1=R1+U:R2=R2+V:R3=R3+W:NEXT I
S1=0:S3=0
IF S=0 GOTO 580
S1=S*TS^B:S3=S1*LOG(TS):GOTO 580
FOR I=1 TO S
U=S(I)^B:V=LOG(S(I)0:W=U*V
S1=S1+U:S3=S3+W:NEXT I
580 : F=R/B+R2-R*(R3+S3)/(R1+S1)
RETURN
```

2. No-Failure Weibull Model

Sometimes when high reliability items are subjected to a time-censored test, *no failures* occur at all. This presents a real dilemma, because the traditional methods for estimating the parameters of a distribution cannot be applied. However, when it is assumed that the exponential model holds, one can determine an *approximate* lower $(1-\alpha)$ confidence limit for the mean life of the component, which is given as:

$$\mu > \frac{2T}{\chi^2_\alpha} \text{ with } \nu = 2 \text{ degrees of freedom.} \qquad (430)$$

Here, T is the fixed accumulated lifetime of the components tested and $\mu > \frac{2T}{\chi^2_\alpha}$ cuts off a right-hand tail of the χ^2 distribution with 2 degrees of freedom. This distribution is identical to the exponential distribution $f(t)=2\ e^{-t/2}$. Its critical value can be easily determined to be $\chi^2_\alpha = -2ln\ \alpha$. If there are n components, each having a lifetime t_i on test, then the accumulated lifetime is:

$$T = \sum_{i=1}^{n} t_i \ . \tag{431}$$

The lower $(1-\alpha)$ confidence interval is then given by:

$$\mu \geq \frac{\sum_{i=1}^{n} t_i}{-\ell n \alpha} \ . \tag{432}$$

The upper confidence interval is, of course, infinity.

PROBLEM: A sample of 300 units was placed on life test for 2,000 hours without a failure. Determine an approximately lower 95-percent confidence limit for its mean life.

SOLUTION:

$$\mu = \frac{(300)(2{,}000)}{-\ell n(1-0.95)} = \frac{600{,}000}{2.995732274} = 200{,}284 \text{ hours.} \tag{433}$$

It is not known whether the approximation of the lower bound is conservative in the sense that the actual confidence is higher than the calculated one. In fact, some practicing statisticians advise against using this approximation at all. Admittedly, it is an act of desperation, but it appears to give reasonable results.

The above method can be extended to the Weibull distribution by making use of the following relationship: If the failure time T has a Weibull distribution with shape parameter β and characteristic life η, then the random variable $Y=T^\beta$ has an exponential distribution with mean $\mu=\eta^\beta$. To show this, we apply the Jacobian method, according to which we have:

$$g(y) = f(t)\left|\frac{dt}{dy}\right| = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}\left|\frac{1}{\beta t^{\beta-1}}\right| . \tag{434}$$

Expressing the above equation in terms of the new variable y, we obtain:

$$g(y) = \alpha e^{-\alpha y} \text{ with } \mu_y = \frac{1}{\alpha} = \eta^\beta \ . \tag{435}$$

Thus, if no failures have occurred in a time-censored test and the n components have lifetimes $t_1, t_2 \ldots t_n$, the approximate lower confidence for μ_y is:

$$\mu_y = \eta^\beta \geq \frac{\sum_{i=1}^{n} t_i^\beta}{-\ell n \alpha} \,. \tag{436}$$

From this we obtain the lower confidence limit for the characteristic life:

$$\eta \geq \left(\frac{\sum_{i=1}^{n} t_i^\beta}{-\ell n \alpha} \right)^{\frac{1}{\beta}} \,. \tag{437}$$

Notice that the shape parameter β is still unknown and cannot be determined from the existing no-failure data set. To overcome this obstacle, it has been advocated to use historical failure data from similar components or from engineering knowledge of the failure physics under consideration to supply the missing Weibull shape parameter. This method is known as *Weibayes analysis* because it is regarded as an informal Bayesian procedure.

Sometimes the denominator of the above equation is set equal to one. This corresponds to a $(1 - e^{-1})$ or 63.2-percent confidence interval for the lower bound. It has been said that for this condition "the first failure is imminent." This statement must surely rank among the most bizarre ever made in life testing analysis.

K. Error Propagation Law

Instead of aspiring to calculate the probability distribution of an engineering system response or performance from the distribution of the component variables (life drivers), it is frequently adequate to generate some of its lower order moments. It is known, for instance, that the knowledge of the first four moments of a distribution enables one to establish a Pearson-type distribution or a generalized lambda distribution from which percentiles of the distribution can be estimated. Since these two families of distributions can accommodate a wide variety of distributions, the technique is often quite effective and the result is usually better than expected. Generating approximate system performance moments by this method is generally known as *statistical error propagation*. Sometimes it is referred to as the *Delta Method* or the *Root-Sum-Square (RSS) Method*.

In order to apply this technique, the functional relationship between the system output (performance) and system parameters (life drivers) must be explicitly known or at least be approximated. The essence of the technique consists in expanding the functional relation in a multivariate Taylor series about the design point or expected value of the system parameters and retaining only lower order terms. The goodness of the approximation depends strongly on the magnitude of the coefficient of variation of the system parameter distributions; the smaller the coefficient of variation, the better is the approximation.

The analysis is actually rather straightforward, albeit tedious, for moments higher than the second. A software package called Second Order Error Propagation (SOERP) relieves one of this drudgery. Despite its misleading acronym, it performs higher order error propagation analysis for complicated engineering systems. The following exposition is limited to calculating only the first two moments, namely the mean and variance, of a system output.

Let the performance of a system in terms of its system parameters $x_1, x_2 \ldots x_n$ be given by:

$$z = f(x_1, x_2 \ldots x_n) \,. \tag{438}$$

Let $\mu=(\mu_1, \mu_2 \ldots \mu_n)$ and $\sigma=(\sigma_1, \sigma_2 \ldots \sigma_n)$ denote the vector of the mean and standard deviation of the system parameters, respectively.

The Taylor series expansion of the function z about the point $\mu=(\mu_1, \mu_2 \ \ldots \ \mu_n)$ can be symbolically defined by:

$$z=f(x_1,x_2,\ldots x_n)=\sum_{n=0}^{\infty}\frac{1}{n!}\left[\Delta x_1\frac{\partial}{\partial x_1}+\Delta x_2\frac{\partial}{\partial x_2}+\ldots\Delta x_n\frac{\partial}{\partial x_n}\right]^n_{\mu_1,\mu_2\ldots\mu_n} f(x_1,x_2\ldots x_n) \quad (439)$$

where $\Delta x_i=(x_i-\mu_i)$ and the partial derivatives are evaluated at the design point $\mu=(\mu_1, \mu_2, \ ...,\mu_n)$. Retaining only terms up to second order we have:

$$z=f(\mu_1,\mu_2\ldots\mu_n)+\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_1}\right)_{\mu}(x_i-\mu_i)+\frac{1}{2!}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2 f}{\partial x_i\partial x_j}\right)_{\mu}(x_i-\mu_i)(x_j-\mu_j)\,. \quad (440)$$

1. The Mean of the System Performance

To calculate the mean of the system performance we take the expectation of the above equation and observing that $E((X_i-\mu_i))=0$, we obtain:

$$E(z)=\mu_z=f(\mu)+\frac{1}{2!}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2 f}{\partial x_i\partial x_j}\right)_{\mu}E[(x_i-\mu_i)(x_j-\mu_j)]\,. \quad (441)$$

Deleting the second-order term, we get the first-order approximation for the mean system performance:

$$E(z)=\mu_z=f(\mu)\ . \quad (442)$$

2. The Variance of the System Performance

The calculation of the variance is more complex than the one of the mean. In fact, the calculations become substantially more lengthy with increasing order of the moments. To begin with, we calculate the variance for the case of two system parameters x and y. This is not a real loss in generality. Once the result for this case has been obtained, the extension to more than two variables can be rather easily recognized. Let the system performance function be given by:

$$z=f(x, y)\ . \quad (443)$$

Furthermore, let us denote the variance of a random variable X by $V(X)=E(X-E(X))^2=E(X-\mu_X)^2=\sigma^2$ and the partial derivative of a function with respect to a variable, say x, by f_x.

(a) First-Order Taylor Series Approximation. Here we take only the first terms of the above Taylor series. The first-order approximation of the variance of the system performance is then obtained by:

$$V(z)=E(z-\mu_z)^2=E\left[\, fx(x-\mu_x)+fy(y-\mu_y)\right]^2$$
$$=f_x^2\,\sigma_x^2+f_y^2\,\sigma_y^2+2\,f_x\,f_y\,\sigma_{xy} \tag{444}$$

where σ_{xy} is the covariance of x and y. The partial derivatives f_x and f_y are sometimes called *sensitivity* coefficients.

Recalling that the covariance is related to the correlation coefficient by $\sigma_{xy}=\rho\,\sigma_x\,\sigma_y$, we obtain the *standard deviation* of the system performance by taking the square root of the above equation:

$$\sigma_z=\sqrt{f_x^2\,\sigma_x^2+f_y^2\,\sigma_y^2+2\,f_x\,f_y\,\rho\sigma_x\sigma_y}\ . \tag{445}$$

Extension of this expression to more than two system parameters is straightforward. An important, because often encountered, situation arises if the system parameters are *independent* (ρ=0). For this special case the standard deviation of the system performance simplifies to:

$$\sigma_z=\sqrt{f_x^2\,\sigma_x^2+f_y^2\,\sigma_y^2}\ . \tag{446}$$

This formula is the reason why this method is also called *Root-Sum-Square Method*. If correlation is suspected but unknown, one can use the upper limit of the standard deviation which is given by:

$$\sigma_z<|f_x|\,\sigma_x+|f_y|\,\sigma_y\ . \tag{447}$$

This corresponds to a "worst-on-worst" type situation.

Before proceeding to the second-order approximation, we examine some special cases of practical importance.

Multi-Output Linear System $\underline{y}=A\,\underline{x}$. Let the system performance function be a vector function of the following form:

$$\underline{y}=\underline{f}(\underline{x})\ . \tag{448}$$

where the system performance (output) vector $\underline{y}$ is an $(m\times 1)$ vector and the system parameter vector $\underline{x}$ be an $(n\times 1)$ vector. We expand now the system performance function in a Taylor series. Without loss of generality we set the value of the performance function at the design point equal to zero and obtain the linearized system performance function as:

$$\underline{y}=A\,\underline{x} \tag{449}$$

where A is an $(m\times n)$ matrix. It is known as the Jacobian matrix or system *sensitivity* matrix and its elements are the first partial derivatives of the system performance vector function $\underline{y}=\underline{f}(\underline{x})$.

To calculate the first-order approximation of the system performance variance, we introduce the covariance matrix which is defined as:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \cdots \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 \cdots \sigma_{2n} \\ \cdots\cdots\cdots\cdots\cdots \\ \sigma_{1n} & \sigma_{2n} \cdots \sigma_n^2 \end{bmatrix} \tag{450}$$

and is symmetric.

The covariance matrix of the system performance is then obtained by taking the following steps:

$$\begin{aligned} \text{Cov}\,(\underline{y}) &= E\,[(\underline{y}-\underline{\mu_y})\,(\underline{y}-\underline{\mu_y})^T] = E\,[\underline{y}\underline{y}^T] - E[\underline{y}]\,E[\underline{y}^T] \\ &= E\,[A\,\underline{x}\,\underline{x}^T\,A^T] - E\,[A\underline{x}]\,E\,[\underline{x}^T\,A^T]\ . \end{aligned} \tag{451}$$

Continuing from the previous page:

$$\begin{aligned} \text{Cov}\,(\underline{y}) &= A\,E\,[\underline{x}\,\underline{x}^T]\,A^T - A\,E\,[\underline{x}]\,E\,[\underline{x}^T]\,A^T \\ &= A\,[E\,(\underline{x}\,\underline{x}^T) - E\,(\underline{x})\,E\,(\underline{x}^T)]\,A^T \end{aligned} \tag{452}$$

and finally

$$\text{Cov}\,(\underline{y}) = A\,[\text{Cov}(\underline{x})]\,A^T\ . \tag{453}$$

As an example:

$$y_1 = a_1\,x_1 + a_2\,x_2 \tag{454}$$

$$y_2 = b_1\,x_1 + b_2\,x_2\ . \tag{455}$$

We assume independent system parameters; i.e., ρ=0.

$$\text{Cov}\,(\underline{y}) = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \tag{456}$$

$$= \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} a_1\sigma_1^2 & b_1\sigma_1^2 \\ a_2\sigma_2^2 & b_2\sigma_2^2 \end{bmatrix} \tag{457}$$

or

$$\text{Cov}\,(\underline{y}) = \begin{bmatrix} \left(a_1^2\sigma_1^2 + a_2^2\sigma_2^2\right) & \left(a_1 b_1\sigma_1^2 + a_2 b_2\sigma_2^2\right) \\ \left(a_1 b_1\sigma_1^2 + a_2 b_2\sigma_2^2\right) & \left(b_1^2\sigma_1^2 + b_2^2\sigma_2^2\right) \end{bmatrix}. \tag{458}$$

NOTE: Although the system parameters were assumed to be independent, the system performance output variables become dependent because of the “cross-coupling” (off-diagonal) terms of the sensitivity matrix.

Power Function $z=x^m y^n$. Let us denote the coefficient of variation as $\eta=\sigma/\mu$. The partial derivatives of the system parameters evaluated at the design point $\mu=(\mu_x, \mu_y)$ are:

$$f_x=m\,(\mu_x)^{m-1}\,(\mu_y)^n \quad \text{and} \quad f_y=n\,(\mu_x)^m\,(\mu_y)^{n-1} \ . \tag{459}$$

Assuming again independence of the system parameters, the system performance variance is then given by:

$$\sigma_z^2=\left[m(\mu_x)^{m-1}(\mu_y)^n\right]^2\sigma_x^2+\left[n(\mu_x)^m(\mu_y)^{n-1}\right]^2\sigma_y^2 \ . \tag{460}$$

Dividing both sides by $\mu_z^2=\left[(\mu_x)^m(\mu_y)^n\right]^2$ we obtain an expression for the coefficient of variation for the system performance which is:

$$\eta_z=\sqrt{m^2\eta_x^2+n^2\eta_y^2} \ . \tag{461}$$

We see that the coefficient of variation of the system performance can be expressed in terms of the coefficient of variation of the system parameters and that it is magnified by the power of the system parameters.

For example, the volume of a sphere is $V=4/3\pi R^3$ where R is its radius. Therefore a 10-percent error in the radius will show up as a 30-percent error in the volume of the sphere.

(b) Second-Order Taylor Series Approximation. Let us first consider the case of only one system parameter. The Taylor series expansion about the mean μ_x is:

$$z=f(\mu_x)+f_x\,(x-\mu_x)+\tfrac{1}{2}\,f_{xx}\,(x-\mu_x)^2 \ . \tag{462}$$

The variance is best calculated by using the relationships

$$V(a+x)=V(x) \qquad a=\text{constant} \tag{463}$$

and

$$V(x)=E\,(x^2)-(E\,(x))^2 \ . \tag{464}$$

The mean is simply given by the expectation of the Taylor series, which is:

$$\mu_z=f(\mu_x)+\tfrac{1}{2}\,f_{xx}\sigma_x^2 \ . \tag{465}$$

Therefore, the variance of the system performance is

$$\begin{aligned} V(z)&=E\left[f_x(x-\mu_x)+\tfrac{1}{2}f_{xx}(x-\mu_x)^2\right]^2-E^2\left[f_x(x-\mu_x)+\tfrac{1}{2}f_{xx}(x-\mu_x)^2\right] \\ &=f_x^2\,E(x-\mu_x)^2+\tfrac{1}{4}f_{xx}^2\,E(x-\mu_x)^4+f_x f_{xx}\,E(x-\mu_x)^3-\tfrac{1}{4}f_{xx}^2\,E^2(x-\mu_x)^2 \end{aligned} \tag{466}$$

or

$$\sigma_z = f_x^2 \sigma_x^2 + \tfrac{1}{4} f_{xx}^2 \mu_4 + f_x f_{xx} \mu_3 - \tfrac{1}{4} f_{xx}^2 \sigma_x^4 \tag{467}$$

where μ_3=skewness and μ_4=kurtosis of the system parameter.

For example: $z=x^2$.

We assume the system parameter to be normally distributed with $\mu_x=0$ and $\sigma_x^2=1$. We have the derivatives $f_x=2x$, $f_{xx}=2$ and the moments $\mu_x=0$, $\sigma_x^2=1$, $\mu_3=0$, and $\mu_4=3$. From this we obtain the mean and variance of the system performance as

$$\mu_z=1 \text{ and } \sigma_x^2=2 \quad . \tag{468}$$

It can be shown that these moments agree with the exact moments of the variable z, which in this case has a χ^2 distribution with 1 degree of freedom.

The case of several parameters becomes more complicated and will not be further pursued. For uncorrelated system parameters, the resulting expressions for the system performance moments, retaining only third-order terms in the final result, yield:

Variance:
$$\sigma_z^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma^2(x_i) + \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right) \left(\frac{\partial^2 f}{\partial x_i^2} \right) \mu_3(x_i) \tag{469}$$

Skewness:
$$\mu_3(z) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^3 \mu_3(x_i) \tag{470}$$

Kurtosis:
$$\mu_4(z) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^4 \mu_4(x_i) + 6 \sum_{i} \sum_{\substack{j \\ i>j}} \left(\frac{\partial f}{\partial x_i} \right)^2 \left(\frac{\partial f}{\partial x_j} \right)^2 \sigma^2(x_i) \sigma^2(x_j) \quad . \tag{471}$$

All derivatives are again evaluated at the design point of the system.

BIBLIOGRAPHY

Abramowitz, M., and Stegun, A., editors: *Handbook of Mathematical Functions*. Dover, 1964.

D'Agostino, R.B., and Stephens, M.A., editors: *Goodness-of-Fit Methods*. Marcel Dekker, New York, NY, 1986.

Freund, J.E., and Walpole, R.E.: *Mathematical Statistics*. Third Edition, Prentice-Hall, Englewood Cliffs, NJ, 1980.

Gibra, Isaac N.: *Probability and Statistical Inference for Scientists and Engineers*. Prentice-Hall, Inc., 1973.

Guttman, I.: *Statistical Tolerance Regions: Classical and Bayesian*. Hafner, Darien, CN, 1970.

Hahn, G.J., and Shapiro, S.S.: *Statistical Models in Engineering*. John Wiley & Sons, New York, NY, 1967.

Hall, M.: *Combinatorial Analysis*. Blaisdell Publishing Co., 1967.

Hampel, F.R.; Ronchetti, E.M.; Rosseeuw, P.J.; and Stahel, W.A.: *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, Inc., New York, NY, 1986.

Hogg, R.V., and Craig, A.T.: *Introduction to Mathematical Statistics*. MacMillan Publishing Company, New York, NY, 1978.

Johnson, N.L., and Leone, F.C. *Statistics and Experimental Design*. John Wiley & Sons, 1964.

Kapur, K.C., and Lamberson, L.R.: *Reliability in Engineering Design*. John Wiley & Sons, New York, NY, 1977.

Kendall, M.G., and Stuart, A.: *The Advanced Theory of Statistics*. Vol. 1, 4th ed., vol. 2, 4th ed., vol. 3, 4th ed., MacMillan Publishing Company, New York, NY, 1977, 1979, and 1983.

Kennedy, W.J., Jr., and Gentle, J.E.: *Statistical Computing*. Marcel Dekker, New York, NY, 1980.

Lawless, J.F.: *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York, NY, 1981.

Meyer, Stuart L.: *Data Analysis*. John Wiley & Sons, New York, NY, 1975.

MIL–HDBK–5F, vol. 2, 1 November 1990.

Miller, I.R., Freund, J.E., and Johnson, R.: *Probability and Statistics for Engineers*. Fourth Edition, Prentice-Hall, Englewood Cliffs, NJ, 1990.

Mosteller, F., and Tukey, J.W.: *Data Analysis and Regression*. Addison-Wesley Publishing Co., Reading, MA, 1977.

Nelson, W.: *Applied Life Data Analysis*. John Wiley & Sons, New York, NY, 1982.

Scarne, John: *Scarne's New Complete Guide to Gambling*. Simon and Schuster, 1986.

Shooman, M.L.: *Probabilistic Reliability: An Engineering Approach*. Second Edition, Robert E. Krieger Publishing Company, Malabar, FA, 1990.

Spiegel, M.R.: *Theory and Problems of Statistics*. Second edition, Schaum's Outline Series, McGraw-Hill, New York, NY, 1988.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1998	Technical Publication

4. TITLE AND SUBTITLE
Probability and Statistics in Aerospace Engineering

5. FUNDING NUMBERS

6. AUTHORS
M.H. Rheinfurth and L.W. Howell

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

8. PERFORMING ORGANIZATION REPORT NUMBER
M–856

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546–0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA/TP—1998–207194

11. SUPPLEMENTARY NOTES
Prepared by Systems Analysis and Integration Laboratory,
Science and Engineering Directorate

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified–Unlimited
Subject Category 65
Standard Distribution

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This monograph was prepared to give the practicing engineer a clear understanding of probability and statistics with special consideration to problems frequently encountered in aerospace engineering. It is conceived to be both a desktop reference and a refresher for aerospace engineers in government and industry. It could also be used as a supplement to standard texts for in-house training courses on the subject.

14. SUBJECT TERMS
statistics, probability, Bayes' Rule, engineering

15. NUMBER OF PAGES
134

16. PRICE CODE
A07

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited