UNIT FIVE: ELEMENTARY PROBABILITY

Objectives:

Having studied this unit, you should be able to

- ✓ understand the elements of probability
- ✓ calculate some probabilities of events associated with random experiments
- ✓ apply the concept of probability in some biological phenomena

5.1 Introduction

Without some formalism of probability theory, the student cannot appreciate the true interpretation from data analysis through modern statistical methods. It is quite natural to study probability prior to studying statistical inference. Elements of probability allow us to quantify the strength or "confidence" in our conclusions. In this sense, concepts in probability form a major component that supplements statistical methods and helps us to gauge the strength of the statistical inference. The discipline of probability, then, provides the transition between descriptive statistics and inferential methods. Elements of probability allow the conclusion to be put into the language that the science or engineering practitioners require. An example follows that will enable the reader to understand the notion of a *P*-value, which often provides the "bottom line" in the interpretation of results from the use of statistical methods.

5.2 Definition of some probability terms

Definition 5.1: **Random experiment** is an experiment in which the outcome cannot be determined or predicted exactly in advance, i.e. it is the process of observing or measuring the outcome of a chance event.

Some of the characteristics of a random experiment are

- ✓ All the possible outcomes of the experiment can be specified in advance.
- ✓ The experiment can be repeated indefinitely.
- ✓ There is a sort of regularity in the outcomes observed in large repetitions of the experiment.

Examples of random experiments includes throwing a fair coin and observing the outcome, throwing a fair die and observing the number on the top face, taking a student at random from science class and noting the sex of the student.

All of these examples satisfy the above characteristics of a random experiment.

Definition 5.2:

Sample point (outcome): The individual result of a random experiment.

Sample space: The set containing all possible sample points (out comes) of

the random experiment. The sample space is often called the *universe* and denoted by S.
Event: The collection of outcomes or simply a subset of the sample space. We denote events with capital letters, A, B, C, etc.

Example 5.1: If an experiment consists of flipping of a coin once, then $S = \{H, T\}$ where H means that the outcome of the toss is a head and T that it is a tail. A= {H} represents the event of head occurring.
Example 5.2: If an experiment consists of rolling a die once and observing the number on top, then the sample space is $S = \{1, 2, 3, 4, 5, 6\}$ where the outcome $i$means that $i$appeared on the die, i= 1, 2, 3, 4, 5, 6. {1}, {2},{3}, {4},{5} and {6}are elementary events i.e. events consisting of a single outcome. Let A represents the event of an odd number will occur, then A is simply the set containing 1, 3 and 5 i.e. A= {1, 3, 5}.

Review of set theory

Concepts of set theory are important in understanding probability. Given A,B and C are events associated with a sample space S and ω represents an elementary event (outcome) in S, then the following are some useful definitions and results in set theory.

Definitions 5.3:
1. Union: The union of A and B, A u B, is the event containing all sample points in either
A or B or both. Sometimes we use A or B for union.
2. Intersection: The intersection of A and B, A n B, is the event containing all sample points that are both in A and B. Sometimes we use AB or A and B for intersection.
3. Subset: If for any ω ∈ A, then ω ∈ B. Then $A \subseteq B$.
4. Empty set: If a set A contains no points, it will be called the null set, or empty set, and denoted by φ.
5. Complement: The complement of a set A denoted by A^c is the set where ω∈ S, ω ∈ A^c but, $\omega \notin A$.
6. Mutually Exclusive Events: Two events are said to be mutually exclusive (or disjoint) if their intersection is empty. (i.e. A n B = φ). Subsets A_1, A_2,… are defined to be mutually exclusive if A_i n A_j = φ for every i ≠ j.

Theorem 5.1:Important elementary set theory results
i) Au B=B u A and A n B = B n A
ii) Au (B u C) = (Au B) u C and A n (B n C) = (A n B) n C
iii) An (B u C) = (A n B) u (A n C) and Au (B n C) = (A u B) n (A u C)
iv) $(A^c)^c$ = A

v) An S = A; A u S = S; A n$\varphi = \varphi$; and A u A =A
vi) $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

5.3 Counting rules

Combinatorics refers to the methods used to count things. If a sample space contains a finite set of outcomes, determining the probability of an event often is a counting problem. But often the numbers are just too large to count in the 1, 2, 3, 4 ordinary ways. For example, if you put a grain of rice on the first square of a chessboard, then two grains on the second square, four on the third square, and continue doubling until all 64 squares are filled, how many grains of rice would you have in all? The number is so large that it is difficult to handle without a systematic enumeration technique.

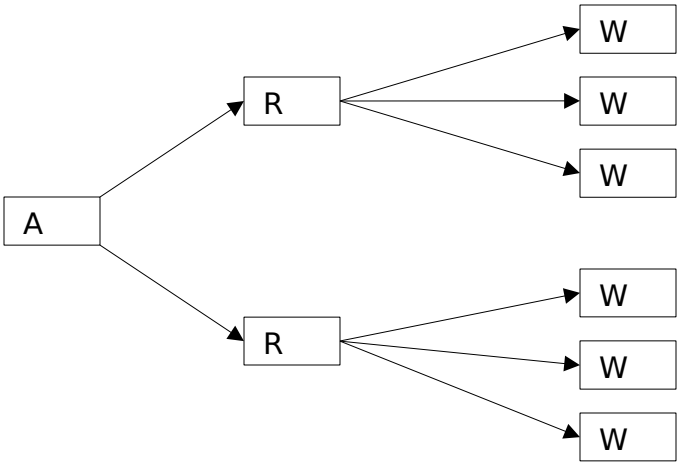
In short, to assign probabilities for an event, we might need to enumerate the possible outcomes of a random experiment and need to know the number of possible outcomes favoring the event. The following principles will help us in determining the number of possible outcomes favoring a given event.

Theorem 5.2:Addition principle
If a task can be accomplished by k distinct procedures where the i^{th}procedure has n_i alternatives, then the total number of ways of accomplishing the task equals
$n_1 + n_2 + \ldots + n_k$.

Example 5.3: Suppose one wants to purchase a certain commodity and that this commodity is on sale in 5 government owned shops, 6 public shops and 10 private shops. How many alternatives are there for the person to purchase this commodity?
Solution: Total number of ways =5+6+10=21 ways

Theorem 5.3: Multiplication principle
If a choice consists of k steps of which the first can be made in n_1 ways, for each of these the second can be made in n_2 ways… and for each of these the k^{th} can be made in n_k ways, then the whole choice can be made in $n_1.n_2.\ldots.n_k$ ways.

Example 5.4: If we can go from Addis Ababa to Rome in 2 ways and from Rome to Washington D.C. in 3 ways then the number of ways in which we can go from Addis Ababa to Rome to Washington D.C. is 2x3 ways or 6 ways. We may illustrate the situation by using a tree diagram below:



Example 5.5: If a test consists of 10 multiple choice questions, with each permitting 4 possible answers, how many ways are there in which a student gives his/her answers?
Solution: There are 10 steps required to complete the test.
First step: To give answer to question number one. He/she has 4 alternatives.
Second step: To give answer to question number two, he/she has 4 alternatives……
Last step: To give answer to last question, he/she has 4 alternatives.
Therefore, he/she has $4x4x4x...x4=4^{10}$ ways or1, 048, 576 ways of completing the exam. Note that there is only one way in which he /she can give correct answers to all questions and that there are 3^{10} ways in which all the answers will be incorrect.
Example 5.6: A manufactured item must pass through three control stations. At each station the item is inspected for a particular characteristic and marked accordingly. At the first station, three ratings are possible while at the last two stations four ratings are possible. Hence there are 48 ways in which the item may be marked.
Example 5.7: Suppose that car plate has three letters followed by three digits. How many possible car plates are there, if each plate begins with a H or an F?
2x 26x 26x 10x 10x 10 or 1, 352, 000 different plates.

Definition 5.4: If n is a positive integer, we define n!= n(n-1)(n-2)...1 and call it *n-factorial* and 0!=1.

Permutations
Suppose that we have n different objects. In how many ways, say $_nP_n$, may these objects be arranged (permuted)? For example, if we have objects a, b and c we can consider the following arrangements: abc, acb, bac, bca, cab, and cba. Thus the answer is 6. The following theorem gives general result on the number of such arrangements.

Theorem 5.4: Permutation

i) The number of permutations of n different objects is given by ${}_nP_n = n!$

ii) A permutation of n objects, arranged in groups of size r, without repetition, and order being important is:

$${}_nP_r = \frac{n!}{(n-r)!}$$

Example 5.8: Suppose that we have five letters a, b, c, d.

i) What is the number of possible arrangements of these letters taken all at a time?

ii) What is the number of possible arrangements of these letters if we use only three of the letters at a time?

Solution:

i) Using (i) of theorem 5.4, we have 4! ways of arranging the 4 letters, i.e. we have 24 possible arrangements.

ii) Using (ii) of theorem 5.4, we have ${}_4P_3$ ways of arranging 3 letters taken from the four letters, i.e. we have 24 possible arrangements.

Example 5.9: In a class with 8 boys and 8 girls

i) In how many ways can the children line up if they alternate girl-boy-girl-boy-… ?

ii) In how many ways can the children line up so that no two of the same sex are next to each other?

Solution:

i) The 8 girls can line-up in 8! ways, and likewise the 8 boys can line-up in 8! ways. For any single arrangement of the girls, all possible arrangements of the boys are possible, thus by multiplication principle we have 8!x 8! ways to arrange the children in girl-boy lines.

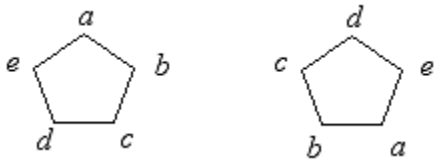
ii) Now we must include the case of boy-girl. So we have 2x8!x 8! ways of arranging.

Example 5.10: If I have 5 different books on my shelf, in how many ways can I arrange these books? Solution: We can arrange the books in 5! different ways or 5x4x3x2x1 ways or 120 ways.

Remarks

i) The number of permutations of n distinct objects arranged in a circle is (n-1)!.

This is because we consider two permutations the same if one is a rotation of the other. For n objects arranged around a circle, there a n rotations that give the same permutation. Dividing n! by n gives (n - 1)!. The two circular permutations below are considered the same; their order is a, b, c, d, e.



ii) Permutations when not all objects are different

Given n objects of which n_1 are one kind, n_2 are another kind, …,n_k of another kind, then the total number of distinct permutations that can be made from these objects is $\frac{n!}{n_1!n_2!\ldots n_k!}$.

Example 5.11

i) How many "words" (text strings or distinct arrangements) can be made from the letters b,k,o,o?

ii) How many permutations are there for the letters in the word banana?

Solution:

i) If we label the two o's as o_1 and o_2, and think of them as distinct, then the number of permutations is 4!. For each permutation there will be a matching permutation that switches the o's, that is for $o_1o_2$bk there is the matching $o_2o_1$bk permutation. We can see then that if we divide the number of distinct permutations by two, we have a count of the number of permutations of the 4 letters where we do not distinguish between the two o's. Therefore, there are distinct4!/2 text strings or 12 text strings.

ii) If we think of all 6 letters as distinct, then we would have 6! permutations. As in the preceding example for the two n's, we would need to divide 6! by 2. For the 3 a's, we would have 6 counts for a single permutation. For instance, each of the following would be a single word if the a's were not distinct. $a_1a_2a_3$bnn, $a_1a_3a_2$bnn, $a_2a_1a_3$bnn, $a_2a_3a_1$bnn, $a_3a_1a_2$bnn, and $a_3a_2a_1$bnn. Hence the number of distinct permutations of the word banana is

$\frac{6!}{2!3!}=60$.

Combinations

Consider n different objects. This time we are concerned with counting the number of ways we may choose r out of these n objects without regard to order. For example, we have the objects a, b, c and d, and r=2; we wish to count ab, ac, ad, bc, bd, and cd. In other words, we do not count ab and ba since the same objects are involved and only the order differs.

There are many problems in which we are interested in determining the number of ways in which r objects can be selected from n distinct objects without regard to the order in which they are selected. Such selections are called *combinations or r-sets*. It may help to think of combinations as committees. The key here is without regard for order.

To obtain the general result we recall the formula derived above: the number of ways of choosing r objects out of n and permuting the chosen r equals n!/(n-r)!. Let C be the number of ways of choosing r out of n,

disregarding order. C is the number required. Note that once the r items have been chosen, there are r! ways of permuting them. Hence applying the multiplication principle again, together with the above result, we obtain

C.r! = n!/(n-r)!. Therefore, $C=\frac{n!}{r!(n-r)!}$. This number arises in many contexts in mathematics and hence a special symbol is used for it. We shall write

$$\binom{n}{r}={}_nC_r=\frac{n!}{r!(n-r)!}.$$

Theorem 5.5: Combination
The number of ways of choosing r out of n different objects, disregarding order, is given by $\binom{n}{r}=\frac{n!}{r!(n-r)!}$.

Example 5.12: How many different committees of 3 can be formed from Hawa, Segenet, Nigisty and Lensa?
Solution: The question can restated in terms of subsets from a set of 4 objects, how many subsets of 3 elements are there? In terms of combinations the question becomes, what is the number of combinations of 4 distinct objects taken 3 at a time? The list of committees:{H,S,N}, {H,S,L}, {H,N,L}, {S,N,L}.Therefore, we have ${}_4C_3$ or 4 possible number of committees.
Example 5.13:
(i) A committee of 3 is to be formed from a group of 20 people. How many different committees are possible?
(ii) From a group of 5 men and 7 women, how many different committees consisting of 2 men and 3 women can be formed?
Solution: (i) There are $\binom{20}{3}=\frac{20!}{3!17!}=1140$ possible committees.
(i) $\binom{5}{2}\binom{7}{3}=\frac{5!}{2!3!}\frac{7!}{3!4!}=350$ possible committees.
Remarks:
i) $\binom{n}{r}=\binom{n}{n-r}$

ii) A set with n elements has 2^n subsets.

5.4 Probability of an event

Definition 5.5: **The Axioms of Probability**

Probabilities are real numbers assigned to events (or subsets) of a sample space. We can think of the assignment of probabilities to events, or probability measure, as a function between the collection of subsets of the sample space and the real numbers. Mathematically, a probability measure P for a random experiment is a real-valued function defined on the collection of events that satisfies the following axioms:

Axiom 1: The probability of an event is a nonnegative real number; that is, $P(A) \geq 0$ for any subset A of S.

Axoim 2: $P(S) = 1$

Axiom 3: If A_1, A_2, A_3 ... is a finite or infinite sequence of mutually exclusive events of S, then $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots = \sum P(A_i)$

It is rather surprising that with only these three axioms, we can construct the "entire" theory of probability! The next theorems and definitions help in assigning probabilities of events.

Theorem 5.6 : If *A* is an event in a discrete sample space *S*, then *P(S)* equals the sum of the probabilities of the individual outcomes comprising *A*.

Theorem 5.7: Suppose that we have a random experiment with sample space S and probability function P and A andB are events. Then we have the following results:

i) $P(\varphi) = 0$

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

iii) $P(B \cap A^c) = P(B) - P(A \cap B)$

iv) If A subset of B then $P(A) \leq P(B)$.

Definition 5.6: **The classical definition of probability**

If an experiment can result in any one of N equally likely and mutually exclusive outcomes, and if n of these outcomes constitute the event A, then the probability of event A is $P(A) = \frac{n}{N}$.

Example 5.14: Consider the experiment of tossing a fair die. A fair die means that all six numbers are equally likely to appear. Calculate the probabilities of the following events:

a) A=One will occur ={1}
b) B=Even number will occur ={2, 4, 6}
c) C=Odd number will occur ={1, 3, 5}
d) D=A number less than 3 will occur ={1,2}

Solution:

a) Since the die is fair

$P(A)=P(\{1\})=\frac{1}{6}$

$P(\{2\})=P(\{3\})=P(\{4\})=P(\{5\})=P(\{6\})=\frac{1}{6}$

$b)P(B)=\frac{3}{6}=0.5;\ \ c)P(C)=\frac{3}{6}=0.5;\ \ d)P(D)=\frac{2}{6}$

Example 5.15: Suppose that we toss two coins, and assume that each of the four outcomes in the sample space *S = {(H,H),(H, T), (T ,H), (T , T)}* are equally likely and hence has probability ¼. Let A= *{(H, H),(H, T)}* and B= *{(H,H), (T ,H)}* that is, Ais the event that the first coin falls heads, and Bis the event that the second coin falls heads. Then, calculate the probabilities of A, B, A^c, B^c, and S^c. The event that none of the outcomes will occur is the same as S^c.

Solution:

$P(A)=\frac{2}{4}=0.5$

$P(B)=\frac{2}{4}=0.5$

$P(A^c)=1-P(A)=1-0.5=0.5$

$P(B^c)=1-P(B)=1-0.5=0.5$

$P(S^c)=1-P(S)=1-1=0=P(\varphi)$

Example 5.16: From a group of 5 men and 7 women, it is required to form a committee of 5 persons. If the selection is made randomly, then

i) what is the probability that 2 men and 3 women will be in the committee?
ii) what is the probability that all members of the committee will be men?
iii) what is the probability that at least three members will be women?

Solution: The total number of possible committees is $\binom{12}{5}=\frac{12!}{5!7!}=792$, i.e. the number of possible out comes in the sample space is 792.

i) Let A be the event that the committee will consist of two 2 men and 3 women. We need to know the number of possible outcomes favoring this event. The number of ways we can select 2 men from 5 men is $\binom{5}{2}=\frac{5!}{2!3!}=10$ and the number of ways of selecting 3 women out of 7 women is $\binom{7}{3}=\frac{7!}{3!4!}=35$. Using the multiplication principle, the number of elements favoring event A is 10x35 or 350.
Hence, using the classical definition of probability,

$$P(A)=\frac{\binom{5}{2}\binom{7}{3}}{\binom{12}{5}}=\frac{350}{792}=0.44$$

ii) Let B be the event that all members of the committee will be men. Hence

$$P(A)=\frac{\binom{5}{5}\binom{7}{0}}{\binom{12}{5}}=\frac{1}{792}$$

iii) Let C be the event that at least three of the committee members will be women.

Basically, three different compositions of committee members can be formed in terms of sex: 3 women and 2 men, 4 women and 1 man, and all are women. Hence the number of possible outcomes favoring event C using the principle of combination together with the addition principle is $\binom{5}{2}\binom{7}{3}+\binom{5}{1}\binom{7}{4}+\binom{5}{0}\binom{7}{5}=350+175+21=546$.

Therefore,

$$P(C)=\frac{\binom{5}{2}\binom{7}{3}+\binom{5}{1}\binom{7}{4}+\binom{5}{0}\binom{7}{5}}{\binom{12}{5}}=\frac{546}{792}=0.69$$

Definition 5.7: **Relative Frequency Definition of probability**
If an experiment is repeated a large number, n, of times and the event A is observed $n_A$times, the probability of A is P(A) ≈ n_A/n.

The above definition of probability is based on empirical data accumulated through time or based on observations made from repeated experiments for a large number of times.

5.5 Some probability rules

Theorem 5.8: If A and B , thenP(A u B) = P(A) + P(B) − P(A n B).

Example 5.17: Consider the experiment of tossing a fair die. Let
A = Even number occurring = {2,4,6}
B = A number greater than 2 occurring ={3, 4, 5, 6}
C = Odd number occurring ={1, 3, 5}

i) What is the probability that A and B will occur?
ii) What is the probability that A or B will occur?

Solution: We use the concept of set theory to help us solve probability questions very easily and vein diagrams are useful tools to depict the relations between events within the sample space. The shaded region on Fig 1.shows the event that both A and B will occur.

i) A and B ≡ AnB ={4,6}
Thus P(AnB)=2/6.

ii) A or B ≡ AUB ={2,3,4,5,6}

AnB={4,6} Hence,

$$P(AUB) = P(A) + P(B) - P(AnB) = \frac{3}{6} + \frac{4}{6} - \frac{2}{6} = \frac{5}{6}.$$

Example 5.18: Sixty percent of the families in a certain community own their own car, thirty percent own their own home, and twenty percent own both their own car and their own home. If a family is randomly chosen,

a) what is the probability that this family do not have a car?
b) what is the probability that this family owns a car or a house?
c) what is the probability that this family owns a car or a house but not both?
d) what is the probability that this family owns only a house?
e) what is the probability that this family neither owns a car nor a house?

Solution: Let A represents that the family owns a car and B represents that the family owns a house. Given information: P(A)=0.6,P(B)=0.3, and P(AnB)=0.2.

a) Required: $P(A^c)$ = ?
$P(A^c)$=1-P(A) = 1-0.6 = 0.4

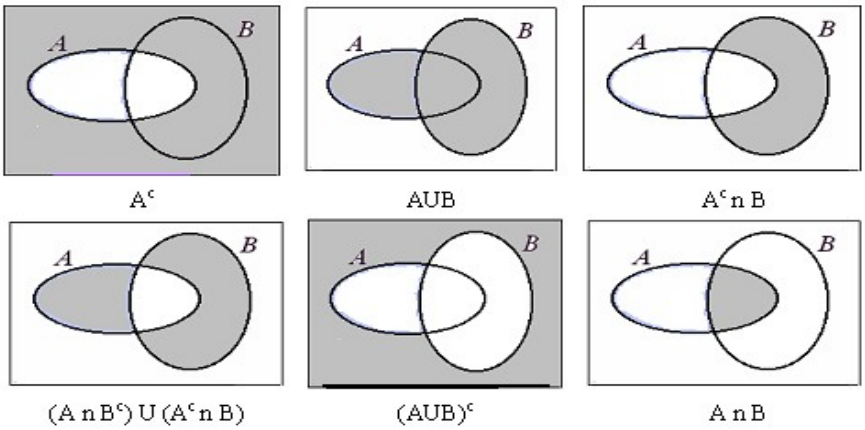
b) Required: P(AUB) = ?
P(AUB) = P(A)+P(B)-P(AnB) = 0.6+0.3-0.2 = 0.7

c) Required: $P((AnB^c)U(A^cnB))$ = ?
$P((AnB^c)U(A^cnB)) = P(AnB^c)+P(A^cnB)$ = [P(A)-P(AnB)]+[P(B)-P(AnB)]
= [0.6-0.2]+[0.3-0.2]=0.5

d) Required: $P(A^cnB)$ =?
$P(A^cnB)$ = P(B)-P(AnB) = 0.3-0.2 = 0.1

e) Required: $P(A^cnB^c)$ = ?
$P(A^cnB^c) = P((AUB)^c)$ = 1-P(AUB) = 1-0.7 = 0.3

We can represent various events by an informative diagram called vein *diagram.* If properly and correctly drawn, a vein diagram helps to calculate probabilities of events easily. The figure below shows various events represented by shaded regions. Note that the rectangle in each figure represents the sample space.



5.6 Conditional probability and independence

Conditional Probability

Conditional probability provides us with a way to reason about the outcome of an experiment, based on partial information. Here are some examples of situations we may have in our mind:

(a) What is the probability that a person will be HIV-Positive given he has tuberculosis?

(d) A spot shows up on a radar screen. How likely is it that it corresponds to an aircraft?

In more precise terms, given an experiment, a corresponding sample space, and a probability law, supposes that we know that the outcome is within some given event B. We wish to quantify the likelihood that the outcome also belongs to some other given event A. We thus seek to construct a new probability law, which takes into account this knowledge and which, for any event A, gives us the conditional probability of A given B, denoted by P($A|B$).

Definition 5.8: If P(B) > 0, the conditional probability of A given B, denoted by P(A|B), is $P(A/B)=\frac{P(AnB)}{P(B)}$.

Example 5.19: Suppose cards numbered one through ten are placed in a hat, mixed up, and then one of the cards is drawn at random. If we are told that the number on the drawn card is at least five, then what is the conditional probability that it is ten?

Solution : Let A denote the event that the number on the drawn card is ten, and Bbe the event that it is at least five. The desired probability is $P(A|B)$.

$$P(A/B)=\frac{P(AnB)}{P(B)}=\frac{P(\{10\}n\{5,6,7,8,9,10\})}{P(\{5,6,7,8,9,10\})}=\frac{P(\{10\})}{P(\{5,6,7,8,9,10\})}=\frac{1/10}{6/10}=\frac{1}{6}$$

Example 5.20: A family has two children. What is the conditional probability that both are boys given that at least one of them is a boy? Assume that the sample space S is given by $S = \{(b, b), (b, g), (g, b), (g, g)\}$,

and all outcomes are equally likely. *(b, g)* means, for instance, that the older child is a boy and the younger child is a girl.

Solution: Letting *A* denote the event that both children are boys, and *B* the event that at least one of them is a boy, then the desired probability is given by

$$P(A/B)=\frac{P(AnB)}{P(B)}=\frac{1/4}{3/4}=\frac{1}{3}$$

Law of Multiplication

The defining equation for conditional probability may also be written as:

P(AnB) = P(B) P(A|B)

This formula is useful when the information given to us in a problem is P(B) and P(A|B) and we are asked to find P(AnB). An example illustrates the use of this formula. Suppose that 5 good fuses and two defective ones have been mixed up. To find the defective fuses, we test them one-by-one, at random and without replacement. What is the probability that we are lucky and find both of the defective fuses in the first two tests?

Example 5.21: Suppose an urn contains seven black balls and five white balls. We draw two balls from the urn without replacement. Assuming that each ball in the urn is equally likely to be drawn, what is the probability that both drawn balls are black?

Solution: Let *A* and *B* denote, respectively, the events that the first and second balls drawn are black. Now, given that the first ball selected is black, there are six remaining black balls and five white balls, and so *P (B|A)* = 6/11. As *P(A)* is clearly 7/12 , our desired probability is

$$P(AnB)=P(A)P(B/A)=\frac{7}{12}.\frac{6}{11}=\frac{7}{22}$$

Independence

We have introduced the conditional probability P (A|B) to capture the partial information that event B provides about event A. An interesting and important special case arises when the occurrence of B provides no information and does not alter the probability that A has occurred, i.e., P(A|B) = P(A).When the above equality holds, we say that A is independent of B. Note that by the definition P(A|B) = P(A ∩ B)/P(B), this is equivalent to P(A ∩ B) = P(A)P(B).

Definition 5.9: **Independence**
Two events *A* and *B* are said to independent if P (*A* ∩ *B*) = P (*A*)P(*B*). If in addition, P (*B*) >0, independence is equivalent to the condition P(*A*|*B*) = P(*A*).

UNIT SIX: PROBABILITY DISTRIBUTIONS

Objectives:

Having studied this unit, you should be able to

- ✓ compute probabilities of events using the concept of probability distributions.
- ✓ compute expected values and variances of random variables.
- ✓ apply the concepts of probability distributions to real-life problems.

Introduction

In many applications, the outcomes of probabilistic experiments are numbers or have some numbers associated with them, which we can use to obtain important information, beyond what we have seen so far. We can, for instance, describe in various ways how large or small these numbers are likely to be and compute likely averages and measures of spread. For example, in 3 tosses of a coin, the number of heads obtained can range from 0 to 3, and there is one of these numbers associated with each possible outcome. Informally, the quantity "number of heads" is called a random variable, and the numbers 0 to 3 its possible values. The value of a random variable is determined by the outcome of the experiment. Thus, we may assign probabilities to the possible values of the random variable.

6.1 Definition of random variables and probability distributions

Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome. Mathematically, a random variable is a real-valued function of the experimental outcome. The following are some examples of random variables:

(a) In an experiment involving a sequence of 5 tosses of a coin, the number of heads in the sequence is a random variable.

(b) In an experiment involving two rolls of a die, the following are examples of random variables: (1) The sum of the two rolls, (2) The number of sixes in the two rolls.

(c) In an experiment involving the transmission of a message, the time needed to transmit the message, the number of symbols received in error, and the delay with which the message is received are all random variables.

Notation: We will use capital letters to denote random variables, and lower case characters to denote real numbers such as the numerical values of a random variable.

Types of random variables: Generally, two types of random variables exist: discrete and continuous. A random variable is called *discrete*if its range (the set of values that it can take) is finite or at most countably infinite. For instance, the number of children in a family, number of car accidents within given period of time in a certain locality, the number of bacteria in a cubic mm of agar, etc. If random variable assumes any numerical value in an interval or collection of intervals, then it is called a

continuous random variable. Examples include body weight of new born baby, life time of a human being, height of a person, etc.
The most important way to characterize a random variable is through the probabilities of the values that it can take. For a discrete random variable X, these are captured by the probability mass function (p.m.f. for short) of X, denoted $P_X(x)$. For a continuous random variable X it is done by the probability density function (p.d.f.), denoted $f_X(x)$.

Definition 6.1: Probability mass function
If x is any possible value of X, the probability mass of x, denoted $P_X(x)$, is the probability of the event {X = x} consisting of all outcomes that give rise to a value of X equal to x. A probability mass function must satisfy the following conditions:

i. $P_X(x) \geq 0$ for any value of x of X.
ii. $\sum P_X(x) = 1$ where the summation is over all values of x

.

Example 6.1: Consider an experiment of tossing two fair coins. Letting X denote the number of heads appearing on the top face, then X is a random variable taking on one of the values 0, 1, 2 . The random variable X assigns a 0 value for the outcome (T,T), 1 for outcomes (T ,H) and (H, T), and 2 for the outcome (H,H). Thus, we can calculate the probability that X can take specific value/s as follows:
P(X = 0) = P({(T , T)}) = ¼
P(X = 1) = P({(T ,H),(H, T)}) = 2/4,
P(X = 2) = P({(H,H)}) = ¼
The table below shows the probability mass function X.

X	0	1	2
$P_X(x)$	¼	2/4	¼

We can justify that $P_X(x)$ is probability mass function.
$P_X(x) \geq 0$ for x=0,1,2 and
P(X = 0) + P(X = 1)+P(X = 2) = ¼ + 2/4 + ¼=1
Suppose we are interested to calculate the probability that X≥1. The values of X which are greater than or equal to 1 are 1 and 2. Thus, the probability that X is greater than or equal to 1, denoted P(X≥1), is found as P(X≥1) = P(X = 1) + P(X = 2)=3/4.

Definition 6.2: Continuous random variable
A random variable X is called *continuous* if there exists a function $f_X(x)$ called the probability density function of X which satisfies

a. $f_X(x) \geq 0$ for all x.

b. $\int_{-\infty}^{\infty} f_X(x)dx=1$

We can use the probability density function to calculate probabilities of events expressed in terms of the random variable X. For instance, if we are interested in the probability that X lies between two points, say a and b, we can find it using integration of $f_X(x)$ on the interval [a,b],i.e.

$$P(a\le X\le b)=\int_a^b f_X(x)dx$$

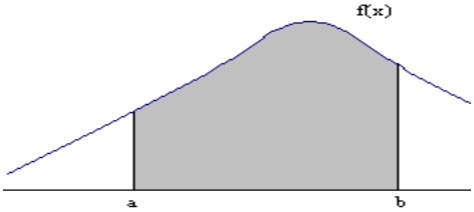


Figure: P(a≤ X ≤ b) is the shaded region

Remarks:

i) The area bounded under the graph of a probability density function and below by the horizontal axis is 1.

ii) The probability that a continuous random variable X will assume a specific value is zero, i.e. $P(X=c)=\int_c^c f_X(x)dx=0$ where c is a constant.

iii) The probability that a continuous random variable X will assume a value in a closed intervals is the same as the probability that it will assume in open interval or half open intervals, i.e. , P(a≤X≤b) = P(a<X<b) = P(a≤X<b) = P(a<X≤b), P(X≤c) = P(X<c) , P(X≥c) = P(X>c) where a, b, and c are constants.

6.2 Introduction to expectation: mean and variance

We can associate with each random variable certain "averages" of interest, such as mean and variance which give useful summary of a probability distribution.

Mean

Definition 6.3: The (mean) expected value of a random variable X denoted by E(X) or μ is given by

i) $E(X)=\sum xP_X(x)$ if X is discrete r.v.

ii) $\int_{-\infty}^{\infty} xf_X(x)dx$ if X is continuous r.v.

It is useful to view the mean of X as a "representative" value of X, which lies somewhere in the middle of its range. We can make this statement more precise, by viewing the mean as the center of gravity of the distribution.

Variance

Definition 6.4: The variance of a random variable X denoted V(X) or σ^2 is defined as $V(X)=E[(X-\mu)^2] = E(X^2) - \mu^2$.

i) if X is discrete, $V(X)=[\sum x^2 P_X(x)]-\mu^2$

ii) if X is continuous, $V(X)=[\int_{-\infty}^{\infty} x^2 f_X(x)dx]-\mu^2$

The variance provides a measure of dispersion of X around its mean. Another measure of dispersion is the standard deviation of X, which is defined as the square root of the variance and is denoted by σ.

Example 6.2: Calculate the mean and variance of the random variable X in example 7.1.

$$E(X)=\sum xP_X(x)=0\times\frac{1}{4}+1\times\frac{1}{2}+2\times\frac{1}{4}=1$$

$$E(X^2)=\sum x^2P_X(x)=0^2\times\frac{1}{4}+1^2\times\frac{1}{2}+2^2\times\frac{1}{4}=1.5$$

$$V(X)=E(X^2)-\mu^2=1.5-1^2=0.5$$

6.3 Common discrete probability distributions – binomial and Poisson

The Binomial distribution

Many real problems (experiments) have two possible outcomes, for instance, a person may be HIV-Positive or HIV-Negative, a seed may germinate or not, the sex of a new born bay may be a girl or a boy, etc. Technically, the two outcomes are called *Success and Failure.* Experiments or trials whose outcomes can be classified as either a "success" or as a "failure" are called *Bernoulli trails*.

Suppose that n independent trials, each of which results in a "success" with probability p and in a "failure" with probability 1 − p, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to have binomial distribution with parameters n and p. The probability mass function of a binomial distribution with parameters n and p is given by

$$P_X(x)=\binom{n}{x}p^x(1-p)^{n-x}, x=0,1,2,\ldots,n$$

The mean and variance of the binomial distribution are np and np(1-p), respectively. Note that the binomial distributions are used to model situations where there are just two possible outcomes, success and failure. The following conditions also have to be satisfied.

i) There must be a fixed number of trials called n
ii) The probability of success (called p) must be the same for each trial.
iii) The trials must be independent

Example 6.3: A fair coin is flipped 4 times. Let X be the number of heads appearing out of the four trials. Calculate the following probabilities:
i) 2 heads will appear
ii) No head will appear
iii) At least two heads will appear
iv) Less than two heads will appear
v) At most heads 2 will appear

Solution: We can consider that the outcomes of each trial are independent to each other. In addition the probability that a head will appear in each trial is the same. Thus, X has a binomial distribution with number of trials 4 and probability of success (the occurrence of head in a trial) is ½. The probability mass function of X is given by

$P_X(x)=\binom{n}{x}0.5^x(1-0.5)^{n-x}=\binom{n}{x}0.5^n, x=0,1,2,3,4$, Note that n = 4 and p = 1/2

i) $P(X=2)=\binom{4}{2}0.5^2(1-0.5)^{4-2}=0.3750$

ii) $P(X=0)=\binom{4}{0}0.5^0(1-0.5)^{4-0}=0.0625$

iii) $P(X\geq 2)=P(X=2)+P(X=3)+P(X=4)=0.3750+0.2500+0.0625=0.6875$

iv) $P(X<2)=P(X=0)+P(X=1)=0.0625+0.2500=0.3125$

v) $P(X\leq 2)=P(X=0)+P(X=1)+P(X=2)=0.0625+0.2500+0.3750=0.6875$

Example 6.4: Suppose that a particular trait of a person (such as eye color or left handedness) is classified on the basis of one pair of genes and suppose that d represents a dominant gene and r a recessive gene. Thus a person with ddgenes is pure dominance, one with $rr$is pure recessive, and one with $rd$is hybrid. The pure dominance and the hybrid are alike in appearance. Children receive one gene from each parent. If, with respect to a particular trait, two hybrid parents have a total of four children, what is the probability that exactly three of the four children have the outward appearance of the dominant gene?

Solution: If we assume that each child is equally likely to inherit either of two genes from each parent, the probabilities that the child of two hybrid parents will have dd, rr, or rdpairs of genes are, respectively, ¼, ¼,½. Hence, because an offspring will have the outward appearance of the dominant gene if its gene pair is either ddor rd, it follows that the number of such children ,say X, is binomially distributed with parameters n equals

4 and p equals ¾. Thus the desired probability is $P(X=3)=\binom{4}{3}0.75^3(1-0.75)^{4-3}=0.421875.$

Example 6.5: Suppose it is known that the probability of recovery for a certain disease is 0.4. If random sample of 10 people who are stricken with the disease are selected, what is the probability that:
(a) exactly 5 of them will recover?
(b) at most 9 of them will recover?

Solution: Let X be the number of persons will recover from the disease. We can assume that the selection process will not affect the probability of success (0.4) for each trial by assuming a large diseased population size. Hence, X will have a binomial distribution with number of trials equal to 10

$$P(X=k)=\binom{10}{k}0.4^k 0.6^{10-k}, k=0,1,2,\ldots 10$$

and probability of success equal 0.4.

(a) $P(X=5)=\binom{10}{5}0.4^5 0.6^{10-5}=0.200658$

(b) $P(X\leq 9)=1-P(X=10)=1-\binom{10}{10}0.4^{10}0.6^{10-10}=1-0.000105=0.9999$

The Poisson Random Variable

A random variable *X*, taking on one of the values 0, 1, 2, . . . , is said to have a Poisson distribution if its probability mass function is given by

$$P_X(x)=\frac{e^{-\lambda}\lambda^x}{x!}, x=0,1,2,3,\ldots and\ \lambda>0$$

.

λ is the parameter of this distribution. The mean and variance of the Poisson distribution are equal and their values are equal to λ. Note that poison distributions is used to model situations where the random variable X is the number of occurrences of a particular event over a given period of time (or space). Together with this , the following conditions must also be fulfilled: events are independent of each other, events occur singly, and events occur at a constant rate (in other words for a given time interval the mean number of occurrences is proportional to the length of the interval).

The poisson distribution is used as a distribution of rare events such as telephone calls made to a switch board in a given minute, number of misprints per page in a book, road accidents on a particular motor way in one day, etc. The processes that give rise to such events are called poisson *processes*.

Example 6.6: Suppose that the number of typographical errors on a single page of this lecture note has a Poisson distribution with parameter $\lambda = 1$. if we randomly select a page in this lecture note, calculate the probability that

a) no error will occur.
b) exactly three errors will occur.

c) less than 2 errors will occur.
d) there is at least one error.

Solution: Let X= Number of errors per page

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!},\lambda=1,k=0,1,2,\ldots$$

a) Required P(X≥1)=? $P(X=0)=\frac{e^{-1}1^{0}}{0!}=\frac{1}{e}=0.367879$

b) $P(X=3)=\frac{e^{-1}1^{3}}{3!}=0.061313$

c) $P(X<2)=P(X=0)+P(X=1)=0.73576$

D) $P(X\geq 1)=1-P(X=0)=1-0.367879=0.632121$

Example 6.7: If the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$, what is the probability that no accidents will occur on a randomly selected day in the future?

Solution: Let X= number of accidents per day

$$P(X=k)=\frac{e^{-3}3^{k}}{k!},k=0,1,2,\ldots$$

Required P(X= 0) = ? $P(X=0)=\frac{e^{-3}3^{0}}{0!}=e^{-3}=0.05$

Note: The Poisson random variable has a wide range of applications in a diverse number of areas. An important property of the Poisson random variable is that it may be used to approximate a binomial random variable when the binomial parameter n is large and p is small. The probability that X will be k can be approximated by substituting λ by np in the poisson distribution, i.e. $P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!},\lambda=np$.

6.4 Common continuous probability distributions

Normal distribution

The normal distribution plays an important role in statistical inference because many real-life distributions are approximately normal; many other distributions can be almost normalized by appropriate data transformations (e.g., taking the log) and as a sample size increases, the means of samples drawn from a population of any distribution will approach the normal distribution.

A continuous random variable X is said to follow normal distribution , if and only if , its probability density function (p.d.f.) is $f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}$ where x ∈(-∞,∞), μ∈ (-∞,∞) and σ ∈ (0,∞). There are infinitely many normal distributions since different values of μ and σ define different

normal distributions. For instance, when μ= 0 and σ =1 , the above density will have the following form $f_Z(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$. This particular distribution is called *the standard normal distribution* and sometimes known as Z-distribution.. The random variable corresponding to this distribution is usually denoted by Z. If X has a normal distribution with mean μ and variance σ^2, we denote it as $X \sim N(\mu, \ \sigma^2)$.

Properties of normal distribution

i) The normal distribution curve is a bell shaped, symmetrical about μ and mesokurtic. The p.d.f. attains its maximum value at x= μ.
ii) Since for x= μ divides the area under the normal curve into two equal parts, μ is the mean, the median and the mode of the distribution.
iii) The mean and variance of the normal distribution are μ, and σ^2, respectively.
iv) The total area under the curve and bounded from below by the horizontal axis is 1, i.e. $\int_{-\infty}^{\infty} f_X(x)dx=1$

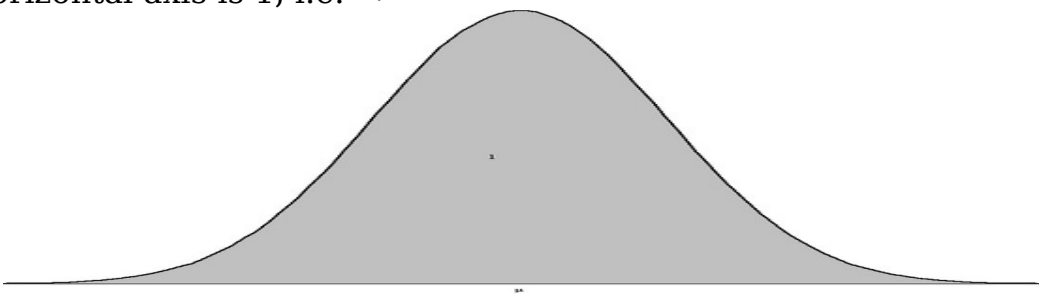


Figure: The shaded area under the normal curve is one

Since a normal distribution is a continuous probability distribution, the probability that X lies between a and b is the area bounded under the curve, from left to right by the vertical lines x = a and x = b and below by the horizontal axis.

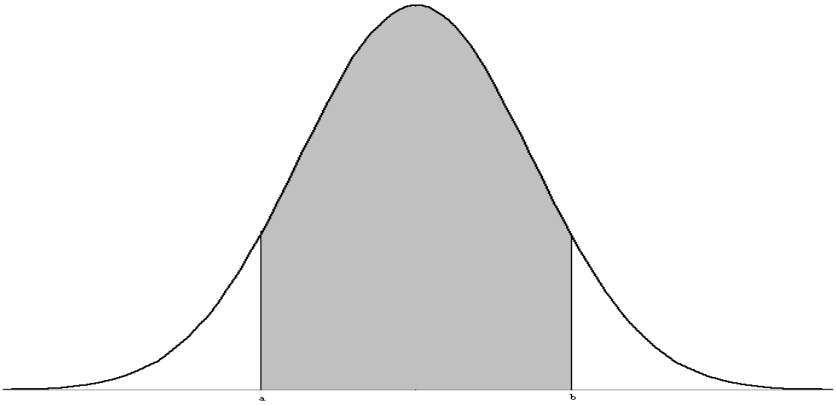


Figure: P(a<X<b) equals the shaded region

$$P(a \le X \le b) = \int_a^b f_X(x)\,dx$$

However, evaluating is very complicated. To facilitate this problem, we use the standard normal table which gives area values bounded by two points. Areas under the standard normal distribution curve are tabulated in various ways. The most common tables give areas bounded between Z=0 and a positive value of Z. In addition to the standard normal table, the properties of normal distribution and the following theorem are useful to make probability calculations very easy for any normal distribution.

Theorem 6.1: Standardization of a normal random variable
If X has a normal distribution with mean, μ and standard deviation ,σ , then

i) $Z = \dfrac{X-\mu}{\sigma}$ will have a standard normal distribution.

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

ii) $$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

Example 6.8: Let Z be the standard normal random variable. Calculate the following probabilities using the standard normal distribution table: a) P(0<Z<1.2) b) P(0<Z<1.43) c) P(Z≤0) d) P(-1.2<Z<0) e) P(Z≤-1.43) f)P(-1.43≤Z<1.2) g) P(Z≥1.52) h)P(Z≥-1.52)

Solution:

a) The probability that Z lies between 0 and 1.2 can be directly found from the standard normal table as follows: look for the value 1.2 from z column (first column) and then move horizontally until you find the value of 0.00 in the first row. The point of intersection made by the horizontal and vertical movements will give the desired area (probability). Hence P(0<Z<1.2)= 0.3849. Refer the table below as a guide to find this probability.

z	0.00	0.01	0.02	0.03	0.04	. . .	0.09
	0.0000	0.0040	0.0080	0.0120	0.0160	. . .	
0.1	0.0398	0.0438					
0.2	0.0793	0.0832					
0.3							
.							
.							
.							
1.2	0.3849	0.3869					
1.3	0.4032	0.4049					
1.4	0.4192	0.4207	0.4222	0.4236			
1.5							
.							
.							
.							
3.4	0.4997	.	.	.	.	.	. . 0.4998

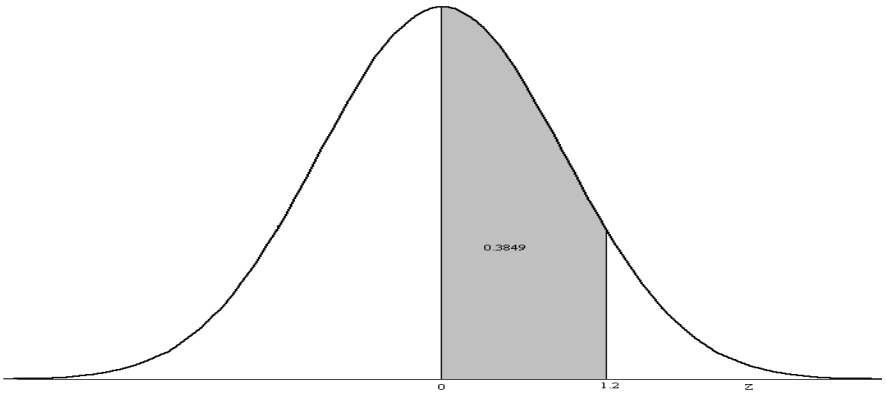


Figure: P(0<Z<1.2) is the shaded area

b) In a similar way P(0<Z<1.43)= 0.4236.

c) We know that the normal distribution is symmetric about its mean. Hence the area to the left of 0 and the to the right of zero are 0.5 each. Therefore P(Z≤0)=P(Z≥0)=0.5

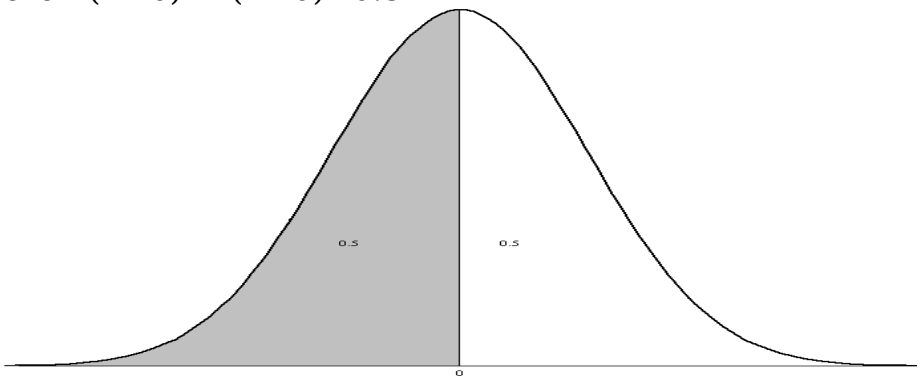


Figure: The area to the left and the right of 0 for z-distribution

d) P(-1.2<Z<0)=P(0<Z<1.2)= 0.3849 due to symmetry

e) P(Z<-1.43)= 1- P(Z ≥ -1.43) Using the probability of the complement event.

= 1-[P(-1.43<Z<0)+P(Z≥0)] Since a region can be broken down

=1-[P(0<Z<1.43)+P(Z ≥0)] into non overlapping regions.

=1-[0.4236 + 0.5]

=1-0.9236=0.0764

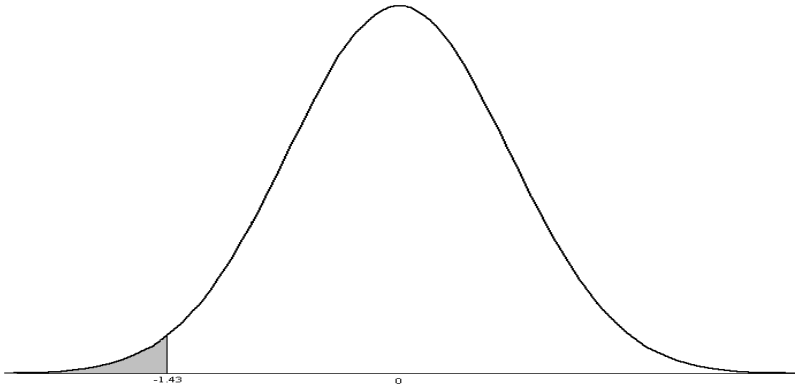


Figure: P(Z<-1.43) is the shaded region

f) P(-1.43≤Z<1.2) = P(-1.43≤Z<0) + P(0≤Z<1.2)=P(0<Z≤1.43) + 0.3849= 0.4236 + 0.3849 =0.8085

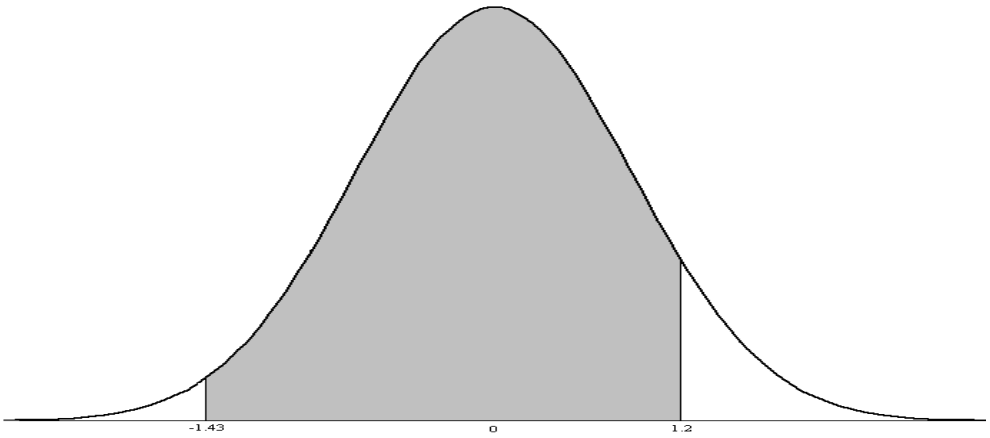


Figure: P(-1.43≤Z<1.2) is the shaded region

g) P(Z≥1.52) = 0.5 – P(0≤ Z<1.52)=0.5 – 0.4357=0.0643

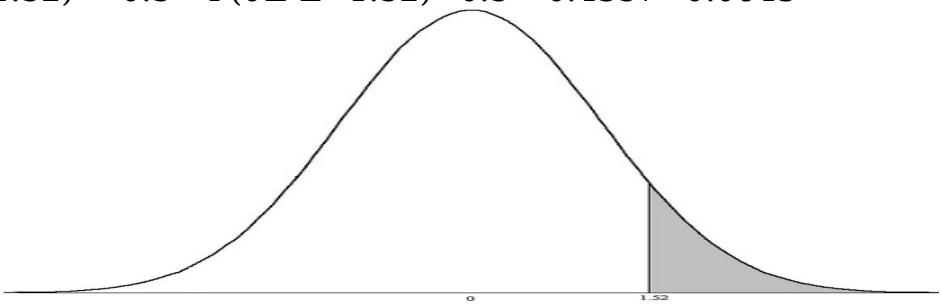


Figure: P(Z≥1.52) is the shaded region

h) P(Z≥-1.52) = P(-1.52≤Z<0) + P(Z ≥0)= P(0 < Z≤1.52) + 0.5
=0.4357 +0.5=0.9357

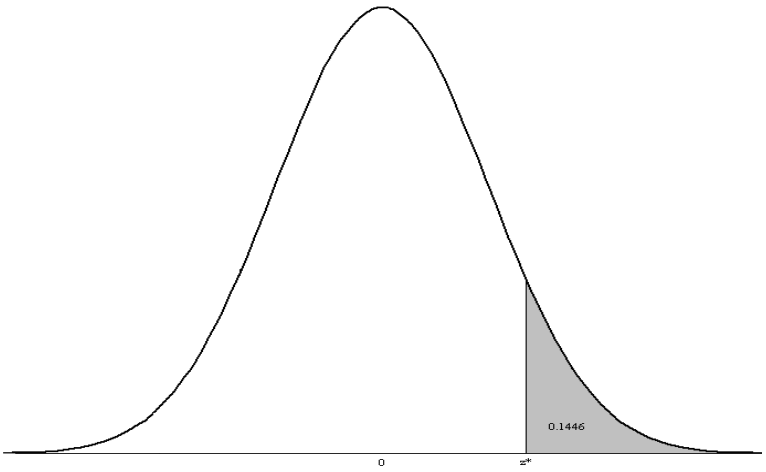
Example 7.11: Find the following values of z* of a standard normal random variable based on the given probability values:

a) P(Z > z*) =0.1446
b) P(Z>z*) = 0.8554

Solution: We need to find specific values of Z given some probability values.

a) If the probability that Z>z* is 0.1446 implies that z* is to the right of zero because

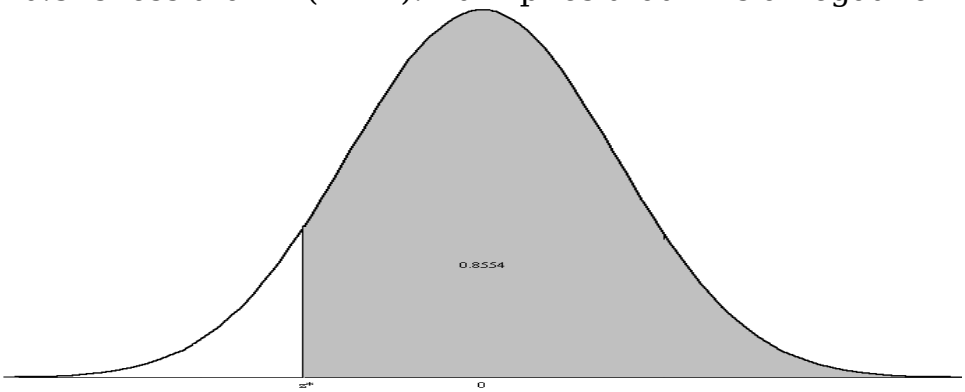
P(Z>0) = 0.5 is greater than P(Z>z*).



P(Z > z*) = 0.1446 implies that P(0<Z≤z*) = 0.5 -0.1446=0.3554. Hence we can look for the value of z* satisfying the above condition form the standard normal table. Thus z* =1.06

b) If the probability that Z>z* is 0.8554 implies that z* is to the left of zero because

P(Z>0) = 0.5 is less than P(Z>z*). It implies that z* is a negative number.



P(Z>z*) = 0.8554 = P(z*≤ Z <0) + P(Z ≥ 0) = P(0 ≤ Z ≤ - z*) + 0.5

Implies P(0 ≤ Z ≤ - z*) = 0.8554 – 0.5=0.3554. Hence the value –z* form the table satisfying the above condition is 1.06. Therefore z* = -1.06.

Example 6.9: If the total cholesterol values for a certain target population are approximately normally distributed with a mean of 200 (mg/100 ml) and a standard deviation of 20 (mg/100 ml), calculate the probability that a person picked at random from this population will have a cholesterol value

a) greater than 240 (mg/100 ml)
b) between 180 and 220(mg/100 ml)
c) less 200 (mg/100 ml)

Solution: Let X be the cholesterol values in mg/100 ml, then $X \sim N(200,\ 400)$

a) $P(X>240)=P(\frac{X-\mu}{\sigma}>\frac{b-\mu}{\sigma})$

$=P(Z>\frac{240-200}{20})=P(Z>2)=0.5-P(0<Z<2)=0.5-0.4772=0.0228$

$$P(180<X<220)=P(\frac{X-\mu}{\sigma}<\frac{X-\mu}{\sigma}<\frac{b-\mu}{\sigma})$$

$$=P(\frac{180-200}{20}<Z<\frac{220-200}{20})=P(-1<Z<1)$$

b) $=2P(0<Z<1)=2\times 0.3413=0.6826$

$$P(X<200)=P(Z<\frac{200-200}{20})=P(Z<0)=0.5$$

c)

Example 6.10: Assume that the test scores for a large class are normally distributed with a mean of 74 and a standard deviation of 10.
(a) Suppose that you receive a score of 88. What percent of the class received scores higher than yours?
(b) Suppose that the teacher wants to limit the number of A grades in the class to no more than 20%. What would be the lowest score for an A?
Solution: Let X be the score of a randomly picked student, then $X\sim N(74,\ 100)$

$$P(X>88)=P(\frac{X-74}{10}\geq\frac{88-74}{10})=P(Z>1.4)$$

a) $=0.5-P(0<Z\leq 1.4)=0.5-0.4192=0.0808$

Hence 8.08 percent of the students score more than you did?

b) Let X_A be the lowest mark to get letter grade A. We are given that

$$P(X\geq x_A)=0.2=P(\frac{X-74}{10}\geq\frac{x_A-74}{10})=P(Z>z_A)$$

$$\Rightarrow P(0<Z\leq z_A)=0.5-0.2=0.3\Rightarrow z_A=0.85\Rightarrow z_A=0.85=\frac{x_A-74}{10}$$

Hence, the lowest mark to get letter grade A is 82.5.

The chi-square and t distributions

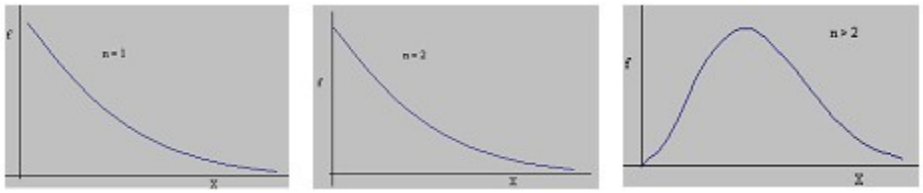
The chi-square and t distributions are important continuous distributions which are useful in statistical inference. In this section we will see a brief introduction of these distributions. In later chapters, we are going to see in detail on how to use these distributions in estimation and hypotheses testing.

Chi-square distribution

A random variable X is said to have a chi-square distribution with n degrees of freedom (denoted by χ^2_n) if its probability density function is given by

$$f_X(x)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{\frac{-x}{2}},\ x>0.$$

The chi-square distribution has one parameter called the degrees of freedom, n. Depending on the values of n, we can have many different chi-square distributions. The mean and the variance of chi-square distribution are n, and 2n, respectively.


Figure: The chi-square distribution

Because of its importance, the chi-square distribution is tabulated for various values of the parameter n (refer table). Thus we may find in the table that value, denoted by $\chi^2_\alpha(n)$, satisfying $p(X \geq \chi^2_\alpha(n)) = \alpha, 0 < \alpha < 1$. The example below helps on how to read chi-square distribution values.

Example 6.11: To read the chi-square value with 2 degrees of freedom where the area to the right of this value is 0.005.Look the degrees of freedom, 2, in the first column (df column) and then move horizontally until you find the value of α , 0.005 in the first row. The point of intersection made by the horizontal and vertical movement will give the desired chi-square value, 10.597. This value satisfies the following: $P(X \geq 10.597) = 0.005$.
In a similar way,The chi-square value with 100 degrees of freedom where the area to the right of this value is 0.975 is 74.222.

df	0.995	0.99	0.975	.	.	.	0.01	0.005
1	---	---	0.001				6.635	7.879
2	0.01	0.02	0.051				9.21	10.597
3	0.072							
4	0.207							
.								
.								
.								
100	67.328	70.065	74.222	.	.	.	135.807	140.169

The t distribution

The t distribution is an important distribution useful in inference concerning population mean/means. This distribution has one parameter called the *degrees of freedom*. Depending on the values of the degrees of freedom, we may have different t distributions. The degrees of freedom is usually denoted by n. In inference on the population mean, the degrees of freedom is related to sample size. As the sample size or degrees of freedom increases, the t distribution approaches the standard normal distribution.

The t- distribution shares some characteristics of the normal distribution and differs from it in others. The t distribution is similar to the standard normal distribution in the following ways.

i) it is bell-shaped
ii) it is symmetrical about the mean
iii) the mean, median, and mode are equal to 0 and are located at the center of the distribution.
iv) The curve never touches the x-axis

The t distribution differs from the standard normal distribution in the following ways.

i) the variance is greater than 1.
ii) The t distribution is actually a family of curves based on the concept of degrees of freedom.

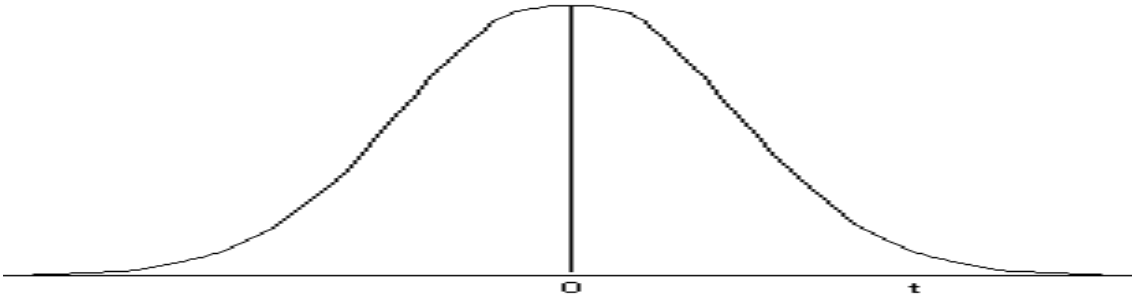


Figure: The t distribution

Due to its importance in inference values of t distribution is tabulated for some values of n (refer table). Thus we may find in the table that value, denoted by $t_\alpha(n)$, satisfying $p(t(n)\geq t_\alpha(n))=\alpha, 0<\alpha<1$ and t(n) represents the t random variable with n degrees of freedom. The following example will help you to read t distribution values.

Example 6.12: To find the t value with 3 degrees of freedom where the area to the right of this value is 0.05.Look the degrees of freedom, 3, in the first column (df column) and then move horizontally until you find the value of α , 0.05 in the first row. The point of intersection made by the horizontal and vertical movement will give the desired t value 2.353. This value satisfies the following: $P(t\geq 2.353)=0.05$

df	0.1	0.05	. .	0.0005
	t_α=1.282	1.645		
1	3.078	6.314		
2	1.886	2.92		
3	1.638	2.353		
.				
.				
.				3.66

UNIT SEVEN: SAMPLING AND SAMPLING DISTRIBUTION OF SAMPLE MEAN

Objectives:
After a successful completion of this unit, students will be able to:

- ✓ Differentiate the two major sampling techniques: probabilistic and non-probabilistic
- ✓ Apply simple random sampling technique to select sample
- ✓ Define sampling distribution of the sample mean

Introduction to sampling and sampling distribution
In our daily life we are forced to make decision based on small scale study. For instance, a laboratory technician take small droplets of blood to examine the presence disease; we examine fruits before we purchase it; zoologists use the concept of sampling to estimate the population of rodents, e t c. This process of inspection is very wide and is commonly used on various occasions. But this job is difficult to implement on large scale. On the basis of small study, we make inference about the entire population.

7.1 Methods of sampling

Definition of some basic terms
Sampling:is the technique of selecting representative sample from the whole.
Population: is the totality of elements or units under study.
Sample: is the part of the population.
Sampling Frame: A complete list of all the units of the population is called the sampling frame. A unit of population is a relative term. If all the workers in a factory make a population, then a worker is a unit of the population. If all the factories in a country are being studied for some purpose, then a factory is a unit of the population of factories. The frame provides a base for the selection of a sample.
Major reasons to use sampling

1. **Saves Time and Cost:** As the size of the sample is small as compared to the population, the time and cost involved on sample study are much less than the complete counts. Hence a sample study requires less time and cost.
2. **To prevent destruction:** The destructive nature of some experiments (or inspection) do not allow to carryout complete enumeration, for instance, to check quality of beers, to study the efficacy of new drugs, testing the life length of a bulb, e t c.
3. **Sample survey provides higher level of accuracy:** This accuracy can be achieved through more selective recruiting of interviewers and supervisors, more extensive training programs, a closer supervision of the personnel involved and a more efficientmonitoring of the field work.

Types of sampling

Generally, two types of sampling methods exist: probability and non-probability sampling.

Probability Sampling

The term probability sampling (or random sampling) is used when the selection of the sample is purely based on chance. There is no subjective bias in the selection of units. Every unit of the population has a known nonzero probability to be in the sample. The following are some of the t random sampling methods: Simple random sampling, Stratified random sampling, Cluster sampling, Systematic random sampling.

Simple random sampling

Simple random sampling is a method of selecting a sample from a population in such a way that every unit of the population is given an equal chance of being selected. In practice, you can draw a simple random sample of elements using either the 'lottery method' or 'tables of random numbers'.

For example, you may use the lottery method to draw a random sample by using a set of 'N' tickets, with numbers ' 1 to N' if there are 'N' units in the population. After shuffling the tickets thoroughly, the sample of a required size, say n, is selected by picking the required n number of tickets.

The best method of drawing a simple random sample is to use a table of random numbers. After assigning consecutive numbers to the units of population, the researcher starts at any point on the table of random numbers and reads the consecutive numbers in any direction horizontally, vertically or diagonally. If the read out numbers corresponds with the one written on a unit card, then that unit is chosen for the sample.

Suppose that a sample of 6 study centers is to be selected at random from a serially numbered population of 60 study centers. The following table is portion of a random numbers table used to select a sample.

Row¿ Colum n∀	1	2	3	4	5	……	N
1	2315	7548	5901	8372	5993	…..	6744
2	0554	5550	4310	5374	3508	…..	1343
3	1487	1603	5032	4043	6223	…..	0834

4	3897	6749	5094	0517	5853	.....	1695
5	9731	2617	1899	7553	0870	.....	0510
6	1174	2693	8144	3393	0862	.....	6850
7	4336	1288	5911	0164	5623	.....	4036
8	9380	6204	7833	2680	4491	.....	2571
9	4954	0131	8108	4298	4187	.....	9527
10	3676	8726	3337	9482	1569	.....	3880
11	.....	.....	.....	.....	.....	.....	.....
12	.....	.....	.....	.....	.....	.....	.....
13	.....	.....	.....	.....	.....	.....	.....
14	.....	.....	.....	.....	.....	.....	.....
15	.....	.....	.....	.....	.....	.....	.....
N	3914	5218	3587	4855	4888	.....	8042

If you start in the first row and first column, centers numbered 23, 05, 14, ..., will be selected. However, centers numbered above the population size (60) will not be included in the sample. In addition, if any number is repeated in the table, it may be substituted by the next number from the same column. Besides, you can start at any point in the table. If you chose column 4 and row 1, the number to start with is 83. In this way you can select first 6 numbers from this column starting with 83.

The sample, then, is as follows:

~~83 75~~

53 **33**
40 **01**
05 **26**

Hence, the study centers numbered 53, 40, 05, 33, 01 and 26 will be in the sample.

Simple random sampling ensures the best results. However, from a practical point of view, a list of all the units of a population is not possible to obtain. Even if it is possible, it may involve a very high cost which a researcher or an organization may not be able to afford. In addition, it may result an unrepresentative sample by chance.

Stratified sampling

Stratified random sampling takes into account the stratification of the main population into a number of sub-populations, each of which is homogeneous with respect to one or more characteristic(s). Having ensured this stratification, it provides for selecting randomly the required number of units from each sub-population. The selection of a sample from each subpopulation may be done using simple random sampling. It is useful in providing more accurate results than simple random sampling.

Systematic sampling

In this method, samples are selected at equal intervals from the listings of the elements. This method provides a sample as good as a simple random sample and is comparatively easier to draw a sample. For instance, to study

the average monthly expenditure of households in a city, you may randomly select every fourth households from the household listings

Cluster sampling

Cluster sampling is used when sampling frame is difficult to construct or using other sampling techniques (simple random sampling) is not feasible or costly. For instance, when the geographic distribution of units is scattered it is difficult to apply simple random sampling. It involves division of the population of elementary units into groups or clusters that serve as primary sampling units. A selection of the clusters is then made to form the sample. The precision of estimates made based on samples taken using this method is relatively low.

Non-probabilily sampling techniques

In non-probability sampling, the sample is not based on chance. It is rather determined by personal judgment. This method is cost effective; however, we cannot make objective statistical inferences. Depending on the technique used, non-probability samples are classified into quota, judgment or purposive and convenience samples.

Sampling and non-sampling errors

Sampling error is the difference between the value of a sample statistic and the value of the corresponding population parameter. On the other hand, non-sampling error is an error that occurs in the collection, recording and tabulation of data. Sampling error can be minimized by using appropriate sampling methods and/or increasing the sample size. The non-sampling error is likely to increase with increase in sample size.

7.2 Sampling distribution of the sample mean $\overline{x}$

The value of the sample mean for any sample will depend on the elements included in that sample. Consequently, the sample mean is a random variable. Therefore, like other random variable, the sample means possess a probability distribution which is more commonly called the *sampling distribution of sample mean*. In general, the probability distribution of a sample statistic is called its sampling distribution. Sampling distribution is important in statistical inference. The important characteristics of the sampling distribution of the sample mean are its mean, variance and the form of the distribution.

Example 7.1: Suppose we have a hypothetical population of size 3, consisting of three children namely: A is 3 years old, B is 6 years old and C is 9 years old. Construct sampling distribution of the sample mean of size 2 using sampling without replacement and with replacement.

Solution: The mean and variance of the population are 6 and 6, respectively.

1. If sampling is without replacement we will have ${}_3C_2 = 3$ possible samples: (A, B), (A, C) and (B, C) and their corresponding sample means are (3+6)/2 = 4.5, 6 and 7.5, respectively. Hence the

probability distribution (sampling distribution) of the sample mean is:

$\bar{x}$	4.5	6	7.5
$P(\bar{X} = \bar{x})$	1/3	1/3	1/3

$E(\bar{X}) = \sum \bar{x} P(\bar{x}) = 4.5(1/3) + 6(1/3) + 7.5(1/3) = 6$

$V(\bar{X}) = (\sum \bar{x}^2 P(\bar{x})) - \mu_{\bar{x}^2} = (6.75 + 12 + 18.75) - 36 = 1.5$

2. If sampling is with replacement we will have $N^n = 3^2 = 9$ possible samples: (A, A), (A, B), (A, C), (B, A), (B, B), (B, C), (C, A), (C, B) and (C, C). Hence the probability distribution (sampling distribution) of the sample mean is:

$\bar{x}$	3	4.5	6	7.5	9
$P(X = \bar{x})$	1/9	2/9	3/9	2/9	1/9

$E(\bar{X}) = \sum \bar{x} P(\bar{x}) = 3(1/9) + 4.5(2/9) + 6(3/9) + 7.5(2/9) + 9(1/9) = 6$

$V(\bar{X}) = (\sum \bar{x}^2 P(\bar{x})) - \mu_{\bar{x}^2} = (1 + 4.5 + 12 + 12.5 + 9) - 36 = 3$

Note:

- ✓ The mean of the sampling distribution of the sample mean is the same as the population mean irrespective of the sampling procedure.
- ✓ The variance of the sampling distribution of the sample mean is:

$$\begin{cases} \frac{\sigma^2}{n}, & \text{if sampling is with replacement} \\ \end{cases}$$

- ✓ The problem with using sample mean to make inferences about the population mean is that the sample mean will probably differ from the population mean. This error is measured by the variance of the sampling distribution of the sample mean and is known as the *standard error*. The standard error is the average amount of sampling error found because of taking a sample rather than the whole population. As sample size increases, the standard error decreases.

7.3 Central Limit Theorem

If X_1, X_2, …, X_n is a random sample from a population with mean μ and variance σ^2, then as *n* goes to infinity the distribution of the sample mean, $\overline{X}$, approximates normal distribution with mean μ and variance σ^2/n. That is, as n gets large, $\overline{X} \sim N$ (μ, σ^2/n) and its standardized form is $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Note: The central limit theorem is useful for approximating the distribution of the sample mean based on a large sample size and when the population distribution is non normal; however, if the population is normal, then the

sampling distribution of the sample mean will be normal regardless of the sample size.

Example 7.2: If the uric acid values in normal adult males are normally distributed with mean 5.7 mgs and standard deviation of 1mg. Find the probability that

a) a sample of size 4 will yield a mean less than 5
b) a sample of size 9 will yield a mean greater than 6

Solution: Let X be the amount of uric acids in normal adult males with mean 5.7 and variance 1.

a) If a sample of size 4 is taken, then $\bar{X} \sim$ N (5.7, 0.25) since the population is normally distributed.

$$P(\bar{X}<5)=P\left(Z<\frac{5-5.7}{0.5}\right)=P(Z<-1.4)$$
$$=0.5-P(0<Z<1.4)=0.0808$$

b) If a sample of size 9 is taken, then $\bar{X} \sim$ N (5.7, 1/9) since the population is normally distributed.

$$P(\bar{X}>6)=P\left(Z>\frac{6-5.7}{1/3}\right)=P(Z>0.9)$$
$$=0.5-P(0<Z<0.9)=0.1841$$

UNIT EIGHT: SIMPLE LINEAR REGRESSION AND CORRELATION

Objectives:
Having studied this unit, you should be able to:

- ✓ Formulate a simple linear regression model.
- ✓ express quantitatively the magnitude and direction of the association between two variables

Introduction
The statistical methods discussed so far are used to analyze the data involving only one variable. Often an analysis of data concerning two or more variables is needed to look for any statistical relationship or association between them. Thus, regression and correlation analysis are helpful in ascertaining the probable form of the relationship between variables and the strength of the relationship.

8.1 Simple linear regression analysis

Regression analysis is the statistical method that helps to formulate a functional relationship between two or more variables. It can be used for assessment of association, estimation and prediction. For instance one might be interested to formulate a statistical model to relate the height of fathers and their sons, blood pressure and age, fertilizer amount and yield, etc.
A simple model to relate dependent (response) variable Y and with only one predictor variable X is to consider a linear relationship.
The first step in regression analysis involving two variables is to construct a scatter plot (diagram) of the observed data. Scatter diagram is a plot of all ordered pairs (X_i, Y_i) on the coordinate plane which is helpful for determining an apparent relationship between two variables.
The simple linear regression of Y on X can be expressed with respect to the population parameters α and β as

$$Y = \alpha + \beta X + \varepsilon$$

where α = y-intercept that represents the mean value of the dependent variable Y when the independent variable X is zero; β = slope of the regression line that represents the change in the mean of Y for a unit change in the value of X; ε = error term
The population parameters α and β can be estimated from sample data using the least square technique. The estimators of α and β are usually denoted by a and b, respectively. The resulting regression line is

$$\hat{Y} = a + b X$$

and the equation is known as the fitted regression line. The estimated values of Y are denoted by $\hat{Y}$. The observed values of Y are denoted by y. The difference between the observed and the estimated values, $Y-\hat{Y}$, is known as error or residual, and is denoted by $\{\hat{\varepsilon}\}$. The residual can be positive, negative or zero.

A best fitting line is the one for which the sum of squares of the residuals, $\sum \hat{\varepsilon}^2$ has the minimum value. This is called the method of least squares. According to this method, one would select a and b such that $\sum \hat{\varepsilon}^2 = \sum (Y-\hat{Y})^2$ is minimum. The solution of this minimization problem using partial differentiation is as follows:

$$b=\frac{\sum XY-\frac{\sum X\sum Y}{n}}{\sum X^2-\frac{(\sum X)^2}{n}} = \frac{n\sum XY-\sum X\sum Y}{n\sum X^2-(\sum X)^2} \quad \text{and} \quad a=\bar{Y}-b\bar{X}$$

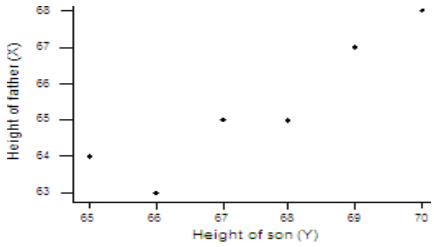
Example 8.1: A researcher wants to find out if there is any relationship between height of the son and his father. He took random sample of 6 fathers and their sons. The height in inch is given in the table below:

Height of father (X)	63	65	64	65	67	68
Height of the son (Y)	66	68	65	67	69	70

i) Draw the scatter diagram and comment on the type of relationship.
ii) Fit the regression line of Y on X.
iii) Predict the height of the son if his father's height is 66 inch.

Solution:

i)



From the scatter plot one can see that the points are roughly on straight line.

ii)

$$n=6, \sum X=392, \sum Y=405, \sum X^2=25628, \sum XY=26476, \sum Y^2=27355$$

$b=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-(\sum X)^2}=\frac{6(26476)-(392)(405)}{6(25628)-(392)^2}$ = 0.923 $a=\bar{Y}-b\bar{X}=\frac{405}{6}-0.923\frac{392}{6}$ = 7.2

Then the fitted (regression) line of Y on X is given by:

$\hat{Y}=a+b\,X$ = 7.2+0.923X

- ✓ The slope of the line, i.e. b=0.923, tells us that a unit (one inch) increase in the height of the father results in 0.923 inch increase in the height of the son.
- ✓ The y-intercept of the line, i.e. a=7.2, is the value of Y when the value of X is zero(do you think that the intercept is meaningful?)

iii) Y=7.2+0.923(66) =68.118, thus the height of the son is 68.118 inch.

8.2 The covariance and the correlation coefficient

Correlation coefficient measures the degree of linear relationship between two variables. The population correlation coefficient is represented by ρ and its estimator is r. For a set of n pairs of sample values X and Y, Pearson's correlation coefficient is calculated as the ratio of the covariance of the variables X and Y to the product of the standard deviations of X and Y. symbolically,

$$r=\frac{Cor(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}.}=\frac{\sum\frac{(X-\bar{X})(Y-\bar{Y})}{n-1}}{\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}\sqrt{\frac{\sum(Y-\bar{Y})^2}{n-1}}}$$

$$=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}$$

Alternatively, the Pearson's correlation coefficient r can be obtained as:

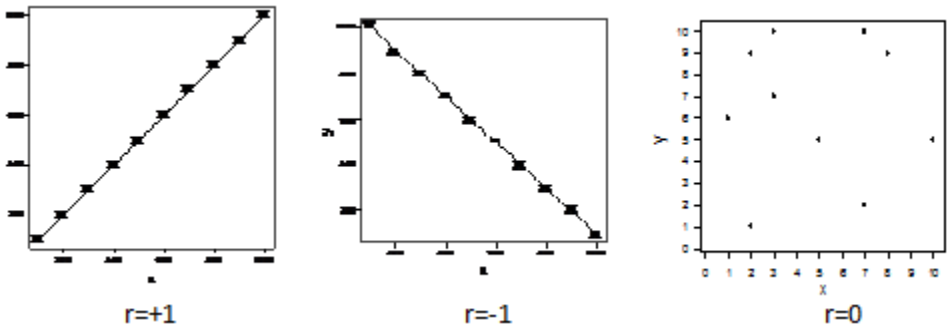
$$r=\frac{n\sum XY-(\sum X)(\sum Y)}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}}$$

Properties of Pearson's correlation coefficient r,

- o It is appropriate to calculate when both variables X and Y are measured on an interval or ratio scale.
- o The value of r is independent of the unit in which X and Y are measured. i.e., it is a pure number.
- o The value of r ranges from +1 to -1.
- o r = +1 indicates a perfect linear relationship between X and Y with positive slope.

- r = -1 indicates a perfect linear relationship between X and Y with negative slope.
- r = 0 indicates no linear relationship between the two variables X and Y.
- as r approaches +1 indicates strong and positive linear relationship between the two variables
- as r approaches -1 indicates strong and negative linear relationship between the two variables
- as r approaches 0 indicates weak linear relationship between the two variables

Examples of correlation coefficients:



Example 8.2: In some locations, there is strong association between concentrations of two different pollutants. An article reports the accompanying data on ozone concentration x (ppm) and secondary carbon concentration y ($\mu g/m^3$):

X	0.066	0.088	0.120	0.050	0.162	0.186	0.057	0.100
Y	4.6	11.6	9.5	6.3	13.8	15.4	2.5	11.8

0.112	0.055	0.154	0.074	0.111	0.140	0.071	0.110
8.0	7.0	20.6	16.6	9.2	17.9	2.8	13

a. Calculate the correlation coefficient and comment on the strength and direction of the relationship between the two variables.

Solution: The summary quantities are

$n=16, \sum x_i=1.656, \sum y_i=170.6, \sum x_i y_i=20.0397, \sum x_{i^2}=0.196912, \sum y_{i^2}=2253.56$

The Person's correlation coefficient is

$$r=\frac{n\sum XY-(\sum X)(\sum Y)}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}}$$
$$=\frac{16(20.0397)-(1.656)(170.6)}{\sqrt{16(0.196912)-(1.656)^2}\sqrt{16(2253.56)-(170.6)^2}}$$
$$=\frac{320.6352-282.5136}{\sqrt{0.408256}\sqrt{6952.6}}=\frac{38.1216}{(.639)(83.38)}$$
$$=0.716$$

The value of 0.716 indicates that there is somehow strong and positive relationship between ozone concentration and secondary carbon concentration.

UNIT NINE: ESTIMATION AND HYPOTHESIS TESTING

Objectives:

Having studied this unit, you should be able to

- ✓ construct and interpret confidence interval estimates
- ✓ formulate hypothesis about a population mean
- ✓ determine an appropriate sample size for estimation

Introduction

We now assume that we have collected, organized and summarized a random sample of data and are trying to use that sample to estimate a population parameter. Statistical inference is a procedure whereby inferences about a population are made on the basis of the results obtained from a sample. Statistical inference can be divided in to two main areas: estimation and hypothesis testing. Estimation is concerned with estimating the values of specific population parameters; hypothesis testing is concerned with testing whether the value of a population parameter is equal to some specific value.

9.1 Point and interval estimation of the mean

Point estimate: In point estimation, a single sample statistic (such as $\bar{x}, s \text{ or } \hat{p}$) is calculated from the sample to provide an estimate of the true value of the corresponding population parameters (such as $\mu, \delta \text{ or } p$). Such a single statistic is termed as point estimator, and the specific value of the statistic is termed as point estimate. For example, the sample mean $\bar{X}$ is an estimator for population mean and $\bar{X} = 10$ is an estimate, which is one of the possible values of $\bar{X}$.

Interval estimate: In most practical problems, a point estimate does not provide information about 'how close is the estimate' to the population parameter unless accompanied by a statement of possible sampling errors involved based on the sampling distribution of the statistic. Hence, an interval estimate of a population parameter is a confidence interval with a statement of confidence that the interval contains the parameter value.

An interval estimate of the population parameter θ consists of two bounds within which the parameter will be contained:

$$L \leq \theta \leq U$$

where L is the lower bound and U is the upper bound.

Case 1: When the population is normal.

- ✓ If the variance σ^2 is known, the sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and variance $\frac{\delta^2}{n}$. i.e., $\bar{X} \sim N\left(\mu, \frac{\delta^2}{n}\right)$ and $Z = \frac{\bar{X} - \mu}{\delta / \sqrt{n}} \sim N(0,1)$.

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

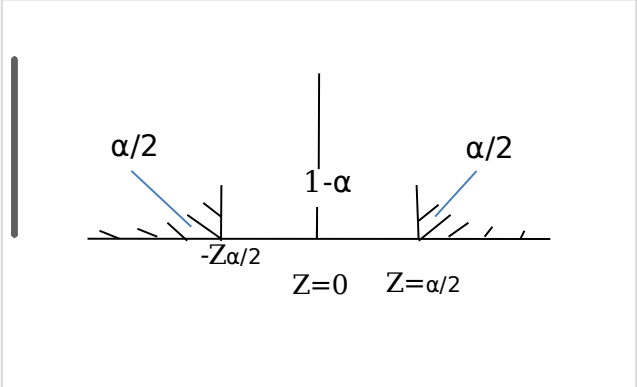
- ✓ If the variance σ^2 is unknown, will have t-distribution with n - 1 degrees of freedom. Moreover, as the sample size increases t is approximately the same as standard normal.

Consider the case σ^2 is known, we can derive a $(1-\alpha)100\%$ confidence interval for the population mean μ.

Let $Z_{\alpha/2}$ be a point on the standard normal curve that cuts an area of $\frac{\alpha}{2}$ to the right. i.e. $P(Z>Z_{\alpha/2}) = \frac{\alpha}{2}$. By the symmetric property of the normal distribution, $P(Z<-Z_{\alpha/2}) = \frac{\alpha}{2}$ (see the diagram below).

From the standard normal distribution, we know that

$$P(-Z_{\alpha/2}<Z<Z_{\alpha/2})=1-\alpha$$



To obtain the limit of the interval estimate, we use the standardized form of $Z=\frac{\bar{X}-\mu}{\delta/\sqrt{n}}$ in the above probability statement. i.e., letting

$P(-Z_{\alpha/2}<Z<Z_{\alpha/2})=1-\alpha$ Becomes

$$\Rightarrow P(-Z_{\alpha/2}<\frac{\bar{X}-\mu}{\delta/n}<Z_{\alpha/2})=1-\alpha$$

$$\Rightarrow P(-Z_{\alpha/2}\frac{\delta}{\sqrt{n}}<\bar{X}-\mu<Z_{\alpha/2}\frac{\delta}{\sqrt{n}})=1-\alpha$$

$$\Rightarrow P(-\bar{X}-Z_{\alpha/2}\frac{\delta}{\sqrt{n}}<-\mu<-\bar{X}+Z_{\alpha/2}\frac{\delta}{\sqrt{n}})=1-\alpha$$

$$\Rightarrow P(\bar{X}-Z_{\alpha/2}\frac{\delta}{\sqrt{n}}<\mu<\bar{X}+Z_{\alpha/2}\frac{\delta}{\sqrt{n}})=1-\alpha$$

We can assert with probability $1-\alpha$ that the interval $(\bar{X}-Z_{\alpha/2}\frac{\delta}{\sqrt{n}}<\mu<\bar{X}+Z_{\alpha/2}\frac{\delta}{\sqrt{n}})$ contains the population mean we are estimating.

Thus, $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\left(\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

The end points of the interval, $\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, are called *confidence limits* and the probability $1-\alpha$ is called the *degree of confidence*. In a similar way a $(1-\alpha)100\%$ confidence interval for the population mean μ with unknown variance σ^2 is given by

$$\left(\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

where $t_{\alpha/2}$ is the critical value of t-test statistic providing an area $\frac{\alpha}{2}$ in the right tail of the t-distribution with $n-1$ degrees of freedom, and $S=\sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}$.

Case 2: When the population is non normal.

We use the central limit theorem to approximate the distribution of the sample mean based on large sample ($n\geq 30$). Large sample size is a necessary condition to use the normal distribution. And hence, $Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ~ N(0,1). If σ is unknown we can replace it by its sample estimate S. The resulting $(1-\alpha)100\%$ confidence interval of μ becomes

$$\begin{cases}\left(X-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ X+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), & \text{when } \sigma \text{ is known}\\ \left(X-Z_{\alpha/2}\frac{S}{\sqrt{n}},\ X+Z_{\alpha/2}\frac{S}{\sqrt{n}}\right), & \text{when } \sigma \text{ is unknown}\end{cases}$$

Example 9.1: A drug company is testing a new drug which is supposed to reduce blood pressure. From the six people who are used as subjects, it is found that the average drop in blood pressure is 2.28 millimeter of mercury (mmHg) with a standard deviation of 0.95 mmHg. What is the 95% confidence interval for the mean change in blood pressure? (Assume that the population is normal).

Solution: Given: $\bar{X}=2.28,\ S=0.95,\ n=6$

$(1-\alpha)100\%=95\%\Rightarrow 1-\alpha=0.95\Rightarrow\alpha=0.05\Rightarrow \alpha/2=0.025$

- ✓ $\bar{X}=2.28$ is a point estimate for the population mean drop in blood pressure μ.

A 95% confidence interval of population mean for unknown δ^2 and small sample size is:

$$\left(\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right).$$

And from the t distribution table, $t_{\alpha/2}(n-1)=t_{0.025}(5)=2.571$

$$\left(2.28-(2.571)\frac{0.95}{\sqrt{6}},\ 2.28+(2.571)\frac{0.95}{\sqrt{6}}\right)$$

$$\Rightarrow (2.28\text{-}0.997,\ 2.28+0.997)$$

$$\Rightarrow (1.28,\ 3.27)$$

We are 95% confident that the mean drop in blood pressure lies in between 1.28 mmHg and 3.27 mmHg for the sampled population.

Example 9.2: Punctuality of patients in keeping appointment is of interest to a research team. In a study of patients flow through the office of general practitioners, it was found that a sample of 35 patients were 17.2 minutes late for appointments, on the average. Previous research had shown the standard deviation to be about 8 minutes. The population distribution was felt to be not normal. What is the 90 percent confidence interval for the true mean amount of time late for appointment?

Solution: Given: $\bar{X}=17.2,\ \delta=8,\ n=35$

$(1-\alpha)100\%=90\%\Rightarrow 1-\alpha=0.90\Rightarrow\alpha=0.1\Rightarrow \alpha/2=0.05$

Since the sample size is fairly large (n > 30), and since the population standard deviation is known, according to the central limit theorem, the sampling distribution of sample mean is approximately normal. Thus, a confidence interval of the population mean is given by:

$$\left(\bar{X}-Z_{\alpha/2}\frac{\delta}{\sqrt{n}},\ \bar{X}+Z_{\alpha/2}\frac{\delta}{\sqrt{n}}\right)$$

And from the standard normal distribution table, $Z_{\alpha/2}=Z_{0.05}=1.65$

$$\left(17.2-(1.65)\frac{8}{\sqrt{35}},\ 17.2+(1.65)\frac{8}{\sqrt{35}}\right)$$

$$\Rightarrow (17.2 - 2.2,\ 17.2 + 2.2)$$

$$\Rightarrow (15.0,\ 19.4)$$

Therefore, the 90% confidence interval for true mean amount of time late for appointment is between 15.0 and 19.4 minutes.

9.2 Hypothesis Testing about the Mean

In many circumstances we merely wish to know whether a certain proposition is true or false. The process of hypothesis testing provides a framework for making decisions on an objective basis, by weighing the relative merits of different hypotheses, rather than on a subjective basis by simply looking at the numbers. Different people can form different opinions by looking at data, but a hypothesis test provides a standardized decision-making process that will be consistent for all people.

Statistical hypothesis: is a claim (belief or assumption) about an unknown population parameter values.

Examples of hypothesis:

- ✓ There is association between lung cancer and number of cigarettes an individual smokes.
- ✓ The proportion of female students in Hawassa University is 0.35.
- ✓ In sub-Saharan Africa 40% of individuals are leaving below poverty line.

Hypothesis testing: is the procedure that enables decision-makers to draw inferences about population characteristics by analyzing the difference between the value of sample statistic and the corresponding hypothesized parameter value.

General procedure for hypothesis testing

To test the validity of the claim or assumption about the population parameter, sample is drawn from the population and analyzed. The result of the analysis are used to decide whether the claim is valid or not.

Step 1: State the null hypothesis (H_0) and alternative hypothesis (H_1)

Null hypothesis (H_0): refers to a hypothesized numerical value of the population parameter which is initially assumed to be true. The null hypothesis is always expressed in the form of an equation making a claim regarding the specific value of the population parameter. That is, for example

$$H_0: \mu = \mu_0$$

where μ_0 is hypothesized value of the population mean.

Alternative hypothesis (H_1): is the logical opposite of the null hypothesis. The alternative hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis. For example,

$$H_1: \mu \neq \mu_0 \quad \text{(Two-sided test)}$$

$$H_1: \mu < \mu_0 \quad or \quad H_1: \mu > \mu_0 \quad \text{(One-sided test)}$$

Step 2: State the level of significance α (alpha) for the test

The level of significance is the probability to wrongly reject the null hypothesis H_0 when it is actually true. It is specified by the statistician or

the researcher before the sample is drawn. The most commonly used values of α are 0.10, 0.50 or 0.01.

Step 3: Calculate the appropriate test statistic

Test statistic is a value computed from a sample that is used to determine whether the null hypothesis has to be rejected or not. The choice of suitable test statistic depends on the sampling distribution of the sample statistic. Accordingly, we have the following cases:

Case 1: When the population is normal.

- ✓ If the variance σ^2 is known, the sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and variance $\frac{\sigma^2}{n}$. i.e., $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and the test statistic is $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ~ N(0,1).
- ✓ If the variance σ^2 is unknown the test statistic is, $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ ~t (n-1).

Case 2: When the population is non normal.

We use the central limit theorem to approximate the distribution of the sample mean based on large sample ($n \geq 30$). Large sample size is a necessary condition to use the normal distribution. And hence the test statistic is

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ~ N(0,1). If σ is unknown we can replace it by its sample estimate S.

Step 4: Establish a decision rule (critical or rejection region)

The cut-off point to reject or not reject H_0 depends on the level of significance α, the type of test statistic chosen and the form of the alternative hypothesis. If the value of the test statistic falls in the rejection region, the null hypothesis is rejected, otherwise we do not reject H_0 (see fig 1 below). The value of the sample statistic that separates the regions of acceptance and rejection is called *critical value*. For a specified α, we read the critical values from the Z or t tables, depending on the test statistic chosen.

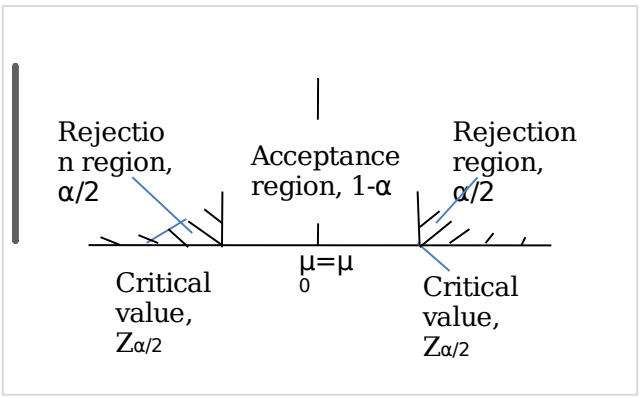
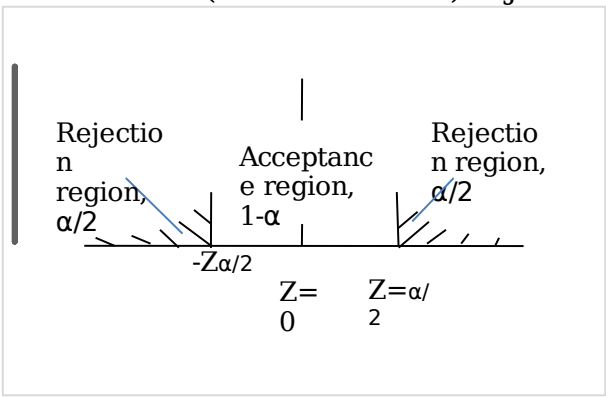


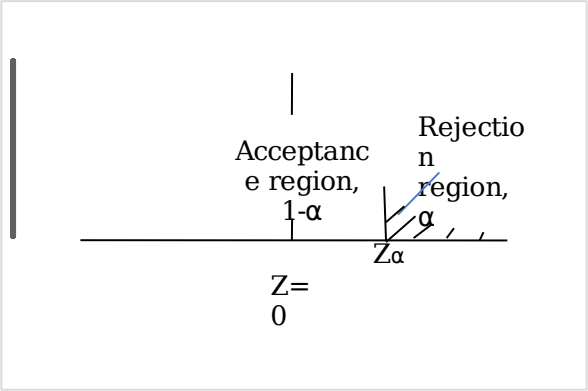
Figure:Area of acceptance and rejection of H_0 (Two-tailed test)

Based on the form of the alternative hypothesis and the test statistic we can make the following decisions:

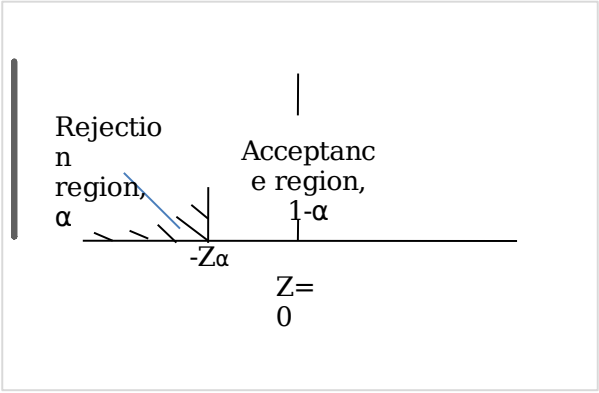
i. For $H_1: \mu \neq \mu_0$ (two-tailed test) reject H_0 if $|Z| > Z_{\alpha/2}$.



ii. For $H_1: \mu > \mu_0$ (right-tailed test) reject H_0 if $Z > Z_\alpha$.



iii. For $H_1: \mu < \mu_0$ (left-tailed test) reject H_0 if $Z < -Z_\alpha$.



We can summarize the decsion rules as follows:

Decision	Alternative hypotheses		
	$H_1: \mu \neq \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$
Reject $H_0: \mu = \mu_0$ if	$\lvert Z \rvert > Z_{\alpha/2}$	$Z > Z_\alpha$	$Z < -Z_\alpha$
Reject $H_0: \mu = \mu_0$ if	$\lvert t \rvert > t_{\alpha/2}(n-1)$	$t > t_\alpha(n-1)$	$t < -t_\alpha(n-1)$

Step 5: Interpret the result.

Errors in Hypotesis Testing

Ideally the hypotesis testing procedure should lead to the rejection of the null hypothesis H_0 when it is false and nonrejection of H_0 when it is true. However, the correct decision is not always possible. Since the decision to reject or do not reject a hypothesis is based on sample data, there is a possibility of committing an incorrect decision or error. Hence, a decision-maker may commit one of the two types of errors while testing a null hypothesis. These errors are summarized as follows:

Decision	Null Hypothesis (H_0)	
	True	False
Reject H_0	Type I error (α)	Correct decision
Accept H_0	Correct decision	Type II error (β)

Type I error is committed if we reject the null hypothesis when it is true. The probability of committing a type I error, denoted by α is called the *level of significance*. The probability level of this error is decided by the decision-maker before the hypothesis test is performed. Type II error is committed if we do not reject the null hypothesis when it is false. The probability of

committing a type II error is denoted by β (Greek letter beta). As type one error increases type two error will decrease (they are inversely proportional). Hence we cannot reduce both errors simultaneously. As the sample size increases both errors will decrease.

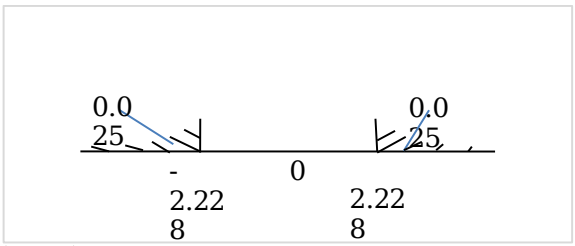
Example 9.3: The life expectancy of people in the year 1999 in a country is expected to be 50 years. A survey was conducted in eleven regions of the country and the data obtained, in years, are given below:

Life expectancy (years): 54.2, 50.4, 44.2, 49.7, 55.4, 47.0, 58.2, 56.6, 61.9, 57.5, and 53.4.

Do the data confirm the expected view? (Assuming normal population) Use 5% level of significance.

Solution: Let μ be the life expectancy of people in the year 1999 in a country.

1. $H_0: \mu = 50$ (The life expectancy of people in the year 1999 in a country is 50 years)
   $H_1: \mu \neq 50$ (The life expectancy of people in the year 1999 in a country is different from 50 years)
2. Level of significance, α = 0.05.
3. Since σ is unknown and the population is normal, the t-test statistic is appropriate.
   Given: n = 11; $\mu_0 = 50$ and we need to compute $\bar{X}$ and s.
   $$\bar{X} = \frac{\sum_{i=1}^{11} x_i}{n} = \frac{54.2 + 50.4 + \ldots.. + 57.5 + 53.4}{11} = \frac{598.5}{11} = 54.41$$
   $$\sum_{i=1}^{11} x_{i^2} = 54.2^2 + 50.4^2 + \ldots.. + 57.5^2 + 53.4^2 = 32799.91$$
   $$S^2 = \frac{1}{n-1}\left[\sum x_{i^2} - \frac{\left(\sum x_i\right)^2}{n}\right] = \frac{1}{10}\left[32799.91 - \frac{(598.5)^2}{11}\right]$$
   $$= \frac{1}{10}(236.07) = 23.607$$
   $$\Rightarrow S = \sqrt{23.607} = 4.859$$
   Then, the t-test statistic is calculated as:
   $$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{54.41 - 50}{4.859/\sqrt{11}} = \frac{4.41}{1.465} = 3.01$$
4. For α = 0.05 and two-tailed test, the critical (table) value is:
   $t_{\alpha/2}(n-1) = t_{0.05/2}(11-1) = t_{0.025}(10) = 2.228$



$|t| = 3.01 > t_{\alpha/2}(n-1) = 2.228$

Since $\Rightarrow$ reject the null hypothesis H_0. That is, the calculated t value lies in the rejection region (the shaded region).

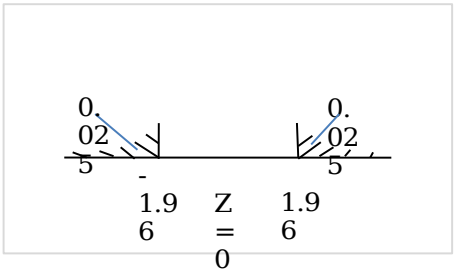
5. **Conclusion:** The data do not confirm the expected view. That is, the life expectancy is different from 50 years at 5% level of significance.

Example 9.4: Suppose that we want to test the hypothesis with a significance level of .05 that the climate has changed since industrialization. Suppose that the mean temperature throughout history is 50 degrees. During the last 40 years, the mean temperature has been 51 degrees and the population standard deviation is 2 degrees. What can we conclude?

Solution:

Let μ be the mean temperature.

1. $H_0: \mu = 50$ (There is no change in temperature since industrialization)
   $H_1: \mu \neq 50$ (There is change in temperature since industrialization)
2. Level of significance, α = 0.05.
3. Since n = 40 is large, the Z-test statistic is appropriate.
   Given: n = 40; $\delta = 2$; $\bar{X} = 51$; $\mu_0 = 50$
   $$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{51-50}{2/\sqrt{40}} = \frac{1}{0.316} = 3.16$$
4. For α = 0.05 and two-tailed test, the critical (table) value is:
   $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$



$|Z| = 3.16 > Z_{\alpha/2} = Z_{0.025} = 1.96$

Since $\Rightarrow$ reject the null hypothesis H_0. That is, the calculated Z value lies in the rejection region (the shaded region).

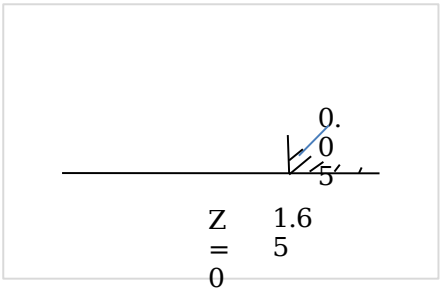
5. **Conclusion:** There has been a change in temperature since industrialization, at 5% level of significance.

Example 9.5: A study was conducted to describe the menopausal status, menopausal symptoms, energy expenditure and aerobic fitness of healthy midwife women and to determine relationship among these factors. Among the variables measured was maximum oxygen uptake (Vo_{2max}). The mean Vo_{2max} score for a sample of 242 women was 33.3 with a standard deviation of 12.14. On the basis of these data, can we conclude that the mean score for a population of such women is greater than 30? Use 5% level of significance.

Solution:

Let μ be the mean Vo_{2max} score for a population of healthy midwife women.

1. $H_0: \mu = 30$ (The mean score for a population of healthy midwife women is 30)
   $H_1: \mu > 30$ (The mean score for a population of healthy midwife women is greater than 30).
2. Level of significance, α = 0.05.
3. Since n = 242 is large, the Z-test statistic is appropriate.
   Given: n = 242; S = 12.14; $\bar{X}$ = 33.3; $\mu_0 = 30$
   $$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{33.3 - 30}{12.14/\sqrt{242}} = \frac{3.3}{0.7804} = 4.23$$
4. For α = 0.05 and right-tailed test, the critical (table) value is:
   $Z_\alpha = Z_{0.05} = 1.65$



   Since $Z = 4.23 > Z_\alpha = 1.65 \Rightarrow$ reject the null hypothesis H_0. That is, the calculated Z value lies in the rejection region (the shaded region).
5. **Conclusion:** The mean Vo_{2max} score for the sampled population of healthy midwife women is greater than 30 at 5% level of significance.

9.3 Test of Association (Independence)

Usually we encounter with nominal scale data. The χ^2 test of association is useful for determining whether there is any relationship or association exists between two nominal variables. For instance, we might be interested in the relationship between HIV status with sex, lung cancer and smoking habit, political affiliation and sex, e t c.

When observations are classified according to two variables or attributes and arranged in a table, the display is called a contingency table as shown below:

	Variable A				
Variable B	A_1	A_2	⋯	A_c	Total
B_1	O_{11}	O_{12}	⋯	O_{1c}	R_1
B_2	O_{21}	O_{22}	⋯	O_{2c}	R_2
.	.	.	⋯	.	.
B_r	O_{r1}	O_{r2}	⋯	O_{rc}	R_r
Total	C_1	C_2		C_c	n

The test of association or independence uses the contingency table format. Here the variables A and B have been classified into mutually exclusive categories. The values O_{ij} in row i and column j of the table shows the observed frequency falling in each joint category i and j. The row and column totals are the sums of their corresponding frequencies. The sum of row or column totals will give grand total n, which represents the sample size. The procedures to test the association between two independent variables is summarized as follows:

Step 1: State the null and alternative hypotheis

H_0**:** There is no association or relationship exists between two variables, that is, the two variables are independent.

H_1**:** There is association or relationship between two variables, that is, the two variables are dependent.

Step 2: State the level of significance, α.

Step 3: Calculate the expected frequencies, E_{ij}, corresponding to the observed frequency in row i and column j. The expected frequencies in each cell are calculated as:

$$E_{ij} = \frac{Row\ i\ total \times Column\ j\ total}{Sample\ size} = \frac{R_i \times C_j}{n}$$

Step 4: Compute the value of test-statistic:

$$\chi_{Cal}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency of row i and coulumn j and E_{ij} is the expected frequency of row i and coulumn j.

Step 5: Find the critical (table) value of $\chi_{\alpha}^2(df)$ (from Appendix..). The value of χ_{α}^2 correponds to an area in the right tail of the distribution.

where df = (Number of rows – 1)(Number of columns – 1) = (r – 1)(c – 1)

Step 6: Compare the calculated and table values of χ^2. Decide wheather the variables are independent or not, using the following decision rule:

Reject H_0 if χ_{Cal^2} is greater than $\chi_{\alpha^2,(df)}$. Otherwise do not reject H_0.

Example 9.6: The following data on the colour of eye and hair for 6800 individuals were obtained from a source:

Hair colour	Eye colour				
	Fair	Brown	Black	red	Total
Blue	1768	808	190	47	2813
Green	946	1387	746	43	3122
Brown	115	444	288	18	865
Total	2829	2639	1224	108	6800

Test the hypothesis that hair colour and eye colour are independently distributed (there is no association between colour of eye and colour of hair) at the level of $\alpha = 0.01$.

Solution:

1. H_0**:** There is no association between hair colour and eye colour.
   H_1**:** There is association between hair colour and eye colour.
2. $\alpha = 0.01$.
3. Calculate the expected frequencies, E_{ij}

$$E_{ij} = \frac{R_i \times C_j}{n}$$

$$E_{11} = \frac{2813 \times 2829}{6800} = 1170.29 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad E_{14} = \frac{2813 \times 108}{6800} = 44.68$$

$$E_{31} = \frac{865 \times 2829}{6800} = 359.87 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad E_{34} = \frac{865 \times 108}{6800} = 13.74$$

Therefore, the contingency table for expected frequencies is as follows:

Hair colour	Eye colour				
	Fair	Brown	Black	red	Total
Blue	1170.29	1091.69	506.34	44.68	2813
Green	1298.84	1211.61	561.96	49.58	3122
Brown	359.87	335.70	155.70	13.74	865
Total	2829	2639	1224	108	6800

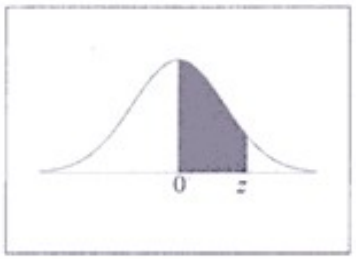
4. Calculate the test statistic:

$$\chi_{Cal^2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi_{Cal^2}=\frac{(1768-1170.29)^2}{1170.29}+\ldots..+\frac{(47-44.68)^2}{44.68}+\frac{(946-1298.84)^2}{1298.84}+\ldots..+$$
$$\frac{(43-49.58)^2}{49.58}+\frac{(115-359.87)^2}{359.87}+\ldots..+\frac{(18-13.74)^2}{13.74}$$
$$\chi_{Cal^2}=1074.43$$

5. Critical value $\chi_{\alpha^2}(df)$
   df = (r – 1) (c – 1) = (3 – 1) (4 – 1) = (2) (3) = 6
   $\chi_{\alpha^2}(df)=\chi_{0.01^2}(6)=16.812$
6. Since $\chi_{Cal^2}=1074.43 > \chi_{\alpha^2}(df)=16.812 \Rightarrow$ Reject H_0.
7. **Conclusion:** There is association between hair colour and eye colour. That is, hair colour and eye colour are dependent.

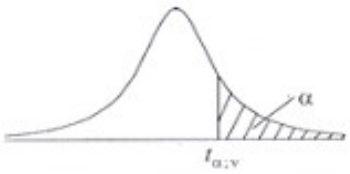
Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998

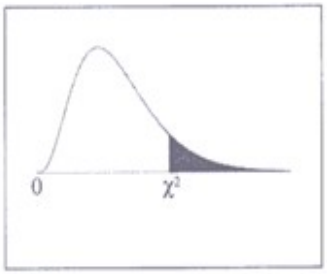
Student t-Distribution Table

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169