1. ELEMENTARY PROBABILITY

1.1 INTRODUCTION

This chapter introduces the basic concepts of probability (or chance). Together with the next chapter it provides a foundation for our study of statistical inference. Here, we examine methods of calculating and using probabilities under various conditions. If you are, say, one of 50 students in a class and it seems that the instructor calls on you each time the class meets, you may accuse that the instructor of not calling on students at random. If, on the other hand, you are one student in a class of ten and you never prepare for class, assuming the instructor will not get around to you, you may be the one who needs to examine probability ideas a bit more.

At the end of the chapter, you will be able to:

- State the fundamentals principles of counting sample points
- Define permutations and combinations
- Understand the basic concepts of probability
- Compute the probability of an event through various probability methods
- Compute and explain what conditional probability is compared to probability of independent events.

History and Relevance of Probability Theory

Early probability theorists

Jacob Bernoulli (1654-1705), Abraham de Moivre (1667 - 1754), the Reverend Thomas Bayes (1702-1761), and Joseph-Lagrange (1736 -1813) developed probability formulas and techniques. In the nineteenth century, Pierre Simon, Marquis de Laplace (1749 - 1827), unified all these early ideas and compiled the first general theory of probability.

Need for probability theory

Probability theory was successfully applied at the gambling tables and, more theory relevant to our study, eventually to other social and economic problems The insurance industry, which emerged in the nineteenth century, required precise knowledge about the risk of loss in order to calculate premiums. Within 50 years, many learning centers were studying probability as a tool for understanding social phenomena. Today, the mathematical theory of probability is the basis for statistical applications in both social and decision-making research.

Examples of the use of probability theory

Probability is a part of our everyday lives. In personal and managerial decisions, we face uncertainty and use probability theory whether or not we admit the use of something so sophisticated. When we hear a weather forecast of a 70 percent chance of rain, we change our plans from a picnic to a pool game. Playing bridge, we make some probability estimate before attempting finesse. Managers who deal with inventories of highly styled women's clothing must wonder about the chances that sales will reach or exceed a certain level, and the buyer who stocks up on skateboards considers the probability of the life of this particular fad. Before Muhammad Ali's highly publicized fight with Leon Spinks, Ali was reputed to have said, "I'll give you **odds** I'm still the greatest when it's over." And when you begin to study for the inevitable quiz attached to the use of probability, you may ask yourself, "What are the chances the professor will ask us to recall something about the history of probability theory?"

We live in a world in which we are unable to forecast the future with complete certainty. Our need to cope with uncertainty leads us to the study and use of probability theory. In many instances we, as concerned citizens, will have some knowledge about the possible outcomes of a decision. By organizing this information and considering it systematically, we will be able to recognize our assumptions, communicate our reasoning to others, and make a sounder decision than we could by using a shot-in-the-dark approach.

❖ SELF TEST ACTIVITY 1-1

1. The insurance industry uses probability theory to calculate premium rates, but life insurer knows for certain that every policyholder is going to die. Does this mean that probability theory does not apply to the life insurance business? Explain.
2. "Use of this product may be hazardous to your health. This product contains saccharin which has been determined to cause cancer in laboratory animals." How might probability theory have played a part in this statement?
3. Is there really any such thing as a "uncalculated risk"? Explain.
4. A well-known soft drink company decides to alter the formula of its oldest and most popular product. How might probability theory be involved in such a decision?

1.2 SOME BASIC DEFINITIONS AND CONCEPTS IN PROBABILITY

In general, probability is the chance something will happen. Probabilities are expressed as fractions (1/6; 1/2, 8/9) or as decimals (0.167, 0.500, 0.889) between zero and 1. Assigning a probability of zero means that something can never happen; a probability of 1 indicates that something will always happen.

Event: In probability theory, an *event* is one or more of the possible outcomes of doing something. If we flip a coin, getting a tail would be an *event,* and getting a head would be another event. Similarly, if we are drawing from a deck of cards, selecting the ace of spades would be an event. An example of an event closer to your life is being picked from a class of 100 students to answer a question. When we hear the frightening predictions of highway traffic deaths, we hope not to be one of those events.

Experiment: The activity that produces such an event is referred to in probability theory' as an *experiment*. In other words statistical experiment is defined as a trial that generates two or more possible outcomes, but as to which outcome occurs is not known in advance. Using this formal language, we could ask the question, "In a coin-flip *experiment,* what is the probability of the event *head?*" And, of course, if it is a fair coin with an equal chance of coming down on either side (or no chance of landing on its edge), we would answer, "½" or "0.5."

Sample space*:* The set of all possible outcomes of an experiment is called the *sample space* for the experiment and we denote it by S.

Example: In the coin-flip experiment, the sample space is: S = {head, tail}

Example: In the card drawing experiment, the sample space has fifty-two members: ace of hearts, deuce of hearts, and so on.

Most of us are less excited about coins or cards than we are interested in questions like, "What are the chances of making that plane connection?" or, "What are my chances of getting a second job interview?" In short, we are concerned with the chances that certain events will happen.

Mutually exclusive events*:* Events are said to be *mutually exclusive* if one and only one of them can take place at a time. Consider again our example of the coin. We have two possible outcomes, heads and tails. On any flip, either heads or tails may turn up, but not both. As a result, the events heads and tails on a single flip are said to be mutually exclusive. Similarly, you will either pass or fail this course or, before the course is over, you may drop it without a grade. Only one of those three outcomes can happen; they are said to be mutually exclusive events. The crucial question to ask in deciding whether events are really mutually exclusive is, "Can two or more of these events occur at one time?" If the answer is yes, the events are *not* mutually exclusive.

Collectively exhaustive list: When a list of the possible events that can result from an experiment includes every possible outcome, the list is said to be *collectively exhaustive*. In our coin example, the list "head and tail" is collectively exhaustive (unless, of course, the coin stands on its edge when we flip it). In a presidential campaign, the list of outcomes "Democratic candidate and Republican candidate" is *not* a collectively exhaustive list of outcomes, since_an independent candidate or the candidate of-another-party-could conceivably win.

❖ Self Test Activity 1-2

1. Give a collectively exhaustive list of the possible outcomes of flipping two dice.
2. Which of the following are pairs of mutually exclusive events in the drawing of a single card from a standard deck of fifty-two?
   a. A heart and a queen
   b. An even number and a spade
   c. A club and a red card
   d. An ace and an even number
3. Which of the following are mutually exclusive outcomes in the rolling of two dice?
   a. A total of 5 and a five on one die
   b. A total of 7 and an even number of points on both dice
   c. A total of 8 and an odd number of points on both dice
   d. A total of9 points and a two on one die
   e. A total of 10 points and a four on one die.

1.3 Techniques of Counting

To evaluate probabilities associated with chance outcomes, we have a number of elaborate counting techniques at our disposal and hence a need to discuss here. However, probabilities obey certain mathematical laws; their computations can often be simplified.

One of the problems in statistics is that we must consider and attempt to evaluate the element of chance associated with the occurrence of certain events when an experiment is performed. These problems belong in the field of probability, a subject to be introduced in this section. In many cases we shall be able to solve a probability problem by counting the number of points in the sample space without actually listing each element. The fundamental principle of counting, often referred to as the multiplication rule, is stated in the following theorem.

Multiplication Rule: *If an operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, then the two operations can be performed together in $n_1 n_2$ ways.*

Example: How many sample points are in the sample space when a pair of coins is flipped once?

Solution: The first coin can land in any of 2 ways. For each of these 2 ways the second coin can also land in 2 ways. Therefore, the pair of coin can land in (2) (2) = 4 ways.

Example: How many sample points are in the sample space when a pair of dice is thrown once?

Solution: The first die can land in any of 6 ways. For each of these 6 ways the second die can also land in 6 ways. Therefore, the pair of dice can land in (6) (6) = 36 ways. In this case list the 36 sample points in the sample space.

The generalized multiplication rule covering *k* operations is stated in the following theorem.

Generalized Multiplication Rule: If an operation can be performed in n_1 ways, if for each of these a second operation can be performed in n_2 ways, if for each of the first two a third operation can be performed in n_3 ways, and so on, then the sequence of k operations can be performed in $n_1 n_2 \ldots n_k$ ways.

Example: A College student must take a social science course, a mathematics course, and a humanities course. If she may select any of 6 social science courses, any of 3 mathematics courses and any of 4 humanities courses, how many ways can she arrange her program?

Solution: The total number of ways she can arrange would be (6)(3)(4) = 72.

Example: How many even four-digit numbers can be formed from the digits 0, 1, 2, 5, 6, and 9 if each digit can be used only once?

Solution: Since the number must be even, we have only 3 choices for the units position. For each of these we have 5 choices for the tens position and 4 choices for the hundreds position, and 3chices for the thousands position. Therefore, we can form a total of (3)(5)(4)(3) = 180 even four-digit numbers.

Frequently, we are interested in a sample space that contains as elements all possible orders or arrangements of a group of objects. For example, we might want to know how many arrangements are possible for sitting 6 people around a table, or we might ask how many different orders are possible to draw 2 lottery tickets from a total of 20. The different arrangements are called **permutations.**

Permutation: A *permutation* is an arrangement of all or part of a set of objects.

Consider the three letters *a, b,* and c. The possible permutations are *abc, acb, bac, bca, cab,* and *cba* Thus we see that there are 6 distinct arrangements. Using the multiplication rule, we could have arrived at the answer without actually listing the different orders. There are 3 positions to be filled from the letters *a, b,* and c. Therefore, we have 3 choices for the first position, and 2 for the second, leaving only 1choice for the last position, giving a total of (3)(2)(1) = 6 permutations. In general, n distinct objects can be arranged in $n(n - l)(n - 2) \ldots (3)(2)(1)$ ways. We represent this product by the symbol $n!$, which is read "*n* factorial." Three objects can be arranged in 3! = (3)(2)(1) = 6 ways. By definition 1! = 1 and 0! = 1.

The number of permutations of the four letters *a, b,* c, and *d* will be 4! = 24. Let us now consider the number of permutations that are possible by taking the 4 letters 2 at a time. These would be *ab, ac, ad, ba, ca, da, bc, cb, bd, db, cd,* and *dc*. Using Multiplication Rule, we have 2 positions to fill with 4 choices for the first and 3 choices for the second, a total of (4)(3) = 12 permutations. In general, n distinct objects taken r at a time can be arranged in $n(n.- l)(n - 2) \ldots (n - r + 1)$ ways. We represent this product by the symbol ${}_nP_r = n!/(n - r)!$.

The number of permutations of n distinct objects taken r at a time is

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

Example: Two lottery tickets are drawn from 30 for first and second prizes. Find the number of ways of doing this.

Solution: The total number of sample points is

$$_{30}P_2 = \frac{30!}{(30-2)!} = \frac{30!}{28!} = (30)(29) = 870.$$

Example: How many ways can a football team schedule 3 holiday games with 3 teams if they are all available on any of 5 possible dates?

Solution: The total number of possible schedules is

$$_5P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = (5)(4)(3) = 60.$$

Permutations that occur by arranging objects in a circle are called circular permutations. Two circular permutations are not considered different, unless corresponding objects in the two arrangements are preceded or followed by a different object as we proceed in a clockwise direction. For example, if 4 people are sitting in a round table, we do not have a new permutation if they all move one position in a clockwise direction. By considering 1 person in a fixed position and arranging the other 3 in 3! ways, we find that there are 6 distinct sitting arrangements.

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

So far we have considered permutations of distinct objects. That is, all the objects were completely different or distinguishable. Obviously, if the letters b and c are both equal to x, .then the 6 permutations of the letters *a*, *b*, and c become *axx*, *axx*, *xax*, *xax*, *xxa*, and *xxa*, of which only 3 are distinct. Therefore, with 3 letters, 2 being the same, we have $\frac{3!}{2!} = 3$ distinct permutations. With the 4 letters *a*, *b*, *c*, and *d* we had 24 distinct permutations. If we let *a* = *b* = x and c = *d* = *y*, we can only list the following: *xxyy*, *xyxy*, *yxxy*, *yyxx*, *xyyx*, and *yxyx*. Thus we have $\frac{4!}{2!2!} = 6$ distinct

The number of distinct permutations of n things of which n_l are of one kind, n_2 of a second kind, …, n_k of a k^{th} kind is

$$\frac{n!}{n_1!n_2!n_3!\ldots n_k!}$$

permutations.

Example: In how many different ways can 3 oaks, 4 pines, and 2 apples be arranged along a property line if one does not distinguish between trees of the kind?

Solution: The total number of distinct arrangements is $\frac{9!}{3!4!2!}=1260$

Often, we are concerned with the number of ways of partitioning a set of *n* objects into *r* subsets, called cells. A partition is said to be achieved if

- the intersection of every possible pair of the *r* subsets is the empty set 0 and
- if the union of all subsets gives the original set.
- The order of the elements within a cell is of no importance.

Example: Consider the set *{a, e, i, o, u}*. The possible partitions into 2 cells, in which the first cell contains 4 elements and the second cell 1 element, are *{(a, e, i, 0), (u)}*, *{(a, i, 0, u), (e}*, *{(e, i, o, u), (a)}*, *{(a, e, o, u),* (i)}, and *{(a, e, i, u), (o)}*. We see that there are 5 such ways to partition a set of 5 elements into 2 subsets or cells containing 4 elements in the first cell and 1 element in the second, . The number of partitions for this illustration is denoted by

$$\binom{5}{4,1}=\frac{5!}{4!1!}=5$$

where the top number represents the total number of elements and the bottom numbers represent the number of elements going into each cell, We state this more generally as follows:

The number of ways of partitioning a set of n objects into r cells with n_1 elements in the first cell, n_2 elements in the second, and so on, is

$$\binom{n}{n_1,n_2,\ldots,n_r}=\frac{n!}{n_1!n_2!\ldots n_r!}$$

Example: How many ways can 10 people be assigned to 2 triples and 2 double rooms?

Solution: The total number of possible partitions would be

$$\binom{10}{3,3,2,\ 2}=\frac{10!}{3!3!2!2!}=25,200$$

In several problems we are interested in the number of ways of *selecting r* objects from *n* without regard to order. These selections are called **combinations**. A combination creates a partition of n distinct objects with 2 cells or subsets, one cell containing the *r* objects selected and the other cell containing the *n* - *r* objects that are not selected or left.

The number of such combinations, denoted by $\binom{n}{r,n-r}$, is usually shortened to $\binom{n}{r}$, since the number of elements in the second cell must be *n* - *r*,

The number of combinations of n distinct objects taken r at a time is

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

Example: From a group of 4 men and 5 women, how many committees of size 3are possible

a) With no restriction?
b) With 2 men and a women?
c) With 2 men and a woman if a certain man must be on the committee?

Solution:

a) The number of ways selecting 3 persons from a total of 9 $\binom{9}{3}=\frac{9!}{3!(9-3)!}=\frac{9x8x7x6!}{3x2x1x6!}=84$

b) The number of ways of selecting 2 men from 4 men is

$$\binom{4}{2}=\frac{4!}{2!\text{x}(4-2)!}=\frac{4\text{x}3}{2}=6\text{ ways}$$

The number of ways of selecting a woman from 5 women is

$$\binom{5}{1}=\frac{5!}{1!\text{x}(5-1)!}=\frac{5\text{x}4!}{4!}=5\text{ ways}$$

Therefore, the number of ways of selecting 2 men and a woman is

$$\binom{4}{2} \mathrm{x} \binom{5}{1} = 6 \mathrm{x} 5 = 30 \text{ ways}$$

c) The number of ways of selecting 2 men where a certain man is selected will leave us to select only 1 man from a total of (4-1) = 3 men and this is done in

$$\binom{3}{1} = \frac{3!}{1! \mathrm{x} 2!} = 3 \text{ ways}$$

The number of ways of selecting a woman from 5 women is

$$\binom{5}{1} = \frac{5!}{1! \mathrm{x} (5-1)!} = \frac{5 \mathrm{x} 4!}{4!} = 5 \text{ ways}$$

Therefore, the number of ways of selecting 2 men and a woman where a certain man is selected is

$$\binom{3}{1} \mathrm{x} \binom{5}{1} = 3 \mathrm{x} 5 = 15 \text{ ways}$$

❖ SELF TEST ACTIVITY 1-3

1. A committee of 3 is to be formed from 7 men and 3 women. The committee must comprise at most 2 men. In how many ways can this be done?

1.4 PROBABILITY OF AN EVENT

Introduction

The statistician is basically concerned with drawing conclusions or inferences from experiments involving uncertainties. For these conclusions and inferences to be accurately interpreted, an understanding of probability theory is essential. What do we mean when we make the statements?

i. "Million will probably win the running match,"
ii. "I have a 50:50 chance of getting an even number when a die is tossed,"
iii. "I am not likely to win at bingo tonight," or
iv. "Most of our graduating class will probably be married within 2 years"?

In each case we are expressing an outcome of which we are not certain, but because of past information or from an understanding of the structure of the experiment, we have some degree of confidence in the validity of the statement.

The mathematical theory of probability for finite sample spaces provides a set of real numbers called weights or probabilities, ranging from 0 to 1, which allow us to evaluate the likelihood of occurrence of events. To every point in the sample space we assign a probability such that the sum of all probabilities is 1. If we have reason to believe that a certain sample point is quite likely to occur when the experiment is conducted, the probability assigned should be close to 1. On the other hand, a probability closer to zero is assigned to a sample point that is not likely to occur. In many experiments, such as tossing a coin or a die, all the sample points have the same chance of occurring and are assigned equal probabilities. For points outside the sample space, that is, for simple events that cannot possibly occur, we assign a probability of zero.

To find the probability of an event *A*, we sum all the probabilities assigned to the sample points in *A*. This sum is called the probability of *A* and is denoted by P(A). Thus the probability of the set ϕ is zero and the probability of S is 1

1.5 SOME PROBABILITY RULES

THREE TYPES OF PROBABILITY

There are three basic different approaches of classifying probability. These are:

1. Classical Approach,
2. Relative Frequency Approach,
3. Subjective Approach

❖ **THE CLASSICAL APPROACH**

This approach defines probability as follows:

Definition: Let a random experiment result in n equally likely and mutually exclusive outcomes and let A be any event having n(A) of these outcomes. Then, the probability of A, denoted by P(A), is defined as

$$P(A)=\frac{n(A)}{n}=\frac{number\, of\; cases\; favorable\, to\; A}{Total\, number\, of\; equally\;\; likely\, cases}$$

It must be emphasized that in order for the equation to be valid, each of the possible outcomes must be equally likely. This is a rather complex way of defining something that may seem intuitively obvious to us, but we can use it to write our coin-flip and dice-rolling examples in symbolic form. First, we would state the question, "What is the probability of getting a head on one flip?" as:

P(Head)

Then, using the equation above, we get:

$$P(Head)=\frac{1}{1+1}$$

$$=0.5\, or\, \frac{1}{2}$$

Number of outcomes of one flip where the event occurs (in this case, the number that will produce a

Total number of possible outcomes of one flip (in this case, a head or a tail)

And for the dice-rolling example:

$$P(5)=\frac{1}{1+1+1+1+1+1}$$

$$=\frac{1}{6}$$

Number of outcomes of one roll of the die that will produce a 5

Total number of possible outcomes of one of one roll of the die (getting a 1, a 2, a 3, a 4, a 5, or a 6)

Classical probability is often called *a priori* probability, because if we keep using orderly examples like fair coins, unbiased dice, and standard decks of cards, we can state the answer in advance (a priori) *without* flipping a coin, rolling a die, or drawing a card. We do not have to perform experiments to make our probability statements about fair coins, standard card decks, and unbiased dice. Instead, we can make statements based on logical reasoning before any experiments take place.

This approach to probability is useful when we deal with card games, dice games, coin flips, and the like but has serious problems when we try to apply it to the less orderly decision problems we encounter in business and management problems. The classical approach to probability assumes a world that does not exist. It assumes away situations that are very unlikely but that could conceivably happen. Such occurrences as a coin landing on its edge, your classroom burning down during a discussion *of* probabilities, or your eating pizza while on a you are attending lecture classes are all extremely unlikely but not impossible. Nevertheless, the classical approach assumes them all away. Classical probability also assumes a kind *of* symmetry about the world, and that assumption can get us into trouble. Real-life situations, disorderly and unlikely as they often are, make it useful to define probabilities in other ways.

RELATIVE FREQUENCY OF OCCURRENCE

Suppose we begin asking ourselves complex questions such as, "What is the probability that I will live to be 500" or, "What are the chances that I will blow one *of* my stereo speakers if turn my 1500-watt amplifier up to wide open?" or, "What is the probability that the location *of* a new paper plant on the river near our town will cause a substantial fish kill?" We quickly see that we may not be able to state in advance, without experimentation, what these probabilities are. Other approaches may be more useful.

In the 1800s, British statisticians, interested in a theoretical foundation for calculating risk *of* losses in life insurance and commercial insurance, began defining probabilities from statistical data collected on births and deaths. Today this approach is called *relative frequency of occurrence*. It defines probability as either:

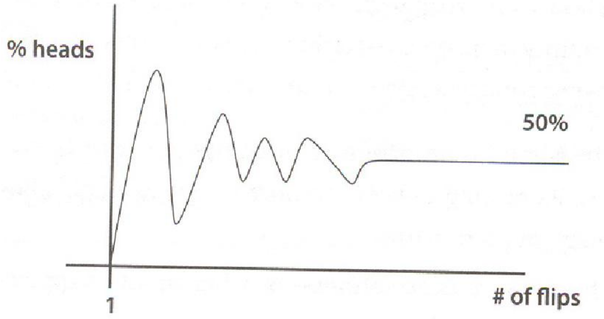
1. The observed relative frequency *of* an event in a very large number of trials, or
2. The proportion *of* times that an event occurs in the long run when conditions are stable.

This method uses the relative frequencies *of* past occurrences as probabilities. We determine how often something has happened in the past and use that figure to predict the probability that it will happen again in the future.

Example: Suppose an insurance company knows from past actuarial data that *of* all males *40* years old, about *60* out *of* every *100,000* will die within a one-year period. Using this method, the company estimates the probability *of* death for that age group as:

$$P(death\ of\ 40\ years\ old) = \frac{50}{100,000} = 0.0005$$

A second characteristic *of* probabilities established by the relative frequency of occurrence method can be shown by flipping one *of* our fair coins *300* times. The figure below illustrates the outcomes *of* these *300* flips.



Note that the x-axis is the number of flops-starting with one flip-and the y-axis the percentage of flips that result in heads. As we can see, the first must have resulted in tails because the percentage of head on the first slip is zero. Here we can see that although the proportion of heads was far from 0.5 in the first few flips, it seemed to stabilize and approach 0.5 as the number of flips increased. In statistical language, we would say that the relative frequency becomes stable as the number of flips becomes large (if we are flipping the coin under uniform conditions). Thus, when we use the relative frequency approach to establish probabilities, our probability figure will gain accuracy as we increase the number of observations. Of course, this improved accuracy is not free; although more flips of our coin will produce a more accurate probability of heads occurring, we must bear both the time and the cost of additional observations.

One difficulty with the relative frequency approach is that people often use it without evaluating a sufficient number of outcomes. If you heard someone say, "My aunt and uncle got the flu this year, and they are both over 65, so everyone in that age bracket will probably get the flu." You would know that your friend did not base his assumptions on enough evidence. He had insufficient data for establishing a relative frequency of occurrence probability.

But what about a different kind of estimate, one that seems not to be based on statistics at all? Suppose your school's basketball team lost the first ten games of the year. You were a loyal fan, however, and bet $100 that your team would beat Indiana's in the eleventh game. To everyone's surprise, you won your bet. We would have difficulty convincing you that you were statistically incorrect. And you would be right to be skeptical about our argument. Perhaps without knowing that you did so, you may have based your bet on the statistical foundation described in the next approach to establishing probabilities.

SUBJECTIVE PROBABILITIES

Subjective probabilities are based on the beliefs of the person making the probability assessment. In fact, subjective probability can be defined as the probability assigned to an event by an individual, based on whatever evidence is available. This evidence may be in the form of relative frequency of past occurrences, or it may be just an educated guess. Probably the earliest subjective probability estimate of the likelihood of rain occurred when a farmer said, "My “teff” hurt; I think we're in for a downpour.” Subjective assessments of probability permit the widest flexibility of the three concepts we have discussed. The decision maker can use whatever evidence is available and temper this with personal feelings about the situation.

Subjective probability assignments are frequently found when events occur only once or at most a very few times.

Example: Suppose it is your job to interview and select a new employee. You have narrowed your choice to three people. Each has an attractive appearance, a high level of energy, great self-confidence, a good record of past accomplishments, and a state of mind that seems to welcome challenges. What are the chances each will relate to clients successfully? Answering this question and choosing among the three will require you to assign a subjective probability to each person's potential.

Example: A Judge is deciding whether to allow the construction of a nuclear power plant on a site where there is some evidence of a geological fault. He must ask himself the question, "What is the probability of a major nuclear accident at this location?" The fact that there is not relative frequency of occurrence evidence of previous accidents at this location does not excuse him from making a decision. He must use his best judgment in trying to determine the subjective probabilities of a nuclear accident.

In many situations, higher level social and managerial decisions are concerned with specific, unique situations, rather than with a long series of identical situations, decision makers at this level make considerable use of subjective probabilities.

❖ SELF TEST ACTIVITY 1-4

1. The ABC Company drafted a set of wage and benefit demands to be presented to management. To get an idea of worker support for the package. He randomly selects the two largest groups of workers at his plant, the machinists (M) and the inspectors (I). He selects 30 of each group with the following results:

OPINION OF PACKAGE	M	I
Strongly support	8	12
Mildly support	13	5
Undecided	1	3
Mildly oppose	3	3
Strongly oppose	5	7
	30	**30**

i. What is the probability that a machinist randomly selected from the selected group mildly supports the package?
ii. What is the probability that an inspector randomly selected from the selected group is undecided about the package?
iii. What is the probability that an employee randomly selected from the group strongly or mildly supports the package?

1. What types of probability estimates are these?

Note: If you can solve the self activity problems 5-4, you may proceed otherwise re-read the previous topic again. Check your answer with the answer key at the back of this text book.

Further Activities

2. Determine the probabilities of the following events in drawing a card from a standard deck of 52 cards:

i. A seven
ii. A black card
iii. An ace or a king
iv. A black two or a black three
v. A red face card (king, queen, or jack)
vi. What type of probability estimates are these?

3. Below is a frequency distribution of annual sales commissions (in Birr) from a survey of 300 salespeople.

ANNUAL COMMISSION	FREQUENCY
0- 4,999	15
5,000- 9,999	25
10,000 -14,999	35
15,000-19,999	125
20,000-24,999	30
25,000 +	70

Based on this information, what is the probability that a media salesperson makes a commission?

i. between Birr5000 and Birr10,000?
ii. less than Birr15,000?
iii. more than Birr20,000?
iv. between Birr15,000 and Birr20,000?

A. Basic Additive Rule: If A and B are any two events, then

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

B. If A, B and C are mutually exclusive, then

1) $P\ (A\cup B) = P\ (A) + P\ (B)$

2) $P\ (A\cup B\cup C) = P\ (A) + P\ (B) + P\ (C)$ also

C. In general, If A_1, A2, A_3, …, A_n B are mutually exclusive, then

$$P\ (A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = P\ (A_1) + P\ (A_2) + P\ (A_3) + \ldots + P\ (A_n)$$

4. The office manager of an insurance company has the following data on the functioning of the copiers in the office:

COPIER	DAYS FUNCTIONING	DAYS OUT OF SERVICE
A	209	51
B	217	43
C	258	2
D	229	31
E	247	13

What is the probability of a copier's being out of service, based on this data?

5. Classify the following probability estimates as to their type (*classical*, *relative frequency*, or *subjective*):
   a. The probability of scoring on a penalty shot in football game is 0.47.
   b. The probability that the current President of the country will resign is 0.15.
   c. The probability of rolling 2 sixes with 2 dice is 1/36.
   d. The probability that you will go to South Africa this year is 0.84.

1.6 SOME BASIC PROPERTIES OF PROBABILITY

Additive Rules: Often it is easier to calculate the probability of an event from known probabilities of other events. This may well be true if the event in question can be represented as the union of other events or as the complement of some event. Several important laws that frequently simplify the computation of probabilities follow. The first, called the additive rule, apply to unions of events.

Example: The probability that a student passes FNDE 101 course is $\frac{13}{52}$, and the probability that she passes BAIS 211 course is $\frac{4}{52}$. If the probability of passing at least one course is $\frac{4}{13}$, what is the probability that she will pass both courses?

Solution: Let F = Event of passing FNDE 211 and B = Event of passing BAIS 211

Then by transposing the terms in Additive Rule A given above we have

$P\ (F \cap B) = P\ (F) + P\ (B) - P\ (E \cup B)$

$= \frac{13}{52} + \frac{4}{52} - \frac{4}{13}$

$= \frac{1}{52}$

Example: What is the probability of getting a total of 9 or 11 when a pair of dice is rolled once?

Solution: Let A = Event that 9 appears and B = Event that 11 occurs. Now a total of 7 occurs for 6 of the 36 outcomes and a total of 11 occurs for 2 of the 36 outcomes. Since all sample points (outcomes) are equally likely to occur, we have P (A) = $\frac{6}{36}$ and P (B) = $\frac{2}{36}$. The events A and B are mutually exclusive, since a total 9 and 11 cannot both occur on the same toss. Therefore,

$P\ (A \cup B) = P\ (A) + P\ (B)$

$= \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}$

1. If A and A′ are complementary events, then $P\ (A') = 1 - P(A)$

Proof:

$A \cup A' = S \quad P\ (A \cup A') = P\ (S)$
$P(A) + P(A') = P(S)$ since A and A′are mutually exclusive
$P(A) + P(A') = 1$ since $P(S) = 1$
$\therefore P(A') = 1 - P(A)$

2. $P\ (\phi) = 0$

Proof :

$S=\phi\cup S$ $P(S) = P(\phi\cup S)$ $1 = P(\phi)+P(S)$ since $\phi\cap S=\phi$

$1= P(\phi)+1$ Therefore, $P(\phi)=1-1=0$

3. Let A and B be subsets of S. If $A\subseteq B$, then $P(A)\leq P(B)$

   Proof:

   $A\cup(A'\cap B)=B$ see the diagram

   $P[A\cup(A'\cap B)]=P(B)$

   $P(A)+P(A'\cap B)=P(B)$ since A and $A'\cap B$ are mutually exclusive

   $P(A)\leq P(B)$ since $P(A'\cap B) \geq 0$

4. Let A be an event of the sample space S. Then $0<P(A)<1$

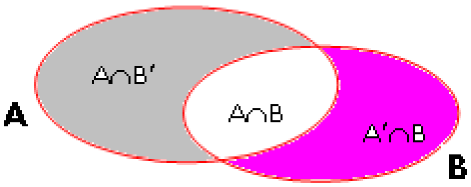
   Proof

   $0<P(A)$ is true by definition

   Since $A\subseteq S$, $P(A) < 1$ by the above result above.

   Therefore, $0 <P(A)<1$

5. Let $A\subseteq S$ and $B\subseteq S$. Then:
   a. $P(A'\cap B)=P(B) - P(A\cap B)$
   b. $P(A\cap B')=P(A) - P(A\cap B)$



Proof: From the Venn diagram above, we have

$B= (A\cap B) \cup (A'\cap B)$ and $A= (A\cap B) \cup (A\cap B')$.

From this it follows that:

$P(B) =P[(A\cap B) \cup (A'\cap B)]$

$P(B) =P(A\cap B) +P(A'\cap B)$ since $A\cap B$ and $A'\cap B$ are mutually exclusive and

$P(A'\cap B) =P(B) - P(A\cap B)$

$P(A) =P[(A\cap B) \cup (A\cap B')]$

$P(A) =P(A\cap B) + P(A\cap B')$ since $A\cap B$ and $A\cap B'$ are mutually exclusive

$P(A\cap B') =P(A) - P(A\cap B)$

❖ SELF TEST ACTIVITY1-5

1. Find the errors in each of the following statements:

   a. The probabilities that an automobile salesperson will sell 0, 1, 2, or 3 cars on any given day in February are, respectively, 0.19, 0.38, 0.29, and 0.15.
   b. The probability that it will rain tomorrow is 0.40 and the probability that it will not rain tomorrow is 0.52.
   c. The probabilities that a printer will make 0, 1, 2, 3, or 4 or more mistakes in printing a document are, respectively, 0.19, 0.34, - 0.25, 0.43, and 0.29.
   d. On a single draw from a deck of playing cards the probability of selecting a heart is $\frac{1}{4}$, the probability of selecting a black card is $\frac{1}{2}$, and the probability of selecting both a heart and a black card is $\frac{1}{8}$.
2. If *A and B are* mutually exclusive events and P(A) = 0.3 and P(B*)* = 0.5, find
   a. P(A U B);
   b. (b) P(A');
   c. P(A'∩ B).

   Hint: Construct Venn diagrams and fill in the probabilities associated with the various regions.

Activity

1. Three men are seeking government office. Candidates *A* and *B* are given about the same chance of winning, but candidate C is given twice the chance of either *A* or *B.*
   a. What is the probability that C wins?
   b. What is the probability that *A* does not win?
2. A box contains 500 envelopes of which 75 contain Birr100 in cash, 150 contain Birr25, and 275 contain Birr10. An envelope may be purchased for Birr25. What is the sample space for the different amounts of money? Assign probabilities to the sample points and then find the probability that the first envelope purchased contains less than Birr100.
3. If *A, B,* and C are mutually exclusive events and P(A) = 0.2, P(B*)* = 0.3, and *P*(C) = 0.2, find
   a. P(A U *B* U C);
   b. (b) P[A' n (B U C)];

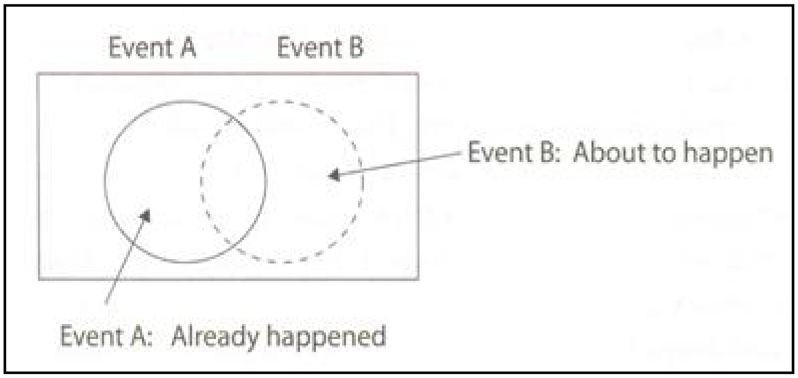
   Hint: Construct Venn diagrams as in Exercise 6.
4. A pair of dice is flipped. Find the probability of getting
   a. a total of 8;
   b. at most a total of 5.

5. Two cards are drawn in succession from a deck without replacement. What is the probability that both cards are greater than 2 and less than 8?
6. If 3 books are picked at random from a shelf containing 6- novels, 3 books of poems, and a dictionary, what is the probability that
    a. the dictionary is selected?
    b. 2 novels and I book of poems are selected?
7. In a college graduating class of 100 students, 54 studied mathematics, 69 studied history, and 35 studied both mathematics and history. If one of these students is selected at random, find the probability that
    a. the student takes mathematics or history;
    b. the student does not take either of these subjects; (c) the student takes history but not mathematics.
8. Suppose that in a senior college class of 500 students it is found that 210 smoke, 258 drink alcoholic beverages, 216 eat between meals, 122 smoke and drink alcoholic beverages, 83 eat between meals and drink alcoholic beverages, 97 smoke and eat between meals, and 52 engage in all three of these bad health practices. If a member of this senior class is selected at random, find the probability that the student
    a. smokes but does not drink alcoholic beverages;
    b. eats between meals and drinks alcoholic beverages but does not smoke; (c) neither smokes nor eats between meals.
9. From past experiences a stockbroker believes that under present economic conditions a customer will invest in tax-free bonds with a probability of 0.6, will invest in mutual funds with a probability of 0.3, and will invest in both tax-free bonds and mutual funds with a probability of 0.15. At this time, find the probability that a customer will invest
    a. in either tax-free bonds or mutual funds;
    b. in neither tax-free bonds nor mutual funds.
10. The odds in favor of an event E are equal as a to b if and only if P(E) = $\frac{a}{a+b}$
    a. If an insurance company quotes odds of 3 to 1 for the event that an individual 65 years of age will survive another 10 years, what is the probability assigned to this event?
    b. If the probability of a successful transplant operation is $\frac{1}{8}$, what are the odds against this type of surgery?

1.7 CONDITIONAL PROBABILITY AND INDEPENDENCE

A conditional probability is written as P (B/A), or the probability of event B occurring given that (or *if)* A has already occurred. We need to define these types of probabilities before we can define joint or intersection probabilities [e.g. P (A and B)].

Assume Event A is that you have your master degree, and that Event B is that you get a higher salary upon graduation. We can express the conditional probability, P (B/A), as the probability of getting a higher salary if (or given that) you have a master degree. Pictorially, we can draw this as shown below. Note that Event A, having your master degree, has already happened. Event B, getting a higher salary, may be about to happen.



Now we must determine the probability of getting this higher salary if you have your master degree.

It should be clear that this conditional probability is not the same as the intersection probability, P(A and B) where, people have both master degrees *and* higher salaries. This is not the same as the conditional probability we seek!

Rather, we want to find the ratio of those with master degrees *and* higher salaries to those with master degrees -the intersection-but only as a fraction of those with the condition we stated; that is, only those with master degrees. Thus, the conditional probability will be the ratio of those with both master degrees and higher salaries as a proportion of those only with master degrees. From this it follows that:

Conditional probabilities have a *reduced sample space—we* care only about results based on what has already happened, nothing else!

The conditional probability of B, given A, denoted by P (B/A) is defined by the equation

$$P(B/A)=\frac{P(BandA)}{P(A)}=\frac{P(B\cap A)}{P(A)}=\frac{\text{Joint } probability}{\text{Pr}iorprobability} \quad if\,P(A)>0$$

From this formula, we can *now* derive a general formula for intersection probabilities. If we cross multiply, we see that P (A) and P (B) = P (B/A) x P (A), or the joint probability is simply the product *of* the conditional and prior probabilities.

Another *note of* caution-order matters with conditional probabilities. Order *does not* matter with unions or intersections:

$P (A\cup B) = P (B \cup A)$

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

$P (B/A) \neq P(A/B)$ ← order *of* events matters!

Special Case: Events where the condition (or prior) doesn't matter: What *if* the conditional probability—for example, the probability that you get a higher salary given that *you* have your master degree-were the same as just the probability *of* getting a higher salary? That is, what *if* P (B/A) = P (B)?

In this special case, the probability *of* getting a higher salary doesn't depend in any way on having a master degree. Thus, in this special case where the conditional probability is the same as the marginal probability, event B does not depend on event A (the prior). Said differently, having a master degree does *not* change the probability that *you* will have a higher salary upon graduation.

In this special case, events A and B are said *to* be ***independent***. But be careful—independence (or dependence) is *not* the same as lack *of* causation. If the events are dependent, this is *not to* say that having a master degree *caused you to* get a higher salary upon graduation, for example.

Dependence ≠ Causation

We can test for this property *of* independence in *two* (equivalent) ways. If events A and B are independent,

1. P(B/A) = P(B), or
2. P (A and B) = P (B) x P (A).

Either test can be used to check for independence. However, we cannot "see" independence in a Venn diagram.

1.8 INDEPENDENT EVENTS

Two events A and B of a sample space are said to be independent if and only if

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

The above rule can be extended to any number of events.

If A and B are independent, then

i. A′ and B are independent
ii. P(A/B)=P(A) and P(B/A)=P(B)
iii. A and B′ are independent
iv. A′ and B′ are independent

Now we can compute intersection probabilities.

INTERSECTIONS:

In general,

P (A and B) = P(A∩B) = P (B/A) x P (A).

P (B/A) = P (B), and P (A and B) = P (B) x P (A). If A and B are independent events You'll notice that we need to be aware of conditional probabilities to find the probability of intersections. Fortunately, these are often given in problems. Look for the words *if* or *given;* these are clues to conditional statements.

Example: Assume we work for a company and are tracking shopping behavior for one of our products. We have men and women as customers, and our product is packaged in two different colors. We'd like to compute some simple probabilities, and we'd also like to see if there is any association-or dependence-between the shopper's gender and the color of the product's packaging. If there is no dependence, then it shouldn't matter what color packaging we use. However, if there is dependence, then we may want to target different packaging to consumers based on gender. The table above summarizes some purchase results from a recent in-store survey.

	Men	Women	Totals
Blue Packaging	25	15	40
Pink Packaging	15	40	55
Totals	40	55	95

Here we see the following probabilities:	
Probability of selecting pink packaging	$=\frac{50}{85}$
Probability of selecting a male *or* blue package shopper	$=\frac{40+35-25}{85}$
Probability of selecting a male *and* pink package shopper	$=\frac{15}{85}$
Probability of selecting a male customer	$=\frac{40}{85}$
Probability of selecting a pink package shopper if male	$=\frac{15}{40}$
Probability of selecting female shopper if blue package	$=\frac{15}{35}$

Are sex and package color independent? Let's test them. Does P (Pink/Male) = P (Pink)? If so, these events are *independent.*

$$\frac{10}{30} \neq \frac{40}{65}$$

Because these probabilities are not the same, we conclude that sex and package color are *dependent events* (but not necessarily causal). That is, package color is not proportionally distributed between men and women. Instead, one sex (or both) may prefer one color over another, although we cannot directly conclude that package color, alone, caused this disproportionate selection. Note, too, that we do not need to test all possible combinations to assess independence. **If one event is dependent, then all events are dependent**.

Try not to confuse mutually exclusive events —a special case relevant to union probabilities-with independent events —a special case relevant to intersection probabilities. They are not the same! You can "see" mutual exclusivity in a Venn diagram. You cannot see independence.

1.9 PROBABILITY TREES

A useful tool for mapping out and computing probabilities is a *probability tree.*

Example: Your procurement organization has proposed new buying policies for office supplies, overnight mail, office furniture, and computers. The likelihood that a department will not adopt the policies is 30 percent. If the new policies are implemented, the likelihood of sustained cost savings for the departments that institute them is 80 percent. If the new policies are not adopted, the chance of sustained cost savings is only 40 percent.

We would like to determine the following:

- How likely is it that a department will adopt the new policies and save money?
- If the new policies are implemented, what's the likelihood that the department will save money?
- What's the likelihood that a department will save money (regardless of whether or not it implements these new purchasing policies)?
- Does policy adoption have anything to do with saving money?

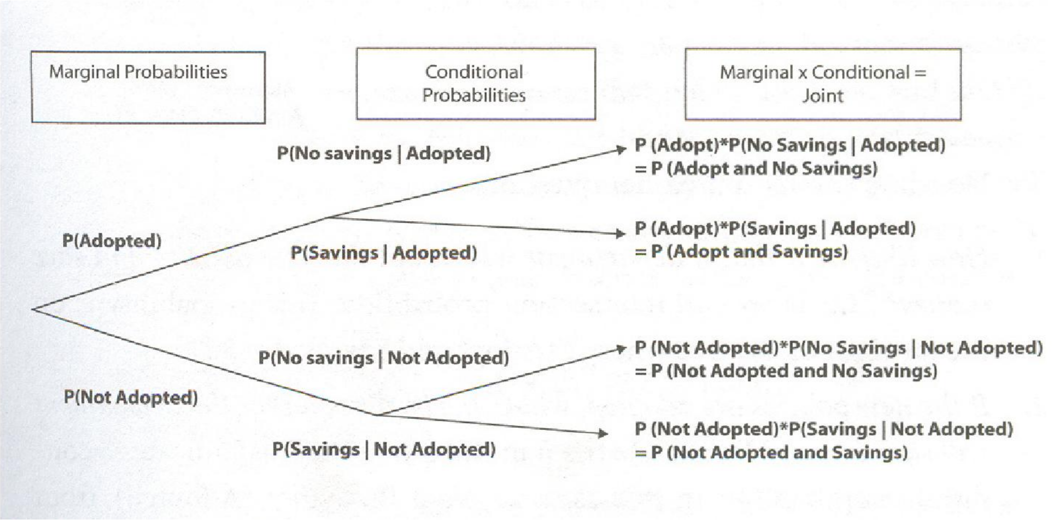
A probability tree will help us map out the events and probabilities so we can answer these questions. The key is to understand what is being asked in each

- Unless a probability is conditional, the denominator is always total sample space.
- Only with conditionals is the denominator smaller than total sample space.

question in probability terms and how to get the solutions from the tree.

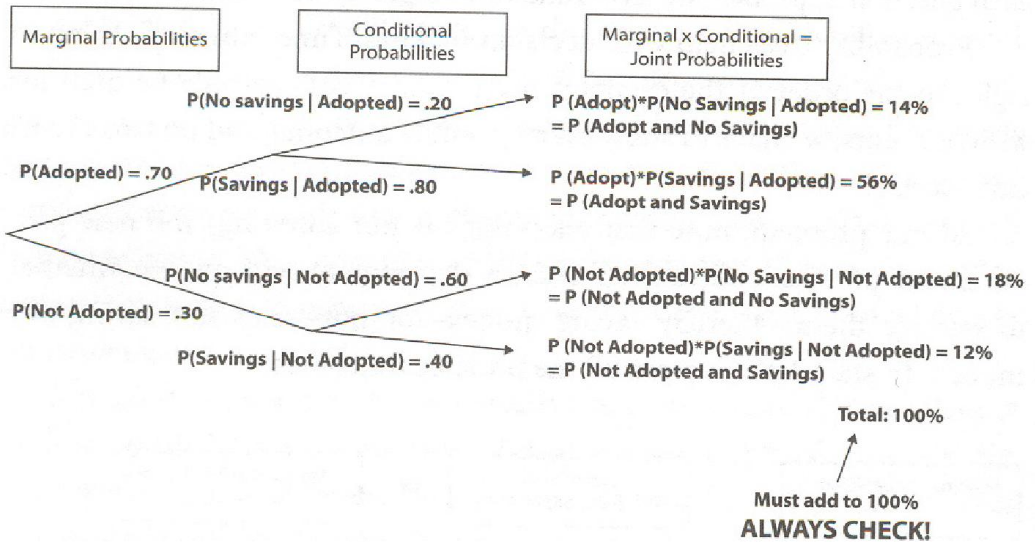
Probability trees help with decision making. Time moves from left to right, so the order of the events is important. Trees can only be used for MECE (mutually exclusive and collectively exhaustive) events; we have to show every possible outcome, and no two events can overlap.

In our problem, note that adopting (or not adopting) the new procurement policies comes first. Then a department will choose whether to enforce them-thereby saving money-or not-thus not saving any money. To start, let's map out all the possible outcomes.


Marginal Probabilities
Conditional Probabilities
Marginal x Conditional = Joint
P(Adopted)
P(Not Adopted)
P(No savings | Adopted)
P(Savings | Adopted)
P(No savings | Not Adopted)
P(Savings | Not Adopted)
P (Adopt)*P(No Savings | Adopted) = P (Adopt and No Savings)
P (Adopt)*P(Savings | Adopted) = P (Adopt and Savings)
P (Not Adopted)*P(No Savings | Not Adopted) = P (Not Adopted and No Savings)
P (Not Adopted)*P(Savings | Not Adopted) = P (Not Adopted and Savings)

We start on the left with two possible outcomes: either a department adopts these new polices, or it does not. These are the simple *marginal* probabilities, or *priors,* since these events happen first. Then, moving to the right, a department that has adopted these new policies either will or won't end up saving money. The same is true for departments that do not adopt these new procurement policies. Thus, we have *conditional probabilities.* Whether a department saves money may or may not be dependent on whether it adopts the new policies.

Finally, at the far right, we multiply the prior branches of the tree —a marginal times a conditional —to get all possible joint probabilities. Because the events are mutually exclusive, the probabilities at each node or "corner" in the tree where branches form to the right-add to one (100 percent). Also, because all possible outcomes are represented on the far right, all of these joint probabilities must also add to one (collectively exhaustive). Filling in the numbers from the tree, we see


Marginal Probabilities
Conditional Probabilities
Marginal x Conditional = Joint Probabilities
P(No savings | Adopted) = .20
P (Adopt)*P(No Savings | Adopted) = 14% = P (Adopt and No Savings)
P(Adopted) = .70
P(Savings | Adopted) = .80
P (Adopt)*P(Savings | Adopted) = 56% = P (Adopt and Savings)
P(No savings | Not Adopted) = .60
P (Not Adopted)*P(No Savings | Not Adopted) = 18% = P (Not Adopted and No Savings)
P(Not Adopted) = .30
P(Savings | Not Adopted) = .40
P (Not Adopted)*P(Savings | Not Adopted) = 12% = P (Not Adopted and Savings)
Total: 100%
Must add to 100%
ALWAYS CHECK!

Now let's answer our earlier questions:

1. *How likely is it that a department will adopt the new policies and save money?* This is an *and* intersection probability. This probability is on the far right of the tree where P(Adopt and Savings) = 56%.
2. *If the new policies are adopted, what's the likelihood that the department will save money?* Here, note the *'if' in* the question. This *indicates* a conditional probability. In this case, we want P (Savings/Adopted), from the middle of the tree, which is 80 percent.
3. What's *the likelihood that a department will save money (regardless of whether it implements the new purchasing policies)?* This one sounds a bit tricky. All we want to know is P (Savings); that is, the *unconditional* probability that a department will save money. Note that this probability is not on the tree. Thus, we have to be a bit clever. Take a look at the far right of the tree-where all possible outcomes are listed-and note that only two outcomes exist where a department saves money. First, a department could save money *and* adopt the new policies. Alternatively, a department could save money and *not* adopt the new policies. Adding these together-without worrying about any overlap because we have MECE events-will give us what we need. That is: P (Savings) = P (Adopted and Savings) or P(Not Adopted and Savings). This is a union probability, but we don't have to worry about overlap because a department cannot exist in both states (i.e. a department cannot save money and both adopt and not adopt the new policies). Thus, we simply add these probabilities together to get: 56% + 12% = 68%.

4. *Does policy adoption have anything to do with saving money?* Finally, we are asked to consider whether policy adoption and saving money are independent events. To test this, we need to perform one of the two independence tests mentioned earlier. For example, does P(Savings/Adopt) = P(Savings)? From the tree, we can determine that the left-hand side of this equation is 80 percent. And from the previous question, we can get the right-hand side: 68 percent. Because these two probabilities are unequal, we surmise that policy adoption and savings are dependent. However, this does not imply causation; just because a department adopts the new policies does not mean that department will save money. On the contrary, as the tree shows, there is a nonzero probability of adopting the new policies but not saving money (14 percent).

❖ SELF TEST ACTIVITY 1-6

1. A random sample of 200 employees is classified below according to sex and the level of education attained.

	Male	*Female*
Elementary	70	35
Secondary	40	25
College	20	10

If an employee is picked at random from this group, find the probability that

   a. the person is a male, given that the person has a secondary education;
   b. the person does not have a college degree, given that the person is a female.

Further Activities

2. In the senior year of a high school graduating class of 150 students, 50 studied mathematics, 70 studied psychology, 54 studied history, 22 studied both mathematics and history, 25 studied both mathematics and psychology, 7 studied history but neither mathematics nor psychology, 10 studied all three subjects, and 8 did not take any of the three. If a student is selected at random, find the probability that
   a. a person enrolled in psychology takes all three subjects;
   b. a person not taking psychology is taking both history and mathematics.
3. A pair of dice is thrown. If it is known that one die shows a 3, what is the probability that
   a. the other die shows a 5?
   b. the total of both dice is greater than 7?
4. The probability that an automobile being filled with gasoline will also need an oil change is 0.25; the probability that it needs a new oil filter is 0.40; and the probability that both the oil and filter need changing is 0.14.

   a. If the oil had to be changed, what is the probability that a new oil filter is needed?
   b. If a new oil filter is needed, what is the probability that the oil has to be changed?
5. The probability that a married man watches a certain television show is 0.4 and the probability that a married woman watches the show is 0.5. The probability that a man watches the show, given that his wife does, is 0.7. Find the probability that
   a. a married couple watches the show;
   b. a wife watches the show given that her husband does; (c) at least 1 person of a married couple will watch the show.
6. 11. The probability that a doctor correctly diagnoses a particular illness is 0.7. Given that the doctor makes an incorrect diagnosis, the probability that the patient e enters a law suit is 0.9. What is the probability that the doctor makes an incorrect diagnosis and the patient sues?
7. One bag contains 4 white balls and 3 black balls, and a second bag contains 3 white balls and 5 black balls. One ball is drawn at random from the second bag and is placed unseen in the first bag. What is the probability that a ball now drawn from the first bag is white?
8. A town has 2 fire engines operating independently. The probability that a specific fire engine is available when needed is 0.96.
   a. What is the probability that neither is available when needed?
   b. What is the probability that a fire engine is available when needed?
9. If the probability that you will be alive in 20 years is 0.85 and the probability that your brother will be alive in 20 years is 0.95, what is the probability that neither will be alive in 20 years?
10. A coin is biased so that a head is twice as likely to occur as a tail. If the coin is tossed 4 times, what is the probability of getting
    a. exactly 3 tails?
    b. at least 2 heads?

1.10 PROBABILITY DISTRIBUTIONS

1.10.1 INTRODUCTION

The generalizations associated with statistical inferences are subject to uncertainties, since we are dealing with only partial information obtained from a subset of the data of interest. To cope with these uncertainties, an understanding of probability theory is essential in order to provide a mathematical model that theoretically describes the behavior of the population associated with the statistical experiment. These theoretical models, which are very similar to relative frequency distributions, are called probability distributions.

At the end of the chapter, you will be able to:

- ❖ Describe what discrete and continuous sample spaces are

- ❖ Define discrete and continuous random variables together with their probability distributions
- ❖ Evaluate the expected values of random variables
- ❖ Explain the use of expected values in decisions

1.10.2 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

The term *statistical experiment* has been used to describe any process by which one or more chance measurements are obtained. Often, we are not interested in the details associated with each sample point but only in some numerical description of the outcome. For example, the sample space giving a detailed description of each possible outcome when one tosses a coin 3 times may be written

S = *{HHH, HHT, HTH, THH, HIT, THT, ITH, TTT}*.

If one is concerned only with the number of heads that fall, then a numerical value of 0, 1, 2, or 3 will be assigned to each sample point.

The numbers 0, 1, 2, and 3 are random quantities determined by the outcome of an experiment. They may be thought of as the values assumed by some random variable X, which in this case represents the number of heads when a coin is tossed 3 times.

RANDOM VARIABLE: A function whose value is a real number determined by each element in the sample space is called a *random variable.*

We shall use a capital letter, say X, to denote a random variable and its corresponding small letter, x in this case, for one of its values. In the coin tossing example above, we notice that the random variable X assumes the value 2 for all elements in the subset

E = *{HHT, HTH, THH}*

of the sample space S. That is, each possible value of X represents an event that is a subset of the sample space for the given experiment.

Example: Two balls are drawn in succession without replacement from an urn containing 4 red balls and 3 black balls. The possible outcomes and the values *y* of the random variable *Y*, where *Y* is the number of red balls, are:

Sample Space	*y*
RR	2
RB	1
BR	1
BB	0

Example: A hat check girl returns 3 hats at random to 3 customers who had previously checked them. If Samuel, Yoseph, and Bedasso in that order, receive one of the 3 hats, list the sample points for the, possible orders of returning the hats and find the values *m* of the random variable *M* that represents the number of correct matches.

Solution: If S, *J,* and *B* stand for Samuel's, Joseph's, and Bedasso's hats, respectively, then the possible arrangements in which the hats may be returned and the number of correct matches are

Sample Space	*m*
SYB	3
SBY	1
YSB	1
YBS	0
BSY	0
BYS	1

In each of the two preceding examples the sample space contains a finite number of elements. On the other hand, when a die is thrown until a 4 occurs, we obtain a sample space with an unending sequence of elements, namely,

S = {*F, NF, NNF, NNNF,* .. . },

where *F* and *N* represent, respectively, the occurrence and nonoccurrence of a 4. But even in this experiment, the number of elements can be equated to the number of whole numbers and in this sense can be counted.

DISCRETE SAMPLE SPACE: If a sample space contains a finite number of possibilities or an unending sequence with as many elements as there are whole numbers, it is called a *discrete sample space.*

The outcomes *of* some statistical experiments may be neither finite nor countable. Such is the case, for example, when one conducts an investigation measuring the distances that a certain make of automobile will travel over a prescribed test course on 5 liters of gasoline. Assuming distance to be a variable measured to any degree of accuracy then clearly we have an infinite number of possible distances in the sample space that cannot be equated to the number of whole numbers. Also, if one were to record the length of time for a chemical reaction to take place, once again the possible time intervals making up our sample space are infinite in number and uncountable. We observe now that all sample spaces need not be discrete.

Continuous Sample Space: If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a *continuous sample space.*

Random *variables* defined over discrete and continuous sample spaces are called, respectively, **discrete random variables** and **continuous random variables.** In most practical problems continuous random variables represent *measured* data, such as all possible heights, weights, temperatures, distances, or life periods, whereas discrete random variables represent *count* data, such as the number of defectives in a sample of n items or the number of highway fatalities per year in a given city. Note that the random variables *Y* and *M* of Examples 1 and 2 given above both represent count data, *Y* the number of red balls and *M* the number of correct hat matches.

1.10.3 Discrete Probability Distribution

A discrete random variable assumes each of its values with a certain probability. In the case of tossing a coin 3 times, the variable X, representing the number of heads, assumes the value 2 with probability $\frac{3}{8}$ since 3 of the 8 equally likely sample points result in 2 heads and 1 tail. By assuming equal probabilities for the simple events in Example 2, the probability that no man gets back his right hat, that is, the probability that M assumes the value 0 is $\frac{1}{3}$. The possible values m of M and their probabilities are given by:

m	0	1	3
P (M=m)	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Note that the values of *m* exhaust all possible cases, and hence the probabilities add to 1.

Frequently, it is convenient to represent all the probabilities of a random variable X by a formula. Such a formula would necessarily be a function of the numerical values x that we shall denote by *f(x), g(x), r(x),* and so forth. Hence we write *f(x)* = *P(X = x);* that is, *f(3)* = *P(X* = 3). The set of ordered pairs *(x, f(x))* is called the **probability function or probability distribution** of the discrete random variable X.

> ***DISCRETE PROBABILITY DISTRIBUTION:*** A table or a formula listing all possible values that a discrete random variable can take on, along with the associated probabilities, is called a *discrete probability distribution.*

Example. Find the probability distribution of the sum of the numbers when a pair of dice is tossed.

Solution. Let X be a random variable whose values x are the possible totals, i.e. X:the possible totals from the toss of the two dice. Then x =2, 3, 4, ..., ,12. Two dice can fall in (6)(6) = 36 ways, each with probability $\frac{1}{36}$. Then P(X = 5) = $\frac{4}{36}$, since a total of 4 can occur in only 4 ways. Consideration of the other cases leads to the following probability distribution:

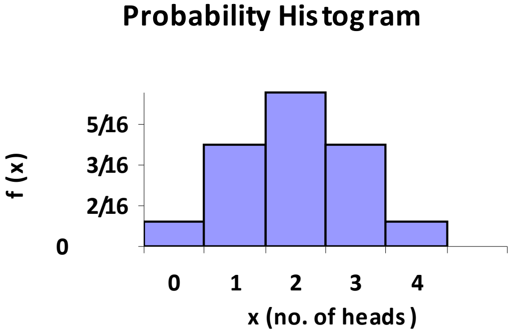
x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example 4. Find a formula for the probability distribution of the number of heads when a coin is tossed 4 times.

Solution: Since there are 2^4=16 points in the sample space representing equally likely outcomes, the denominator for all probabilities, and therefore for our function, will be 16. To obtain the number of ways of getting, say 2 heads, we need to consider the number of ways of partitioning 4 outcomes into 2 cells, with 2 heads assigned to one cell and a tail assigned to the other. This can be done in $\binom{4}{2}$= 6 ways. In

$$f(x) = P(X = x) = \frac{\binom{4}{x}}{16}, \quad x = 0, 1, 2, 3, or \ 4$$

general x heads and 4 – x tails can occur in $\binom{4}{x}$ ways, where x=0, 1 ,2 ,3, or 4. Thus the probability distribution f(x) = P(X = x) is:



It is often helpful to look at a probability distribution graphically in the form of a probability histogram, as demonstrated in the figure below for the above example. The rectangles are constructed so that their bases of equal width are centered at each value of x and their heights are equal to the corresponding probabilities given by f(x).

Since each rectangle in the above figure has a base of unit width, then P(X= 0), P(X = 1), P(X = 2), P(X = 3), and P(X = 4) are equal to the areas of the rectangles centered at x = 0, x = 1, x = 2, x = 3, and x = 4, respectively. Even if the bases were not of unit width, we could adjust the heights of the rectangles to give areas that would still equal the probabilities of X assuming any of its values x. This concept of using areas to represent probabilities is necessary for our consideration of the probability distributions of continuous random variables.

Certain probability distributions are applicable to more than one physical situation. The probability distribution of the previous example above, for example, also applies to the random variable Y, where Y is the number of red cards that occurs when 4 cards are drawn at random from a deck in succession with each card replaced and the deck shuffled before the next drawing. Special discrete probability distributions that can be applied to many different experimental situations will be considered in the next chapter.

1.10.3.1 Continuous Probability Distributions

A continuous random variable has a probability of zero of assuming exactly any of its values. Consequently, its probability distribution cannot be given in tabular form. At first this may seem startling, but it becomes more plausible when we consider a particular example. Let us discuss a random variable whose values are the heights of all people over 21 years of age. Between any two values, say 163.5 and 164.5 centimeters, or even 163.99 and 164.01 centimeters, there are an infinite number of heights, of which only one is 164 centimeters. The probability of selecting a person at random who is exactly 164 centimeters tall, and not one of the infinitely large set of heights so close to 164 centimeters that you cannot humanly measure the difference, is extremely remote. Thus we assign a probability of zero to the event. This is not the case, however, if we talk about the probability of selecting a person who is at least 163 centimeters but not more than 165 centimeters tall. Now we are dealing with an interval rather than a point value of our random variable.

We shall concern ourselves with computing probabilities for various intervals of continuous random variables such as P (a < X < b), P (Y > d), and so forth. Note that when X is continuous

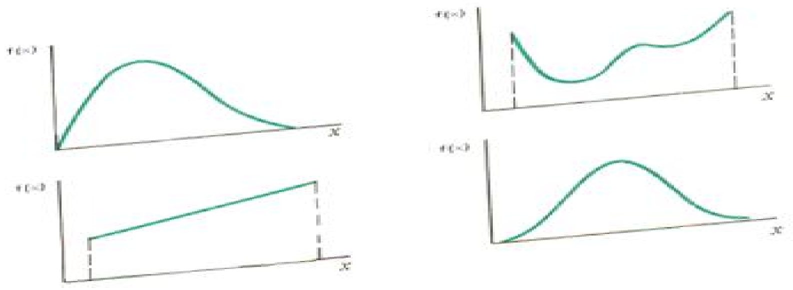
$$P (a < X \leq b) = P (a < X < b) + P(X = b) = P (a < X < b).$$

That is, it does not matter whether we include an end point of the interval or not. This is not true when X is discrete.

Although the probability distribution of a continuous random variable cannot be presented in tabular form, it does have a formula. Such a formula would necessarily be a function of the numerical values of the continuous variable X and as such could be graphed as a continuous curve. The probability function portrayed by this curve is usually called a probability density function (pdf), or simply a density function.

PROBABILITY DENSITY FUNCTION: The function with values *f(x)* is called a *probability density function* for the continuous random variable X if the total area under its curve and above the *x* axis is equal to 1 and if the area under the curve between any two ordinates *x* = *a* and *x* = *b* gives the probability that X lies between *a* and *b.*

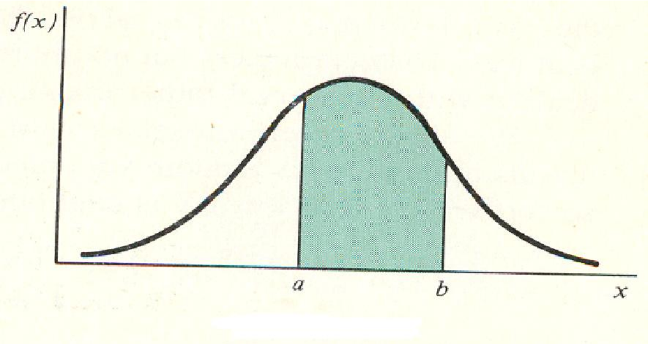
Most density functions that have practical applications in the analysis of statistical data are continuous for all values of X and their graphs may take any of several forms, some of which are shown in figures below. Because areas will be used to represent probabilities and probabilities are positive numerical values, the density function must lie entirely above the x axis.



Typical Density Functions

A probability density function is formed so that the area under its curve bounded by the x-axis is equal to 1. If the density function is represented by the curve below:

Then the probability that X assumes a value between a and b, i.e. P (a<X<b), is equal to the shaded area under the curve between the ordinates at x=a and x=b.



$P(a < X < b)$.

Example: A continuous random variable X that can assume values between x = 2 and x = 4 has a density function given by

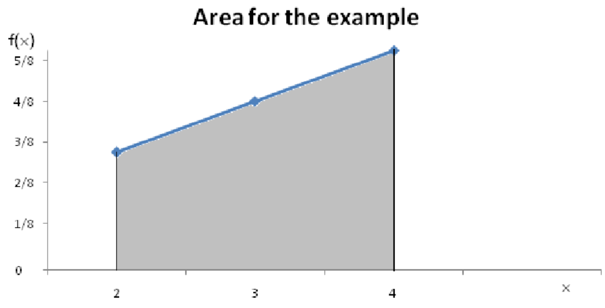
$f(x)=\frac{x+1}{8}$.

i. Show that *P(*2 < X < 4) = 1.
ii. Find *P(X* < 3).
iii. Find *P(*2.6 < X < 3.5).

Solution:

Since the shaded region in the figure is a trapezoid, the area is found by summing the heights of the parallel sides, multiplying by the length of the base, and dividing by 2. That is,

$$area=\frac{(sum\,of\ the\,parallel\,sides)\,X\,base}{2}=\frac{[f(2)+f(4)](2)}{2}$$

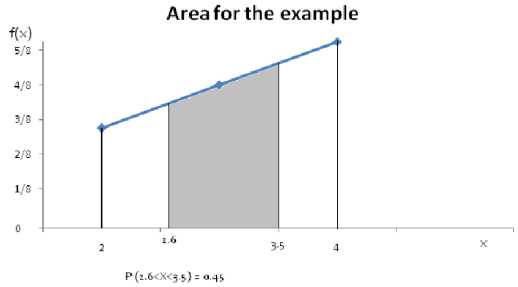


Then substituting the values we get

P (2<X<4) = $P(2<X<4)=\dfrac{(\frac{3}{8}+\frac{5}{8})(2)}{2}=1.$

b) Similarly $P(X<3)=\dfrac{(\frac{3}{8}+\frac{5}{8})(1)}{2}=\dfrac{1}{2}=0.5.$

c) as before $P(2.6<X<3.5)=\dfrac{(\frac{3.6}{8}+\frac{4.5}{8})(0.9)}{2}=\dfrac{0.9}{2}=0.45.$



1.10.3.2 Mean of a Random Variable

If two coins are tossed 20 times and X is the number of heads that occur per toss, then the values of X can be 0, 1, and 2. Suppose that the experiment yields no heads, 1 head, and 2 heads, a total of 6, 8, and 6 times, respectively. The average number of heads per toss of the 2 coins is then

$$\overline{X}=\frac{(0)(6)+(1)(7)+(2)(7)}{20}=1.05$$

This is an average value and is not necessarily a possible outcome for the experiment. For instance, a salesman's average monthly income is not likely to be equal to any of his monthly paychecks.

Let us now restructure our computation of x so as to have the following equivalent form:

$$\overline{X}=0(\tfrac{6}{20})+1(\tfrac{7}{20})+2(\tfrac{7}{20})$$

The numbers $\frac{6}{20},\frac{7}{20},\frac{7}{20}$ are the fractions of the total tosses resulting in 0, I, and 2 heads, respectively. These fractions are also the **relative frequencies** for the different values of X in our sample. In effect, then, one can calculate the mean of a set of data by knowing the distinct values that occur and their relative frequencies, without any knowledge of the total number of measurements or observations in our set of data. Therefore, if $\frac{6}{20}$ of the tosses result in no heads, $\frac{7}{20}$ of the tosses result in I head, and $\frac{7}{20}$

of the tosses result in 2 heads, the mean number of heads per toss would be 1.05 no matter whether the total number of tosses was 16, 1000, or even 10,000.

Let us now use this method of relative frequencies in calculating the population mean for the number of heads per toss of two coins that we might expect in the long run. We shall refer to this population mean as the mean of the random variable X or the mean of its distribution and write it as μ_x or simply μ. It is also common among statisticians to refer to this mean as the mathematical expectation or the expected value of the random variable X and denote it as **E(X)**. In other words, the mean or expected value of a random variable can be interpreted as the mean of the population or distribution whose observations are all the values of X that are generated by repeating the experiment over and over again indefinitely.

Assuming that fair coins were tossed, the sample space for our experiment is given by

$$S = \{HH, HT, TH, IT\}.$$

Since the four sample points are all equally likely, it follows that

$P(X = 0)= P(TT) = \frac{1}{4}$,

$P(X=1)=P(TH)+P(HT)=\frac{2}{4}$, and

$P(X = 2) = P(HH) = \frac{1}{4}$,

where a typical element, say *TH,* indicates that the first toss resulted in a tail followed by a head on the second toss. Now, these probabilities are just the relative frequencies for the given events in the long run. Therefore,

$\mu_x = E(X) = (O)\ (\frac{1}{4}) + (I)\ (\frac{2}{4}) + (2)\ (\frac{1}{4}) = 1.$

This means that a person who tosses 2 coins over and over again will, on the average, get 1 head per toss.

The method described above for calculating the expected number of heads per toss of 2 coins suggests that the mean or expected value of any discrete random variable may be obtained by multiplying each of the values x_1, x_2, ..., x_n of the random variable X by its corresponding probability $f(X_1)$, $f(X_2)$, . . . , $f(x_n)$ and summing the products. This is true, however, only if the random variable is discrete.

Mean of a Random Variable. Let X be a discrete random variable with the probability distribution,

x	x_1	x_2	. . .	x_n
f(x)=P(X=x)	$f(x_1)$	$f(x_2)$	. . .	$f(x_n)$

The mean or expected value of the random variable X is

$$\mu = E(X) = \sum_{x=1}^{n} x f(x)$$

Example: Find the expected value of X, where X represents the outcome when a die is tossed.

Solution: Each of the numbers l, 2, 3, 4, 5, and 6 occurs with probability $\frac{1}{6}$. Therefore,

$E(X) = (\frac{1}{6})(1) + (2)(\frac{1}{6}) + . . . + (6)(\frac{1}{6}) = 3.5.$

This means that a person will, on the average, roll 3.5.

Example: Find the expected number of boys on a committee of 3 selected at random *from* 4 boys and 3 girls.

Solution: Let Y represent the number of boys on the committee. The probability distribution of Y is given by

$$f(x) = \frac{\binom{4}{y}\binom{3}{3-y}}{\binom{7}{3}}, \quad \text{for } x = 0, 1, 2, 3.$$

A few simple calculations give $f(0) = \frac{1}{35}$, $f(1) = \frac{12}{35}$, $f(2) = \frac{18}{35}$, and $f(3) = \frac{4}{35}$.

Therefore,

$E(X) = \underline{(0)(\frac{1}{35})} + (1)(\frac{12}{35}) + (2)(\frac{18}{35}) + \underline{(3)}\ (\frac{4}{35}) = 1.7.$

Thus if a committee of 3 is selected at random over and over again *from* 4 boys and 3 girls, it would contain on the average 1.7 boys.

Example: In a gambling game a man is paid Birr15 if he gets all heads or all tails when 3 coins are tossed, and he pays out Birr13 if either I or 2 heads show. What is his expected gain?

Solution: The sample space *for* the possible outcomes when 3 coins are tossed simultaneously, or equivalently if a coin is tossed three times, is

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

One can argue that each of these possibilities is equally likely and occurs with probability equal to $\frac{1}{8}$. An alternative approach would be to apply the multiplicative rule of probability for independent events to each element of S. For example,

$$P(HHT) = P(H)P(H)P(T)$$
$$= (½)(½)(½) = \tfrac{1}{8}.$$

Let *Y* = the amount the gambler can win.

The values of $Y = \begin{cases} \text{Birr15}, & \text{if } E_1 = \{HHH, TTT\} \\ \text{Birr13}, & \text{if } E2 = \{HHT, HTH, THH, HTT, THT, TTH\}\,\{HHH, TTT\} \end{cases}$

Therefore, *E(Y)* = (15)($\frac{2}{8}$) + (-13)($\frac{6}{8}$) = -6.

In this game the gambler will, on the average, lose Birr 6 per toss of the 3 coins.

Remark! A game is considered "fair" if the gambler will, on the average, come out even. Therefore, an expected gain of zero defines a fair game.

1.10.3.3 THE VARIANCE OF A RANDOM VARIABLE

Example: Calculate the variance of the random variable X, in the previous example, where X is the number of boys on a committee of 3 selected at random from 4 boys and 3 girls.

Variance of a Random Variable: Let X be a discrete random variable whose probability distribution is f(x), for x=x_1, x_2, . . ., x_n

The variance of X, denoted by σ^2 is

$$\sigma^2 = E(X-\mu)^2 = \sum_{x=1}^{n} (x-\mu)^2 f(x)$$

Solution: Previously it has been shown that μ= E(X) = $\frac{12}{7}$.

Therefore,

$$\sigma^2 = E(X-\mu)^2 = \sum_{x=0}^{3} (x-\mu)^2 f(x)$$

$$= (0-\tfrac{12}{7})^2 \tfrac{1}{35} + (1-\tfrac{12}{7})^2 \tfrac{12}{35} + (2-\tfrac{12}{7})^2 \tfrac{18}{35} + (3-\tfrac{12}{7})^2 \tfrac{4}{35}$$

$$= \tfrac{24}{49}$$

An alternative and preferred formula for finding σ^2, which often simplifies the calculations,

Computing Formula for σ^2: The variance of the random variable X is given by $\sigma^2 = E(X)^2 - \mu^2$

Example: Use the computing formula to recalculate the variance of the random variable in the above example.

Solution: Once again referring to the example above, we find that

$$\sum_{x=0}^{3} x^2 f(x) = (0)(\tfrac{1}{35}) + (1)(\tfrac{12}{35}) + (2)(\tfrac{118}{35}) + (3)(\tfrac{4}{35})$$

$$= \tfrac{24}{7}$$

and μ = $\frac{12}{7}$, it follows that

$$\sigma^2 = \tfrac{24}{7} - \tfrac{12}{7} 2 = \tfrac{24}{49}$$

❖ Self Activity Test 1-7

1. By investing in a particular stock, a person can make a profit in 1 year of Birr5000 with probability 0.4 or take a loss of Birr1000 with probability 0.7. What is this person's expected gain?

Further Activities

2. A shipment of 7 television sets contains 2 defectives. A hotel makes a random purchase of 3 of the sets. If X is the number of defective sets purchased by the hotel, find the mean of X.
3. The probability distribution of the discrete random variable X is

$$f(x)=\binom{4}{x}\left(\frac{1}{5}\right)^{x}\left(\frac{3}{5}\right)^{4-x}, \quad \text{for } x=0,1,2,3,4.$$

   Find the mean and variance of X.

4. In a gambling game a man is paid Birr3 if she draws a jack or king and Birr5 if she draws a queen or ace from an ordinary deck of 52 playing cards. If she draws any other card, he loses. How much should he pay to play if the game is fair?

Summary of Basic Probability Rules

- Probabilities of any outcome(s) must be between 0 and 1 (inclusive).
- The sum of probabilities of all possible outcomes must equal 1(exactly!).

Compound Events

- Union (or) probabilities: The sum of the marginals less the overlap (or intersection) probability
- Special case for unions: Mutually exclusive events-no overlap! Probability simplifies to just the sum of the marginals.
- Intersection *(and)* probabilities: The product of the conditional probability and the prior probability
- Special case for intersections: Independent events. Probability simplifies because conditional probability is simply the marginal. Joint probability here is just the product of the marginals.

Probability Trees

- Easy way to map out MECE-mutually exclusive and collectively exhaustive-events. First come the marginals, then the conditionals, and finally the joints or intersections (which are just the marginals times the conditionals).
- Don't forget that all the joint probabilities (on the far right) must add to 1.

Expected value

- Expected value is the "mean payoff." Expected value is to probability and payoff as mean (or average value) is to a dataset. It's the central tendency-or expected-payoff, which is derived by looking at the weighted average of all possible payoffs that all outcomes produce.
- A payoff table can help us determine expected values. Be sure to keep the payoffs in a specific point of view; otherwise, signs can be erroneous (+/-).

1.10.4 COMMON DISCRETE PROBABILITY DISTRIBUTIONS

Among several discrete probability distributions that can be discussed, we shall introduce the two most commonly used distributions; namely, binomial and Poisson probability distributions.

1.10.4.1 BINOMIAL PROBABILITY DISTRIBUTION

Many probability problems are concerned with experiments in which an event is repeated many times. Binomial distribution is one of the simplest and most frequently used discrete probability distribution and is very useful in many practical situations involving *either/or* types of events. It has certain distinct properties which are enumerated as follows:

An experiment often consists of repeated trials, each with two possible outcomes, which may be labeled **success** or **failure.** This is true in;

- Hitting a target 2 times out of 6.
- Finding one defective item in a sample of 25 items.
- The flipping of a coin 6 times, where each trial may result in a head or a tail. We may choose to define either outcome as a success.
- Drawing 5 cards in succession from an ordinary deck and each trial is labeled "success" or "failure," depending on whether the card is red or black if each card is replaced and the deck shuffled before the next drawing, then the two experiments described above have similar properties, in that repeated trials are independent and the probability of a success remains constant, p=½, from trial to trial. Experiments of this type are known as **binomial experiments.**
- Observe in the card-drawing example that the probabilities of a success for the repeated trials change if the cards are not replaced. That is, the probability of selecting a red card on the first draw is ½, but on the second draw it is a conditional probability having a value of $\frac{26}{51}$ *or* $\frac{25}{51}$, depending on the color that occurred on the first draw. This then would no longer be considered a binomial experiment.

In each case, some outcome is designated a success, and any other outcome is considered a failure. Thus if the probability of a success in a single trial is p, the probability of failure will be q =1-p.

A binomial experiment is one that possesses the following properties:

i. The experiment consists of *n* repeated trials.
ii. Each trial results in an outcome that may be classified as a success or a failure.
iii. The probability of a success, denoted by *p,* remains constant from trial to trial.
iv. The repeated trials are independent.

Binomial Distribution: If a binomial trial can result in a success with probability *p* and a failure with probability = 1 - *p,* then the probability distribution of the binomial random variable *X,* the number of successes in *n* independent trials is,

$$f(x;n,p)=P(X=x)=\left\{\binom{n}{x}p^{x}q^{n-x},\ for\, x=0,1,2\cdots,n\right.$$

Notation: X ~bin (n, p) is used to denote that X has a binomial distribution with parameters n and p.

Remark: $\sum P(X=x)=P\{(X=0)+P(X=1)+...+P(X=n)=1$

Constants of the Binomial Distribution

- If X ~ bin (n, p), then

I. The expected value of X, denoted by E(X) or μ_X, is given by

E(X) =μ_X=np

II. The variance of X, denoted by var(X) or σ_x^2, is given by

Var(X) = σ_x^2 = npq where q = 1-p

III. The standard deviation of X, σ_x^2 = $\sqrt{npq}$

Example: If a fair coin is tossed four times.

What is the probability that exactly 2 heads will show up?

a) What is the probability that at least 2 heads will show up?
b) What is the probability that at most 1head will show up?
c) What is the expected number of heads?
d) What is the standard deviation of the number of heads?

Solution

Given: n= 5 independent tosses

p= 0.5= probability that a head shows up in a single trial

q= 1-p = 0.5

Let X= number of heads in 4 tosses of a coin

$\therefore$X~ bin (n=4, p=0.5)

$$P(X=x)=\binom{n}{x}p^{x}q^{n-x}=\binom{4}{x}(0.5)^{x}(0.5)^{4-x}$$

$$P(X=2)=\binom{4}{2}(0.5)^{2}(0.5)^{2}=0.375$$

a) P(X≥2) = 1–P(X≤1) = 1–P(X=0) –P(X=1)

$$=1-\binom{4}{0}(0.5)^{0}(0.5)^{4}-\binom{4}{1}(0.5)^{1}(0.5)^{3}$$

$$=0.6875$$

b) P(X≦1) = P(X=1) + P(X=0)= 0.3125

c) The expected number heads is E(X) = np= 4x0.5=2

d) The standard deviation of heads $\sigma_{x}=\sqrt{npq}=\sqrt{4\times0.5\times0.5}=\sqrt{1}=1$

Example 2: A test check revealed that milk in 25 per cent of the bottles is unfit for consumption. The salesman at a retail outlet offers 5 bottles for sale on demand.

Find the probability that milk will be unfit for consumption:

a) exactly in 2 bottles,

b) at least in 2 bottles, and

c) at the most in 2 bottles

d) Find the average number of bottles with bad milk

Solution:

Since p=$\frac{1}{4}$ and n=5, the required probabilities are as follows.

a) P(X=2) = $\binom{5}{2}\frac{1}{4}^{2}\frac{3}{4}^{3}=0.2636$

b) P(X≥2) = $1-\left[\binom{5}{0}\frac{1}{4}^{0}\frac{3}{4}^{5}+\binom{5}{1}\frac{1}{4}^{1}\frac{3}{4}^{4}\right]=0.3672$

c) P(X≤2) = $\binom{5}{0}\frac{1}{4}^{0}\frac{3}{4}^{5}+\binom{5}{1}\frac{1}{4}^{1}\frac{3}{4}^{4}+\binom{5}{2}\frac{1}{4}^{2}\frac{3}{4}^{3}=0.8964$

d) The average number of bottles with bad milk is

E(X) = np = 5x$\frac{1}{4}$ = 1.25 in each set of 5 bottles.

❖ SELF TEST ACTIVITY 1-8

1. The Abay car account research team estimated that 20% of all car owners in Addis Ababa prefer Abay. Suppose a random sample of 5 car owners is chosen. Find the probability that;
   a. exactly 4 car owners prefer Abay car
   b. none of the car owners prefer Abay car

2. It is estimated that 4000 of the 10,000 voting residents of a town are against a new sales tax. If 15 eligible voters are selected at random and asked their opinion, what is the probability that at least 2 favor the new tax?

   Solution: P (success) = $p=\frac{4000}{10000}=\frac{2}{5}$, and $q=1-p=1-\frac{2}{5}=\frac{3}{5}$, n = 15

   Let X= the number of eligible voters that favor the new sales tax.

   Then X ~ bin (15, $\frac{2}{5}$), then the required answer for the question is,

$$P(X\geq 2)=\sum_{x=2}^{15}\binom{15}{x}\left(\frac{2}{5}\right)^{x}\left(\frac{3^{15-x}}{5}\right)$$
$$=1-P(X\leq 1)$$
$$=1-P(X=1)-P(X=0)$$
$$=1-\binom{15}{1}\left(\frac{2}{5}\right)^{1}\left(\frac{3}{5}\right)^{14}-\binom{15}{0}\left(\frac{2}{5}\right)^{0}\left(\frac{3}{5}\right)^{15}$$
$$=1-0.004702-0.00047$$
$$=0.005172$$

1.10.5 POISSON DISTRIBUTION

Experiments yielding numerical values of a random variable X, the number of outcomes occurring during a given time interval or in a specified region, are often called **Poisson experiments.**

The given time interval could be of any length, a minute, a day, a week, a month, or even a year. Hence a Poisson experiment might generate observations for the Poisson random variable X representing;

- The number of telephone calls per hour received by an office,
- The number of days school is closed due to snow during the winter

- The number of postponed games due to rain during a football season.
- Number of perfect pins produced by a machine per hour.
- Number of cars washed at a place in every half hour, say between 8:00 am and 8:30am.
- Number of accidents in a certain period in a certain traffic intersection per day.
- The number of heavy trucks arriving at a railway station in an hour
- The number of defective items in a manufacturing process per day.

The specified region could be a line segment, an area, a volume, or perhaps a piece of material. In this case X might represent:

- Number of field mice per acre of land.
- Number of bacteria in a given culture.
- Number of typing errors per page.
- Number of mistakes in a book per page.
- Number of car accidents in a certain traffic intersection per hour

A Poisson experiment is one that possesses the following properties:

1. The number of outcomes occurring in one time interval or specified region is independent of the number that occur in any other disjoint time interval or region of space.
2. The probability that a single outcome will occur during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of outcomes occurring outside this time interval or region.
3. The probability that more than one outcome will occur in such a short time interval or fall in such a small region is negligible.

The number X of outcomes occurring in a Poisson experiment is called a Poisson random variable and its probability distribution is called the Poisson distribution.

Poisson Distribution: The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval or specified region, is

$$f(x;\lambda)=P(X=x)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!}, & x=0,1,\,2,\cdots\\ 0, & Otherwise\end{cases},$$

where μ is the average number of outcomes occurring in the given time interval or specified region and *e* = 2.71828 . . . .

Remark:

- If one is interested to consider successes for a time length of k units while λ is available for a unit length of time, then it is necessary to take a new average which is equal to kλ.
- If X has a **Poisson** distribution with parameter λ, then
  - E (X)= λ_x= λ
  - Var(X)=σ_x^2 = λ and $\sigma_x = \sqrt{\lambda}$
- $\sum_{x=0}^{\infty} P(X = x) = 1$

Example:

The average number of field mice per acre in a 5-acre wheat field is estimated to be 10. Find the probability that a given acre contains more than 15 mice.

Solution:

Let X = the number of field mice per acre.

Then using the Poisson Table given in the appendix, we have

$$P(X > 15) = 1 - \mathrm{P}(\mathrm{X} \le 15)$$

$$= 1 - \sum_{x=0}^{15} \frac{\mathrm{e}^{-10} 10^x}{x!}$$

$$= 1 - 0.9513$$

$$= 0.0487$$

Example:

A certain 500-page book has, on the average, 3 printing errors per page.

a. What is the probability that the book will have at most 3 misprints on the following page?
b. What is the average number of misprints and the standard deviation of the number of misprints per page?

Solution:

Let X = the number of misprints per page

Given: λ =3 misprints per page

X has the **Poisson** distribution with parameter λ =3 misprints/page given by

$P(X=x)=\frac{e^{-\mu}\lambda^{x}}{x!}=\frac{e^{-3}3^{x}}{x!}$, then

a. P(X≤3) =P(X=0)+P(X=1)+P(X=2)+P(X=3)

$$=\frac{e^{-3}3^{0}}{0!}+\frac{e^{-3}3^{1}}{1!}+\frac{e^{-3}3^{2}}{2!}+\frac{e^{-3}3^{3}}{3!}$$

$$=e^{-3}+3e^{-3}+\frac{9}{2}e^{-3}+\frac{27}{6}e^{-3}$$

$$=13e^{-3}=0.647$$

b. E (X)= λ_x= λ =3 misprints per page and $\sigma_x=\sqrt{\lambda}=\sqrt{3}=1.73$

❖ Self Activity Test 1-9

1. If a receptionist's phone rings an average of 4 times an hour, find the probability of
   a. no calls in a randomly selected hour,
   b. exactly 2 calls per hour,
   c. 3 or more calls in a duration of 2 hours.
2. The number of students entering a library of a certain university college is on the average 90 per hour. Find the probability of 1 to 3 students entering the library in a minute.

Further Activities

1. A scientist inoculates several mice, one at a time, with a disease germ until he finds 2 that have contracted the disease. If the possibility of contracting the disease is t what is the probability that 8 mice are required?
2. Three people toss a coin and the odd man pays for the coffee. If the coins all turn up the same, they are tossed again. Find the probability that fewer than 4 tosses are needed.
3. On the average a certain intersection results in 3 traffic accidents per month. What is the probability that in any given month at this intersection
   a. exactly 5 accidents will occur?
   b. less than 3 accidents will occur?
   c. at least 2 accidents will occur?

4. A secretary makes 2 errors per page on the average. What is the probability that on the next page she makes
    a. 4 or more errors?
    b. no errors?
5. A certain area of the world is, on the average, hit by 10 hurricanes a year. Find the probability that in a given year this area will be hit by
    a. fewer than 4 hurricanes;
    b. anywhere from 6 to 8 hurricanes.
6. The average number of oil tankers arriving each day at a certain port city is known to be 10. The facilities at the port can handle at most 15 tankers per day. What is the probability that the port is unable to handle all the tankers that arrive
    a. on a given day?
    b. on one of the next 3 days?

1.11 CONTINUOUS PROBABILITY DISTRIBUTIONS

The probability distribution of a continuous random variable can be visualized as a smooth form of the relative frequency histogram based on large number observations.

The mathematical function denoted by f(x) whose graph produces the limiting form of the relative frequency histogram is called the probability density function (pdf) of the continuous random variable X.

The probability density function f(x) describes the distribution of probability for a continuous random variable. It has the properties:

a. The total area under the density curve is 1. i.e.. $\int_{-\infty}^{\infty} f(x)dx = 1$

b. f (x)$\geq$0

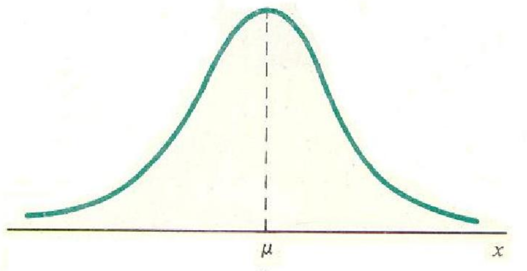
With a continuous random variable X, the probability that X=x is always 0. That is if X is continuous, then P(X=x)=0. Because of this we have P(a<X<b)=P(a$\leq$X<b)=P(a<X$\leq$b)= P(a$\leq$x$\leq$b) = area under the density curve between a and b = $\int_{a}^{b} f(x)dx$.

Areas for important distributions have been extensively tabulated, and all we need to do is to consult these tables.

1.11.1 NORMAL PROBABILITY DISTRIBUTION

Continuous random variables and their associated density functions arise whenever our experimental data are defined over a continuous sample space. Therefore, whenever we measure time intervals, weights, heights, volumes, and so forth, our underlying population is described by a continuous distribution. Just as there are several special discrete probability distributions there are also numerous types of continuous distributions whose graphs may display varying amounts of skewness or in some cases may be perfectly symmetric. Among these, by far the most important is the continuous distribution whose graph is a symmetric bell-shaped curve extending indefinitely in both directions. It is this distribution that provided a basis upon which much of the theory of statistical inference has been developed.

The most important continuous probability distribution in the entire field of statistics is the normal distribution. Its graph, called the normal curve, is the bell-shaped curve of **Figure** that describes so many sets of data that occur in nature, industry, and research. In 1733, DeMoivre derived the mathematical equation of the normal curve. The normal distribution is often referred to as the Gaussian distribution in honor of Gauss (1777-1855), who also derived its equation from a study of errors in repeated measurements of the same quantity.



normal curve

A continuous random variable X having the bell-shaped distribution of Figure above is called a normal random variable. The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.

Definition:*Normal Curve.* If X is a normal random variable with mean μ and variance$\sigma^2 > o$, then the equation of the *normal curve* is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})2} \text{ , for } -\infty < X < \infty \text{ , } -\infty < \mu < \infty \text{ , } \sigma > o$$

Where π = 3.14159… and e = 2.71828…

Properties of the normal curve

i. The *mode,* which is the point on the horizontal axis where the curve is a maximum, occurs at x = μ.

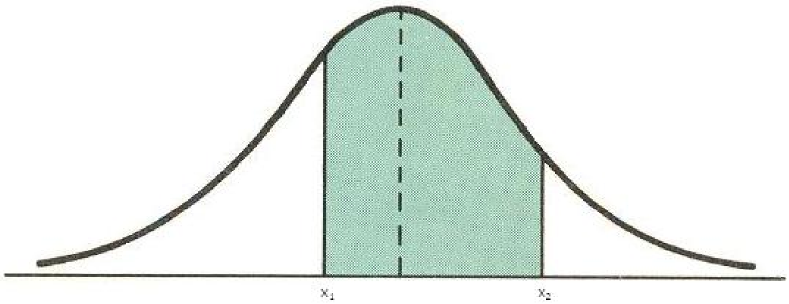
ii. The curve is symmetric about a vertical axis through the mean μ.

iii. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.

iv. The total area under the curve and above the horizontal axis is equal to 1.

Areas under the normal curve

The curve of any continuous probability distribution or density function is constructed so that the area under the curve bounded by the two ordinates $x = x_1$ and $x = x_2$ equals the probability that the random variable X assumes a value between $x = x_1$ and $x = x_2$ Thus, for the normal curve in **Figure,** $P(x_1 < X < x_2)$ is represented by the area of the shaded region below:



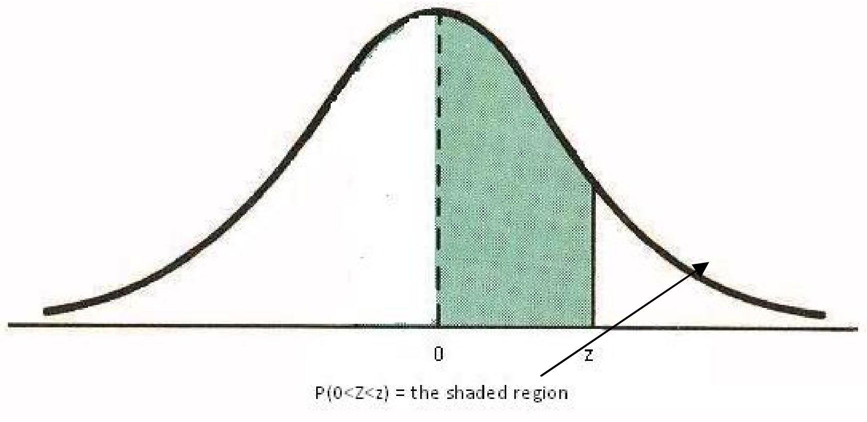
$P(x_1 < X < x_2)$ = shaded region

Standard normal distribution: The distribution of a normal random variable $Z = \frac{x-\mu}{\sigma}$ with mean zero and standard deviation equal to 1 is called a *standard normal distribution*

The standard normal distribution remains unaffected by changes in μ and σ^2 for the normal distribution. The total area under the standard normal curve is 1 and P (Z<0)=P (Z>0)=0.5. Areas under the standard normal curve have been tabulated. These areas (probabilities) are given in the appendix. We shall now learn how to read the table.

Reading the standard normal table

P (0<Z<z) is read from the table corresponding to z. The integer part of Z is given in the first column and the remaining decimal part in the first row. The point joining the row of the integer part of Z and column of the remaining decimal part of Z gives P (0<Z<z).



P(0<Z<z) = the shaded region

The following properties can be derived from the symmetry of the density about 0.

a. P $(0<Z<a)=$ P $(-a<Z<0)$ for some $a>0$

b. Since P $(Z>0)=0.5$, P $(Z>a)=0.5-$P $(0<Z<a)$ for some $a>0$

c. P $(Z<b)=0.5-$ P $(b<Z<0)$

$=0.5-$ P $(0<Z<|b|)$ for some $b<0$

d. P $(a<Z<b)=$P $(0<Z<b)-$P$(0<Z<a)$ for $0<a<b$

e. P$(c<Z<b)=$P $(0<Z<b)+$P$(0<Z<|c|)$ for $c<0$, $b>0$

f. P $(d<Z<c)=$ P $(|c|<Z<|d|)=$ P $(0<Z<|d|)-$ P $(0<Z<|c|)$ for $d<c<0$

g. P $(Z<b)=0.5+$P $(0<Z<b)$ for $b>0$

h) P$(Z>a)=0.5+$P$(0<Z<|a|)$ for $a<0$

Example

1 P $(0<Z<1.96)=0.4750$

2. P $(-1.96<Z<0)=$P $(0<Z<1.9)=0.4750$

3. P $(Z>1.15)=0.5-$ P $(0<Z<1.15)=0.5-0.3749=0.1251$

4. P $(Z<-2.12)$ $=0.5-$P $(-2.12<Z<0)$

$=0.5-$ P $(0<Z<2.12)$

$=0.5-0.4830$

$=0.017$

5. P $(1.59<Z<2.28)$ $=$P $(0<Z<2.28)-$ P $(0<Z<1.59)$

$=0.4887-0.4441=0.0446$

6. P $(-1.05<Z<0.24)$ $=$P $(-1.05<Z<0)+$P $(0<Z<0.24)$

$=$P $(0<Z<1.05)+$P $(0<Z<0.24)$

=0.3531+0.0948= 0.4479

7. P (-2.0<Z<-1.0)=P (1.0<Z<2.0)

=P (0<Z<2.0) – P (0<Z<1.0)

=0.4772 – 0.3413

=0.1359

8. P (Z<1.28)=0.5+P (0<Z<1.28)=0.5+0.3997=0.8997

9. P (Z>-1.25)=0.5+P (-1.25<Z<0)

=0.5+P (0<Z<1.25)

=0.5+0.3944

=0.8944

NOTATION

The notation X ~ N(μ,σ^2) is used to denote that X has a normal distribution with mean μ and variance σ^2.

COMPUTING PROBABILITY USING STANDARD NORMAL TABLE

If X has a normal distribution with mean μ and variance σ^2, i.e. if X~ N (μ, σ^2), then

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

The standard normal table providing areas under the standard normal curve is used to evaluate the right hand side expression given above.

__**Example:**__

1. The heights of 1000 students are normally distributed with a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. How many of these students would you expect to have heights
   (a) less than 160.0 centimeters?
   (b) greater than or equal to 188.0 centimeters?
   (c) between 171.5 and 182.0 centimeters?

__**Solution**__

Let X = the height of a randomly selected student

Given: X~N (174.5, 6.9^2)

(a) 1, 000P (X<160.0)=1, 000P$\left(Z<\frac{160.0-174.5}{6.9}\right)$

$$
\begin{aligned}
&= 1000P(Z < -0.2.10) \\
&= 1000[0.5 - P(0 < Z < 2.10)] \\
&= 1000x0.0.0179 \\
&= 17.9 \approx 18
\end{aligned}
$$

This means approximately 18 students have heights less than 160.0 centimeters.

(b) 1000P (X≥188.0)=1000P (Z> $\frac{188.0-174.5}{6.9}$ =1000P (Z>1.96)

$$
\begin{aligned}
&= 1000[0.5 - \mathrm{P}(0 < \mathrm{Z} < 1.96)] \\
&= 1000[0.5 - 0.4750] \\
&= 1000\mathrm{x}0.0250 \\
&= 25
\end{aligned}
$$

This means approximately 25 students have heights greater than or equal to 188.0 centimeters.

(c)

$$
\begin{aligned}
1000\mathrm{P}(171.5 < \mathrm{X} < 182.0) &= 1000\mathrm{P}\left(\frac{171.5-174.5}{6.9} < \mathrm{Z} < \frac{182.0-174.5}{6.9}\right) \\
&= 1000\mathrm{P}(-0.43 < \mathrm{Z} < 1.09) \\
&= 1000\mathrm{x}[\mathrm{P}(0 < \mathrm{Z} < 0.43) + \mathrm{P}(0 < \mathrm{Z} < 1.09)] \\
&= 1000\mathrm{x}[0.1664 + 0.3621] \\
&= 528.5
\end{aligned}
$$

This means approximately 529 students have heights greater than 171.5 centimeters and less than 182.0 centimeters.

❖ SELF TEST ACTIVITY 1.10

1. Training was designed to upgrade the technical skills of factory workers. A study of past participants indicates that the length of time spent on the training is normally distributed with mean 500 hours and standard deviation of 100 hours.

   What is the probability that a randomly selected worker will need

   a) more than 500 hours?

   b) between 600 and 750 hours?

   c) between 450 and 550 hours?

   d) between 350 and 450 hours?

Further Activities

1. A baseball player's batting average is 0.250. What is the probability that he gets exactly 1 hit in his next 5 times at bat?
2. If we define the random variable X to be equal to the number of heads that occur when a balanced coin is flipped once, find the probability distribution of X. What two well-known distributions describe the values of X?
3. One-fourth of the female freshmen entering Addis Ababa University are out-of-Addis Ababa students. If the students are assigned at random to the dormitories, 3 to a room, what is the probability that in one room at most 2 of the 3 roommates are out-of Addis Ababa students?
4. A multiple-choice quiz has 15 questions, each with 4 possible answers of which only 1 is the correct answer. What is the probability that sheer guesswork yields from 5 to 10 correct answers?
5. The probability that a patient recovers from a delicate heart operation is 0.9. What is the probability that exactly 5 of the next 7 patients having this operation survive?
6. A study conducted revealed that approximately 70% believe "tranquilizers don't really cure anything; they just cover up the real trouble." According to this study, what is the probability that at least 3 of the next 5 people selected at random will be of the opinion that tranquilizers don't really cure the problem but just cover it up?
7. A survey of the residents in a city showed that 20% preferred a mobile phone over any other telephone available. What is the probability that more than one-half of the next 20 telephones in this city will be mobile phones?
8. 15. If 64 coins are tossed a large number of times, how many heads can we expect on the average per toss?
9. A nationwide survey of 12,000 students by the Addis Ababa University shows that almost 75% disapprove of smoking according to a research report. If 20 of these students are selected at random and asked their opinions, what is the probability that more than 9 but less than or equal 14 disapprove of smoking?
10. On the average a certain intersection results in 3 traffic accidents per month. What is the probability that in any given month at this intersection
    (a) exactly 5 accidents will occur?
    (b) less than 3 accidents will occur?
    (c) at least 2 accidents will occur?
11. A secretary makes 2 errors per page on the average. What is the probability that
    (a) on the next page she makes she makes 4 or more errors?
    (b) on the next two pages she makes no errors?
12. A certain area of the world is, on the average, hit by 5 hurricanes a year. Find the probability that in a given year this area will be hit by
    (a) fewer than 4 hurricanes;
    (b) anywhere from 6 to 8 hurricanes.

13. Given a normal distribution with μ= 200 and $\sigma^2 = 100$, find
    a. the area below 214;
    b. the area above 179;
    c. the area between 188 and 206;
    d. the x value that has 80% of the area below it.
14. Given the normally distributed variable X with mean 18 and standard deviation 2.5, find
    a. $P(X < 15)$;
    b. $P(17 < X < 21)$;
    c. the value of k such that $P(X < k) = 0.2578$;
    d. the value of k such that $P(X > k) = 0.1539$.
15. . A soft-drink machine is regulated so that it discharges an average of 200 milliliters per cup. If the amount of drink is normally distributed with a standard deviation equal to 15 milliliters,
    a. what fraction of the cups will contain more than 224 milliliters?
    b. what is the probability that a cup contains between 191 and 209 milliliters?
    c. how many cups will likely overflow if 230-milliliter cups are used for the next 1000 drinks?
    d. below what value do we get the smallest 25% of the drinks

2 SAMPLING AND SAMPLING DISTRIBUTIONS

2.1. Sampling Theory

A sample is a tool to infer something about a population. We begin by discussing methods of selecting a sample from a population. Next, we construct a distribution of the samples to understand how the sample means tend to cluster around the population means and that the shape of this distribution tends to follow the normal distribution.

2.1.1 The Need for Sampling

In many cases sampling is the only way to determine something about the population. Some of the major reasons why sampling is necessary are:

1. The destructive nature of certain tests

   If wine tasters at a quality control drink all the wine to evaluate the vintage content, they will consume the entire wine produced by Awash Wine Factory and none would be available for sale.

2. The physical impossibility of checking all items in the population.

   The population of fish, birds, snakes, mosquitoes, and the like are large and are constantly moving, being born, and dying. So taking population census is impossible.

3. The cost of studying all the items in a population is often prohibitive.

   Television executives of ETV want to know the proportion of television viewers who watch the "Fegegta Talk show". Since more than 20 million people in the country may be watching ETV programs on a given evening, determining the proportions that are watching "Fegegta Talk show" program is prohibitively expensive.

4. The adequacy of sample results

   Even if funds were available (for the case of ETV above), it is doubtful the additional accuracy of a 100 percent sample that is studying the entire population is essential in most problems. As you can see from the above example, the sample results are more than enough to know about the population under consideration.

5. To contact the whole population would often be time consuming

   A candidate for a national office may wish to determine her chances for election. A sample poll using the regular staff and field interviews of a professional polling firm would take only one or two days. By using the same staff and interviewers and working seven days a week, it would take nearly 200 years to contact all the voting population.

2.1.2 Sampling Methods

In general, there are two types of samples: a probability sample and non probability sample.

1. Probability Sample

   A probability samples is a sample selected in such a way that each item or person in the population has a known (nonzero) likelihood of being included in the sample. If probability sampling is done, each item in the population has a chance of being chosen.

2. Non probability Sampling

   If non probability methods are used, not all items or people have a chance of being included in the sample. In such instances the results may be biased, meaning that the sample results may not be representative of the population. Judgmental sampling and convenience sampling are two non probability methods. For example a panel is formed to solicit opinions on a newly developed cat food and there are 2000 cat owners selected for the sample. Selection of the members is based on the judgment of the person conducting the research, and the sample results may therefore not be representative of the entire population of cat owners (since not all cat owners have a chance of being chosen).

The difference between non probability and probability sampling is that non probability sampling does not involve ***random*** selection but probability sampling does. Does that mean that non probability samples aren't representative of the population? Not necessarily. But it does mean that non probability samples cannot depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With non probability samples, we may or may not represent the population well, and it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over non probabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling.

The statistical procedures used in this material are based on probability sampling. Therefore, only the methods of probability sampling will be discussed in the following section.

There is no one “best” method of selecting a probability sample from a population of interest. A method used to select a sample of invoices in a file drawer, might not be the most appropriate method for choosing a national sample of voters. However, all probability sampling methods have a similar goal, namely, to allow chance to determine the items or persons to be included in the sample.

Types of probability sampling

Generally there are four probability samplings. Each of these are discussed as follows:

1. Simple random sampling

The most widely used type of sampling is simple random sample. In simple random sampling, a sample selected so that each item or person in the population has the same chance of being included.

Example: Suppose a population consists of 845 employees of Sheraton Addis hotel. A sample of 52 employees is to be selected from that population. One way of ensuring that every employee in the population has a chance of being chosen is to first write the name of each employee on a small slip of paper and deposit all of the slips in a box. After they have been mixed, the first selection is made by drawing a slip out of the box without looking at it. This process is repeated until the sample size of 52.

2. Systematic random sample

The items or individuals of the population are arranged in some way alphabetically in a file drawer by date received or by some other method. A random starting point is selected, and then every Kth member of the population is selected for the sample. In a systematic sample the first item is chosen at random.

3. Stratified Random Sample

A population is divided in to subgroups, called strata, and a sample is selected from each stratum. After the population has been divided in to strata, either a proportional or a non proportional sample can be selected. As the name implies, a proportional sampling procedure requires that the number of items in each stratum be in the same proportion as in the population. In a non-proportional stratified sample, the number of items chosen in each stratum is disproportionate to the respective numbers in the population. We then weight the sample results according to the stratum's proportion of the total population. Regardless of whether a proportional or a non proportional sampling procedure is used every item or person in the population has a chance of being selected for the sample.

4. **Cluster sampling**

Other common type of sampling is cluster sampling. It is often employed to reduce the cost of sampling a population scattered over a large geographic area. Suppose you want to determine the views of industrialists in a country about Federal and regional government on environmental protection policies. Selecting a simple random of industrialists in the country and personally contacting each one would be time consuming and very expensive. Instead, you could employ cluster sampling by subdividing the country in to small units either regions or zones. These are often called primary units. Suppose you divided the country in to 14 regions, then selected at random four regions -2, 4, 7, and 12 and concentrated your efforts in these primary units you could take a random sample of the industrialists in each of these regions and interview them. (Note that this is a combination of cluster sampling and simple random sampling).

5. **Multi-stage sampling**

The four methods we have covered so far: simple, stratified, systematic and cluster are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

2.2. Sampling Distributions

A sampling distribution is a probability distribution for the possible values of a sample statistic, such as a sample mean.

NOTE: The normal probability distribution is used to determine probabilities for the normally distributed individual measurements, given the mean and the standard deviation. Symbolically, the variable is the measurement X, with the population mean μ and population standard deviation δ. In contrast to such distributions of individual measurements, a sampling distribution is a probability distribution for the possible values of a sample statistic.

A. Sampling Distribution of the Mean

The sampling distribution of the mean is the probability distributions of the means, $\overline{X}$ of all simple random samples of a given sample size n that can be drawn from the population.

NB: the sampling distribution of the mean is not the sample distribution, which is the distribution of the measured values of X in one random sample. Rather, the sampling distribution of the mean is the probability distribution for $\overline{X}$, the sample mean.

For any given sample size n taken from a population with mean μ and standard deviation δ, the value of the sample mean $\overline{X}$ would vary from sample to sample if several random samples were obtained from the population. This variability serves as the basis for sampling distribution.

The sampling distribution of the mean is described by two parameters: the expected value ($\overline{X}$) = $\overline{\overline{X}}$, or mean of the sampling distribution of the mean, and the standard deviation of the mean $\delta_{\bar{x}}$, the standard error of the mean.

Properties of the Sampling Distribution of Means

1. The mean of the sampling distribution of the means is equal to the population mean. μ = $\mu_{\overline{X}} = \overline{\overline{X}}$.
2. the standard deviation of the sampling distribution of the means (standard error) is equal to the population standard deviation divided by the square root of the sample size: $\delta_{\bar{x}}$ = δ/√n. This hold true if and only of n<0.05N and N is very large. If N is finite and n≥ 0.05N, $\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$.
   The expression $\sqrt{\frac{N-n}{N-1}}$ is called finite population correction factor/finite population multiplier. In the calculation of the standard error of the mean, if the population standard deviation δ is unknown, the standard error of the mean $\delta_{\bar{x}}$, can be estimated by using the sample standard error of the mean $S_{\overline{X}}$ which is calculated as follows: $S_{\overline{X}} = S/\sqrt{n}\ or\ S_{\overline{X}} = \frac{S}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$.
3. The sampling distribution of means is approximately normal for sufficiently large sample sizes (n≥ 30).

Example:

A population consists of the following ages: 10, 20, 30, 40, and 50. A random sample of three is to be selected from this population and mean computed. Develop the sampling distribution of the mean.

Solution:

The number of simple random samples of size n that can be drawn without replacement from a population of size N is NCn. With N= 5 and n = 3, 5C3 = 10 samples can be drawn from the population as:

Sampled items	Sample means ($\overline{X}$)
10, 20, 30	20.00
10, 20, 40	23.33
10, 20, 50	26.67
10, 30, 40	26.67
10, 30, 50	30.00
10, 40, 50	33.33
20, 30, 40	30.00
20, 30, 50	33.33
20, 40, 50	36.67
30, 40, 50	40.00

A systematic organization of the above figures gives the following:

Sample mean ($\overline{X}$)	Frequency	Prob. (relative freq.) of $\overline{X}$
20.00	1	0.1
23.33	1	0.1
26.67	2	0.2
30.00	2	0.2
33.33	2	0.2
36.67	1	0.1
40.00	1	0.1
	10.00	1.00

- Columns 1 and 2 show frequency distribution of sample means.
- Columns 1 and 3 show sampling distribution of the mean.

$\mu = \frac{\sum X}{N} = \frac{\sum \overline{x}}{n} = 30,$ regardless of the sample size $\mu = \overline{\overline{X}}$.

$$\sigma = \sqrt{\frac{\sum \left(X_i - \overline{X}\right)^2}{N}} = \sqrt{\frac{1000}{5}} = 14.142$$

.

$$\sigma_{\overline{X}} = \frac{\delta}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} = \frac{14.142}{\sqrt{3}} * \sqrt{\frac{5-3}{5-1}} = 5.774$$

$$= \sqrt{\frac{\sum\left(\overline{X_i} - \overline{\overline{X}}\right)^2}{N}} = \sqrt{\frac{333.4}{10}} = 5.774$$

- Since averaging reduces variability $\delta_{\bar{x}}$ < δ except the cases where δ = 0 and n = 1.

Central Limit Theorem and the Sampling Distribution of the Mean

The Central Limit Theorem (CLT) states that:

1. If the population is normally distributed and population standard deviation is known, the distribution of sample means is normal regardless of the sample size.
2. If the population from which samples are taken is not normal, the distribution of sample means will be approximately normal if the sample size (n) is sufficiently large (n ≥ 30). The larger the sample size is used, the closer the sampling distribution is to the normal curve.

The relationship between the shape of the population distribution and the shape of the sampling distribution of the mean is called the Central Limit Theorem.

The significance of the Central Limit Theorem is that it permits us to use sample statistics to make inference about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample. It also permits us to use the normal distribution (curve for analyzing distributions whose shape is unknown. It creates the potential for applying the normal distribution to many problems when the sample is sufficiently large.

Example:

1. The distribution of annual earnings of all bank tellers with five years of experience is skewed negatively. This distribution has a mean of Birr 15,000 and a standard deviation of Birr 2000. If we draw a random sample of 30 tellers, what is the probability that their earnings will average more than Birr 15,750 annually?

Solution:

Steps:

1. Calculate µ and $\delta_{\bar{x}}$

µ = Birr 15,000

$\delta_{\bar{x}}$ = δ/√n= 2000/√30 = Birr 365.15

2. Calculate Z for $\overline{X}$

$$Z_{\overline{X}} = \frac{\overline{X} - \overline{\overline{X}}}{\delta_{\overline{X}}} = \frac{\overline{X} - \mu}{\delta_{\overline{X}}}$$

$$Z_{15,750} = \frac{15,750 - 15,000}{365} = +2.05$$

3. Find the area covered by the interval

P ($\overline{X}$ > 15,750) = P (Z > +2.05)

= 0.5 - P (0 to +2.05)

= 0.5 – 0.47892

= 0.02018

4. Interpret the results

There is a 2.02% chance that the average earning being more than Birr 15, 750 annually in a group of 30 tellers.

2. Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability of randomly selecting 49 different shopping hours, counting the shoppers, and having the sample mean fall between 441 and 446 shoppers, inclusive?

Solution:

1. Calculate μ and $\delta_{\bar{x}}$

μ = 448 shoppers

$\delta_{\bar{x}}$ = δ/√n= 21/√49 = 3

2. Calculate Z for $\overline{X}$

$$Z_{\overline{X}} = \frac{\overline{X} - \overline{\overline{X}}}{\delta_{\overline{X}}} = \frac{\overline{X} - \mu}{\delta_{\overline{X}}}$$

$$Z_{441} = \frac{441 - 448}{3} = -2.33 \qquad Z_{446} = \frac{446 - 448}{3} = -0.67$$

3. Find the area covered by the interval

P (441 ≤ $\overline{X}$ ≤ 15,750) = P (-2.33 ≤ Z≤ -0.67)

= P (0 to -2.33) - P (0 to - 0.67)

= 0.49010 – 0.24857

= 0.24153

4. Interpret the results

There is a 24.153% chance of randomly selecting 49 hourly periods for which the sample mean falls between 441 and 446 shoppers.

3. A production company's 350 hourly employees average 37.6 year of age, with a standard deviation of 8.3 years. If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

   Solution:

   1. Calculate μ and $\delta_{\bar{x}}$

      μ = 37.6 years   n/N= 45/350 > 5%...... FPCF is needed

      $$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} = \delta_{\bar{x}} = \frac{8.3}{\sqrt{45}} * \sqrt{\frac{350-45}{350-1}} = 1.16$$

   2. Calculate Z for $\overline{X}$

      $$Z_{\overline{X}} = \frac{\overline{X} - \overline{\overline{X}}}{\delta_{\overline{X}}} = \frac{\overline{X} - \mu}{\delta_{\overline{X}}}$$

      $$Z_{40} = \frac{40-37.6}{1.16} = +2.07$$

   3. Find the area covered by the interval

      P ($\overline{X}$ < 40) = P (Z < +2.07)

      = 0.5 + P (0 to +2.07)

      = 0.5 + 0.48077

      = 0.98077

   4. Interpret the results

   There is a 98.08% chance of randomly selecting 45 hourly employees and their mean age be less than 40 years.

4. Suppose that a random sample size of 36 is being drawn from a population with a mean of 278. If 86% of the time the sample mean is less than 280, what is the population standard deviation?

   Solution:

   μ = 278   n = 36   P ($\overline{X}$ < 280) = 0.86   δ =?

   (Z/P=0.36) = +1.08

   $$Z_{\overline{X}} = \frac{\overline{X} - \mu}{\delta_{\overline{X}}} \qquad \delta_{\overline{X}} = \frac{\delta}{\sqrt{n}}$$

   $$Z_{280} = \frac{280-278}{\delta_{\overline{X}}} \qquad 1.85 = \frac{\delta}{\sqrt{36}}$$

   $$+1.08 = \frac{280-278}{\delta_{\overline{X}}} \qquad 1.85 = \frac{\delta}{6}$$

$+1.08 = \frac{2}{\delta_{\overline{X}}}$ $\delta = 6 * 1.85$

$= 11.1$

$\delta_{\overline{X}} = \frac{2}{1.08} = 1.85$

5. A teacher gives a test to a class containing several hundred students. It is known that the standard deviation of the scores is about 12 points. A random sample of 36 scores is obtained.

a) What is the probability that the sample mean will differ from the population mean by more than 6 points?

b) What is the probability that the sample mean will be within 6 points of the population mean?

Solution:

a)

n = 36 δ =12 $\delta_{\overline{X}} = \frac{\delta}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2$

P ($\overline{X}$ > μ +6) + P ($\overline{X}$ < μ - 6) =?

$Z_{\mu+6} = \frac{\mu + 6 - \mu}{2} = +3$ $Z_{\mu-6} = \frac{\mu - 6 - \mu}{2} = -3$

P ($\overline{X}$ > μ +6) + P (Z> μ - 6) = P (Z > 3) + P (Z < - 3)

= [0.5 – P (0 to +3)] + [0.5 – P (0 to -3)]

= (0.5 – 0.49865) + (0.5 – 0.49865)

= 0.00135(2) = 0.00270

b)

n = 36 δ =12 $\delta_{\overline{X}} = \frac{\delta}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2$

P (μ - 6≤ $\overline{X}$ ≤ μ + 6) = P (- 3≤ Z ≤ 3)

= P (0 to 3)*2

= 0.49865*2

= 0.9973

If the population standard deviation is 12, in a random sample of 36 scores there is a 99.73% chance of getting a sample mean score to lie within 6 points of the population mean.

3. STATISTICAL ESTIMATION

Introduction

After developing sampling distributions of different population parameters, it is very important to estimate where these population parameters might be located. That means the population parameters must be estimated relative the samples in hand.

As its name suggests, the objective of estimation is to determine the approximate value of a population parameter on the bases of a sample statistic. An estimator of a population parameter is a random variable that is a function of the sample data. An estimate is the calculation of a specific value of this random variable.

The sampling distribution of the mean shows how far sample means could be from a known population mean. Similarly, the sampling distribution of the proportion shows how far sample proportions could be from a known population proportion. In estimation, our aim is to determine how far an unknown population mean could be from the mean of a simple random sample selected from that population; or how far an unknown population proportion could be from a sample proportion. Those are the concerns of statistical inference, in which a statement about an unknown population parameter is derived from information contained in a random sample selected from the population.

Basic concepts:

- Estimation: is the process of using statistics as estimates of parameters. It is any procedure where sample information is used to estimate/ predict the numerical value of some population measure (called a parameter).

- Estimator- refers to any sample statistic that is used to estimate a population parameter. E.g. $\overline{x}$ for $\mu,$ $\overline{p}$ for p.
- Estimate- is a specific numerical value of our estimator. E.g. $\overline{x}$ =9, 2, 5

$$\begin{cases} \overline{x}, \overline{p}, s^2, s \text{ .... Estimators} \\ \\ \mu, p, \sigma^2, \sigma \text{ .... items being estimated} \end{cases}$$

1, 0.5, 9, 3 …………………... Estimates

Four Important Properties of Estimators

A number of different estimators are possible for the same population parameter, but some estimators are better than others. To understand how, we need to look at four important properties of estimators: unbiasedness, efficiency, consistency, and sufficiency.

Unbiasedness: An estimator exhibits unbiasedness when the mean of the sampling estimator is equal to the population parameter: E (θ) = Θ.

In general, unbiasedness is a desirable property for an estimator. The sample mean is an unbiased estimator of the population mean. Similarly, the sample variance is an unbiased point estimator the population variance because the mean of the sampling distribution of the sample variance is equal to the population variance. And the sample proportion is an unbiased estimator of the population proportion. However, because standard deviation is a nonlinear function of variance, the sample standard deviation is not an unbiased estimator of population standard deviation. The bias of a point estimator is: Bias = E (θ) = Θ. If there are a number of unbiased estimators to choose from, there are three other criteria that could be used to select an estimator.

Efficiency: Efficiency is another standard that can be used to evaluate estimators. Efficiency refers to the size of the standard error of the statistics. The most efficient estimator is the one with the smallest variance. Thus, if there are two estimators for Θ with var (θ_1) and var (θ_2), then the first estimator θ_1 is said to be more efficient than the second estimator θ_2, if var(θ_1) < var (θ_2) although $E(\theta_1) = E(\theta_2) = \Theta$.

Consistency: A third property of estimators, consistency, is related to their behavior as the sample gets large. A statistic is a consistent estimator of a population parameter if, as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter.

An unbiased estimator is a consistent estimator if the variance approaches 0 as n increases. For example, the sample mean is an unbiased and a consistent estimator of population mean. Although the sample standard deviation is not an unbiased estimator of population standard deviation, it is a consistent estimator of population standard deviation.

Sufficiency: The last property of a good estimator is sufficiency. A sufficient statistic is an estimator that utilizes all the information a sample contains about the parameter to be estimated. For example, the sample mean is a sufficient estimator of the population mean. This means that no other estimator of the population mean from the same sample data, such as the sample median, can add any further information about the parameter (population mean) that is being estimated.

Types of Estimates:

We can make two types of estimates about a population: a point estimate and an interval estimate.

A point estimate: - is a single number that is used to estimate an unknown population parameter. It is a single value that is measured from a sample and used as an estimate of the corresponding population parameter.

The most important point estimates (given that they are single values) are:

- Sample mean $(\bar{x})$for population mean (μ);
- Sample proportion $(\bar{p})$for population proportion(p);
- Sample variance (s^2) for population variance (σ^2) and
- Sample standard deviation $(s)$for population standard deviation (σ)

An interval estimate - is a range of values used to estimate a population parameter. It describes the range of values with in which a parameter might lie. Stated differently, an interval estimate is a range of values with in which the analyst can declare with some confidence that the population parameter will fall.

Example:

Suppose we have the sample 10,20,30,40 and 50 selected randomly from a population whose mean μ is unknown.

The sample mean, $\bar{x}$, $\frac{\sum xi}{n}$ = $\frac{10+20+30+40+50}{5}=30$ is a point estimate of μ.

On the other hand, if we state that the mean, μ, is between $\bar{x} \pm 10$, the range of values from 20 (30-10) to 40 (30+10) is an interval estimate.

1. Interval Estimation

Point estimators of population parameters, while useful, do not convey as much information as interval estimators. Point estimation produces a single value as an estimate of the unknown population parameter. The estimate may or may not be close to the parameter value; in other words, the estimate may be incorrect. An interval estimate, on the other hand, is a range of values that conveys the fact that estimation is an uncertain process. The standard error of the point estimator is used in creating a range of values; thus a measure of variability is incorporated into interval estimation. Further, a measure of confidence in the interval estimator is provided; consequently, interval estimates are also called confidence intervals. For these reasons, interval estimators are considered more desirable than point estimators.

Interval estimation for population mean, μ

As a result of the Central Limit Theorem (discussed in Chapter III) the following z formula for sample means can be used when sample sizes are large, regardless of the shape of the population distribution or for smaller sizes if the population is normally distributed.

$$Z = \frac{\overline{X} - \mu}{\sigma/n}$$

Rearranging the formula:

$$\mu = \overline{X} - Z\,\sigma/n$$

Because the sample mean can be greater than or less than the population mean, z can be positive or negative. Thus, the preceding expression takes the form:

$$\mu = \overline{X} \pm Z\,\sigma/n$$

The value of the population mean, μ, lies somewhere within this range. Rewriting this expression yields the confidence interval for population mean:

$$\overline{X} - Z\,\sigma/n \le \mu \le \overline{X} + Z\,\sigma/n$$

The confidence interval for population mean is affected by:

1. The population distribution, i.e., whether the population is normally distributed or not
2. The standard deviation, i.e., whether σ is known or not.
3. The sample size, i.e., whether the sample size, n, is large or not.

Confidence internal estimate of μ - Normal population, σ known

A confidence interval estimate for μ is an interval estimate together with a statement of how confident we are that the interval estimate is correct.

When the population distribution is normal and at the same time σ is known, we can estimate μ (regardless of the sample size) using the following formula[1].

$$\boxed{\mu = \overline{X} \pm Z_{\alpha/2} \sigma / n}$$

Where: $\overline{X}$ = sample mean; Z = value from the standard normal table reflecting confidence level; σ = population standard deviation; n = sample size; a = the proportion of incorrect statements (a = 1 – C); and μ = unknown population mean

From the above formula we can learn that an interval estimate is constructed by adding and subtracting the error term to and from the point estimate. That is, the point estimate is found at the center of the confidence interval.

To find the interval estimate of population mean, μ we have the following steps.

1. Compute the standard error of the mean $(\sigma_{\bar{x}})$
2. Compute $\alpha/2$ from the confidence coefficient.
3. Find the Z value for the $\alpha/2$ from the table
4. Construct the confidence interval
5. Interpret the results

[1] This formula works also for problems which involve large sample size (n>30) even though the population is not normally distributed. And if n>.05N, finite population correction factor may be used.

Examples:

1. The vice president of operations for Ethiopian Tele Communication Corporation (ETC) is in the process of developing a strategic management plan. He believes that the ability to estimate the length of the average phone call on the system is important. He takes a random sample of 60 calls from the company records and finds that the mean sample length for a call is 4.26 minutes. Past history for these types of calls has shown that the population standard deviation for call length is about 1.1 minutes. Assuming that the population is normally distributed and he wants to have a 95% confidence, help him in estimating the population mean.

Solution:

n= 60 calls $\overline{X}$ = 4.26 minutes σ = 1.1 minutes C= 0.95

i. $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.1}{\sqrt{60}} = 0.142$

ii. a = 1 – C = 1- 0.95 = 0.05

$\alpha/2$ = 0.05/2 = 0.025

iii. $Z_{\alpha/2} = Z_{0.025} = 1.96$

iv. $\mu = \overline{X} \pm Z_{\alpha/2}\, \sigma / \sqrt{n}$

= 4.26 ± 1.96(0.142)

= 4.26 ± 0.28

➡ 3.98 ≤ μ ≤ 4.54

The vice-president of ETC can be 95% confident that the average length of a call for the population is between 3.98 and 4.54 minutes.

2. A survey conducted by "Addis Zemen Gazetta" found that the sample mean age of men was 44 years and the sample mean age of women was 47 years. All together, 454 people from Addis were included in the reader poll –340 women and 114 men. Assume that the population standard deviation of age for both men and women is 8 years.

a. Develop a 95% confidence interval estimate for the mean age of the population men who read the gazetta.

b. Develop a 95% confidence interval estimate for the mean age of the population women who read the gazetta.

c. Compare the widths of the two interval estimates form part (a) & (b) which one has a better precision? Why?

Solution:

a.

n= 114 men $\overline{X}$ = 44 years σ = 8 years C= 0.95

i. $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{114}}$ = 0.75

ii. a = 1 – C = 1- 0.95 = 0.05

$\alpha/2$ = 0.05/2 = 0.025

iii. $Z_{\alpha/2} = Z_{0.025} = 1.96$

iv. $\mu = \overline{X} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$

= 44 ± 1.96(0.75)

= 4.26 ± 1.47

➡ 42.53 ≤ μ ≤ 45.47

b.

n= 340 women $\overline{X}$ = 47 years σ = 8 years C= 0.95

i. $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{340}}$ = 0.434

ii. a = 1 – C = 1- 0.95 = 0.05

$\alpha/2$ = 0.05/2 = 0.025

iii. $Z_{\alpha/2} = Z_{0.025} = 1.96$

iv. $\mu = \overline{X} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$

= 47 ± 1.96(0.434)

= 47 ± 0.85

➡ 46.15 ≤ μ ≤ 47.85

i. Part b has a better precision because the sample size is larger as compared with part a.

3. Time magazine reports information on the time required for caffeine from products such as coffee and soft drinks to leave the body after consumption. Assume that the 99% confidence interval estimate of the population mean time for adults is 5.6 hrs to 6.4 hrs.
   a. What is the point estimate of the mean time for caffeine to leave the body after consumption?
   b. If the population standard deviation is 2 hrs, how large a sample was used to provide the interval estimate?

Solution:

C = 0.99 Confidence interval: 5.6 ≤ μ ≤6.4

a. point estimate = $\frac{5.6+6.4}{2} = 6\ hours$

Or;

$$+\begin{cases} 5.6 = \overline{X} - Z_{\alpha/2}\, {}^{\sigma}\!/\!_{\sqrt{n}} \\ 6.4 = \overline{X} + Z_{\alpha/2}\, {}^{\sigma}\!/\!_{\sqrt{n}} \end{cases}$$

12 = 2$\overline{X}$

$\overline{X}$ = 6 hours

b. 0.99 σ = 2 hours Confidence interval: 5.6 ≤ μ ≤6.4

n=?

a = 1- C = 1- 0.99 = 0.01 a/2 = 0.005 $Z_{\alpha/2} = Z_{0.005} = 2.57$

$$6.4 = \overline{X} + Z_{\alpha/2}\, {}^{\sigma}\!/\!_{\sqrt{n}}$$

$6.4 = 6 + 2.57 \; {}^{2}\!\!/\!{}_{\sqrt{n}}$

$0.4 = {}^{5.14}\!\!/\!{}_{\sqrt{n}}$; rearranging the expression

$\sqrt{n} = \dfrac{5.14}{0.4}$

$\sqrt{n} = 12.85$; squaring both sides

n = 165

We state with 99% confidence that the mean time required for caffeine to leave the body after consumption lies between 5.6 and 6.4 hrs.

Confidence interval estimate of μ - Normal population, σ unknown, n large

If we know that the population is normal, and we know the population standard deviation, σ the confidence interval for μ should be constructed in the manner already shown, i.e., $\mu = \overline{X} \pm Z_{\alpha/2} \; {}^{\sigma}\!\!/\!{}_{\sqrt{n}}$. If the population standard deviation is unknown, it has to be estimated from the sample; i.e., when σ is unknown, we use sample standard deviation: $S = \sqrt{\dfrac{\Sigma(X_i - \overline{X})^2}{n-1}}$.

Then, the standard error of the mean, $\sigma_{\overline{X}}$, is estimated by the sample standard error of the mean: $S_{\overline{X}} = {}^{S}\!\!/\!{}_{\sqrt{n}}$.

Therefore, the confidence interval to estimate μ when population standard deviation is unknown, population normal and n is large is[2]

$$\mu = \overline{X} \pm Z_{\alpha/2} \; {}^{S}\!\!/\!{}_{\sqrt{n}} .$$

[2] This formula also works for large sample size even though the parent population is not normally distributed.

Examples:

1. Suppose that a car rental firm in Addis wants to estimate the average number of miles traveled by each of its cars rented. A random sample of 110 cars rented reveals that the sample means travel distance per day is 85.5 miles, with a sample standard deviation of 19.3 miles. Compute a 99% confidence interval to estimate μ.

   Solution:

   n= 110 rented cars $\overline{X}$ = 85.5 miles s = 19.3 miles C= 0.99

   i. $S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{19.3}{\sqrt{110}} = 1.84$

   ii. a = 1 – C = 1- 0.99 = 0.01

   $\alpha/2 = 0.01/2 = 0.005$

   iii. $Z_{\alpha/2} = Z_{0.005} = 2.57$

   iv. $\mu = \overline{X} \pm Z_{\alpha/2}\, {}^{s}\!\big/\!\sqrt{n}$

   $= 85.5 \pm 2.57(1.84)$

   $= 85.5 \pm 4.73$

   ➡ $80.77 \leq \mu \leq 90.23$

   We state with 99% confidence that the average distance traveled by rented cars lies between 80.77 and 90.23 miles.

2. A study is being conducted in a company that has 800 workers. A random sample of 50 of these workers reveals that the average sample age is 34.3 years, and the sample standard deviation is 8 years. Assuming normality, construct a 98% confidence interval to estimate the average age of all workers in this company.

   Solution:

   n= 50 workers N = 800 workers $\overline{X}$ = 34.3 years s = 8 years C= 0.98

i. $S_{\overline{X}} = \frac{S}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$ [3]= $\frac{8}{\sqrt{50}} * \sqrt{\frac{800-50}{800-1}}$ = 1.10

ii. a = 1 – C = 1- 0.98 = 0.02

$\alpha/2$ = 0.02/2 = 0.01

iii. $Z_{\alpha/2} = Z_{0.01} = 2.33$

iv. $\mu = \overline{X} \pm Z_{\alpha/2} \, {}^{s}\!/\!{}_{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$

= 34.3 ± 2.33(1.10)

= 34.3 ± 2.56

➡ 31.74 ≤ μ ≤ 36.86

We state with 98% confidence that the mean age of workers lies between 31.74 and 36.86 years.

Confidence interval for $\mu - \sigma$ unknown, n-small, population normal

If the sample size is small (n<30), we can develop an interval estimate of a population mean only if the population has a normal probability distribution. If the sample standard deviation s is used as an estimator of the population standard deviation σ and if the population has a normal distribution, interval estimation of the population mean can be based up on a probability distribution known as t-distribution.

Characteristics of t-distribution

1. The t-distribution is symmetric about its mean (0) and ranges from - ∞ to ∞.
2. The t-distribution is bell-shaped (unimodal) and has approximately the same appearance as the standard normal distribution (Z- distribution).

[3] Since the sample size is greater than 5% of the population size, finite population multiplier is used to calculate the sample standard error of the mean.

3. The t-distribution depends on a parameter v (Greek Nu)[4], called the degrees of freedom of the distribution. N = n -1, where n is sample size. The degree of freedom, v, refers to the number of values we can choose freely.
4. The variance of the t-distribution is v/ (v-2) for v>2.
5. The variance of the t-distribution always exceeds 1.
6. As v increases, the variance of the t-distribution approaches 1 and the shape approaches that of the standard normal distribution.
7. Because the variance of the t-distribution exceeds 1.0 while the variance of the Z-distribution equals 1, the t-distribution is slightly flatter in the middle than the Z-distribution and has thicker tails.
8. The t-distribution is a family of distributions with a different density function corresponding to each different value of the parameter v. That is, there is a separate t-distribution for each sample size. In proper statistical language, we would say, "There is a different t-distribution for each of the possible degrees of freedom".
9. The t formula for sample when σ is unknown, the sample size is small, and the population is normally distributed is: $t = \frac{\overline{X} - \mu}{S_{\overline{X}}} = \frac{\overline{X} - \mu}{s / \sqrt{n}}$ This formula is essentially the same as the z-formula, but the distribution table values are not.

   The confidence interval to estimate μ becomes:

[4] What are degrees of freedom? We can define them as the number of values we can choose freely. In general, the degrees of freedom for a t statistic are the degrees of freedom associated with the sum of squares used to obtain an estimate of the variance. The variance estimate depends on not only on the sample size but also on how many parameters must be estimated with the sample:

Degrees of = Number of − Number of parameters that
freedom Observations must be estimated beforehand

Here we calculate sample variance by using n observations and estimating one parameter (the mean). Thus, there are (n – 1) degrees of freedom.

$$\mu = \overline{X} \pm t_{\alpha/2,v} \, {}^{s}\!/\!_{\sqrt{n}}$$

Where: $\overline{X}$ = sample mean

α = 1 – C

v = n – 1 (degrees of freedom)

s = sample standard deviation

n = sample size

μ = unknown population mean

Steps:

i. Calculate degrees of freedom (v=n-1) and sample standard error of the mean.
ii. Compute $\alpha/2$
iii. Look up $t_{\alpha/2,V}$
iv. Construct the confidence interval
v. Interpret results

Examples:

1. If a random sample of 27 items produces $\bar{x}$ = 128.4 and s = 20.6. What is the 98% confidence interval for μ? Assume that x is normally distributed for the population. What is the point estimate?

   Solution:

   The point estimate of the population mean is the sample mean, in this case 128.4 is the point estimate.

n= 27 $\overline{X}$ = 128.4 s = 20.6 C= 0.98

i. $S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{20.6}{\sqrt{27}}$ = 3.96 v = n – 1 = 27-1 = 26

ii. a = 1 – C = 1- 0.98 = 0.02

$\alpha/2$ = 0.02/2 = 0.01

iii. $t_{\alpha/2,v} = t_{0.01,26} = 2.479$

iv. $\mu = \overline{X} \pm t_{\alpha/2,v} \, s/\sqrt{n}$

= 128.4 ± 2.479(3.96)

= 128.4 ± 9.82

➡ 118.56 ≤ μ ≤ 138.22

We state with 98% confidence that the population mean lies between 118.56 and 138.23.

2. A sample of 20 cab fares in Bahir Dar city shows a sample mean of Br 2.50 and a sample standard deviation of Br. 0.50. Develop a 90% confidence interval estimate of the mean cab fares in Bahir Dar city. Assume the population of cab fares has a normal distribution.

n= 20 $\overline{X}$ = Birr 2.50 s = Birr 0.50 C= 0.90

i. $S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{0.5}{\sqrt{20}}$ = 0.112 v = n – 1 = 20-1 = 19

ii. a = 1 – C = 1- 0.90 = 0.10

$\alpha/2$ = 0.10/2 = 0.05

iii. $t_{\alpha/2,v} = t_{0.05,19} = 1.729$

iv. $\mu = \overline{X} \pm t_{\alpha/2,v} \, s/\sqrt{n}$

= 2.50 ± 1.729(0.112)

= 2.50 ± 0.194

➡ $2.31 \leq \mu \leq 2.69$

We state with 90% confidence that the mean of cab fares in Bahir Dar city lies between Birr 2.31 and 2.69.

3. Sales personnel for X Company are required to submit weekly reports listing customer contacts made during the week. A sample of 61 weekly contact reports showed a mean of 22.4 customer contacts per week for the sales personnel. The sample standard deviation was 5 contacts.
   a. Develop a 95% confidence interval estimate for the mean number of weekly customer contacts for the population of sales personnel.
   b. Assume that the population of weekly contact data has a normal distribution. Use the t distribution to develop a 95% confidence interval for the mean number of weekly customer contacts.
   c. Compare your answer for parts (a) and (b). What do you conclude from your results?

Solution:

a. n= 61 weekly contact reports[5] $\overline{X}$ = 22.4 contacts s = 5 contacts C= 0.95

i. $S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{5}{\sqrt{61}}$ = 0.64

ii. a = 1 – C = 1- 0.95 = 0.05

$\alpha/2$ = 0.05/2 = 0.025

iii. $Z_{\alpha/2} = Z_{0.025} = 1.96$

iv. $\mu = \overline{X} \pm Z_{\alpha/2} \, {}^{s}\!/\!{}_{\sqrt{n}}$

= 22.4 ± 1.96(0.64)

= 22.4 ± 1.25

➡ $21.15 \leq \mu \leq 23.65$

I state with 95% confidence that the mean weekly contact lies between 21.15 and 23.65 contacts.

[5] Since the sample size is large, we use the Z-distribution to construct the confidence interval.

b. n= 61 weekly contact reports $\overline{X}$ = 22.4 contacts s = 5 contacts C= 0.95

i. $S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{5}{\sqrt{61}}$ = 0.64 v = n – 1 = 61 – 1 = 60

ii. a = 1 – C = 1- 0.95 = 0.05

$\alpha/2$ = 0.05/2 = 0.025

iii. $t_{\alpha/2,v} = t_{0.025,60} = 2.00$

iv. $\mu = \overline{X} \pm t_{\alpha/2,v} \, {}^{s}\!/\!{}_{\sqrt{n}}$

= 22.4 ± 2.00 (0.64)

= 22.4 ± 1.28

➡ 21.12 ≤ μ ≤ 23.68

I state with 95% confidence that the mean weekly contact lies between 21.12 and 23.68 contacts.

c. As the sample size increases, the t-distribution and z (normal) distribution approximate to be equal.

Confidence interval for $\mu - n$ small, σ unknown, population not normal

This is solved by non-parametric tests, which do not require assumption about the underlying form of the population data.

Determination of Sample Size

The reason for taking a sample from a population is that it would be too costly to gather data for the whole population. But collecting sample data also costs money; and the larger the sample, the higher the cost. To hold cost down, we want to use as small a sample as possible. On the other hand, we want a sample to be large enough to provide "good" approximation/estimates of population parameters. Consequently, the question is "How large should the sample be?"

The answer depends on three factors:

1) How precise (narrow) do we want a confidence interval to be?
2) How confident do we want to be that the interval estimate is correct?
3) How variable is the population being sampled?

Sample size for estimating population mean, μ

The confidence interval for μ is $\mu = \overline{X} \pm Z_{\alpha/2} \sigma / \sqrt{n}$.

From the above expression $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called error of estimation (e). That is, the difference between $\overline{x}$ and μ which results from the sampling process. So

$$e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Squaring both sides results in $e^2 = Z^2_{\alpha/2}\frac{\delta^2}{n}$. Solving for n results in, $n = \frac{Z^2_{\alpha/2}\sigma^2}{e^2}$

$$\boxed{n_{\mu} = \left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2}$$

Examples:

1. A gasoline service station shows a standard deviation of Birr 6.25 for the changes made by the credit card customers. Assume that the station's management would like to estimate the population mean gasoline bill for its credit card customers to be within ± Birr 1.00. For a 95% confidence level, how large a sample would be necessary?

Solution:

e = Birr 1.00 σ = Birr 6.25 C = 0.95 $Z_{\alpha/2} = Z_{0.025} = 1.96$

$$n_{\mu} = \left(\frac{Z_{\alpha/2}\sigma}{e} \right)^2$$

$$n_{\mu} = \left(\frac{1.96 * 6.25}{1} \right)^2$$
$$= 150.06 \approx 151$$ [6]

2. The National Travel and Tour Organization (NTO) would like to estimate the mean amount of money spent by a tourist to be with in Birr 100 with 95% confidence. If the amount of money spent by tourist is considered to be normally distributed with a standard deviation of Br 200, what sample size would be necessary for the NTO to meet their objective in estimating this mean amount?

Solution:

e = Birr 100 σ = Birr 200 C = 0.95 $Z_{\alpha/2} = Z_{0.025} = 1.96$

$$n_{\mu} = \left(\frac{Z_{\alpha/2}\sigma}{e} \right)^2$$

$$n_{\mu} = \left(\frac{1.96 * 200}{100} \right)^2$$
$$= 15.37 \approx 16$$

If population standard deviation, σ, is unknown we have to make an educated guess or take a pilot sample and estimate it.

[6] It a procedure for determining sample size produces a non-integer value, always round to the next larger integer.

- The rough approximation is $\sigma = \frac{H-L}{4}$ because 95.4% of the total population falls within $\pm 2\sigma$.

  $\sigma = 1/4\, range$

Relationship between the error term and sample size

Reducing error term in estimation of an interval estimate to 1/a of the original amount, while holding the confidence level constant requires a sample size of a2 times the original sample size.

Sample size for estimating population proportion, p

The confidence interval for p is $P = \bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}\bar{q}}{n}}$. The expression $Z_{\alpha/2}\sqrt{\frac{\bar{p}\bar{q}}{n}}$ is called the error term (e). That is,

$e = Z_{\alpha/2}\sqrt{\frac{\bar{p}\bar{q}}{n}}$, squaring both sides

$e^2 = Z^2_{\alpha/2}\frac{\bar{p}\bar{q}}{n}$, solving for n

$$n_p = \frac{Z^2_{\alpha/2}\bar{p}\bar{q}}{e^2}$$

Since we are trying to determine n, we cannot have $\bar{p}$ *and* $\bar{q}$. Instead, we should have p and q. so it becomes $\boxed{n_p = \left(\frac{Z_{\alpha/2}}{e}\right)^2 pq}$

Example

1. Suppose that a production facility purchases a particular component parts in large lots from a supplier. The production manager wants to estimate the proportion of defective parts received from this supplier. She believes that the proportion of defects is no more than 0.2 and wants to be with in 0.02 of the true proportion of defects with a 90% level of confidence. How large a sample should she take?

   Solution:

   e = 0.02 p = 0.2 q =0.8 C = 0.90

   $Z_{\alpha/2} = Z_{0.05} = 1.64$

$$n_p = \left(\frac{Z_{\alpha/2}}{e}\right)^2 pq$$

$$n_p = \left(\frac{1.64}{0.02}\right)^2 0.2 * 0.8$$
$$= 1075.84 \approx 1076$$

2. What is the largest sample size that would be needed in estimating a population proportion to within ± 0.02, with a confidence coefficient of 0.95?

   Solution:

   e = 0.02 C = 0.95 $Z_{\alpha/2} = Z_{0.025} = 1.96$

   The largest sample size would be obtained when p = 0.5. So,

$$n_p = \left(\frac{Z_{\alpha/2}}{e}\right)^2 pq$$

$$n_p = \left(\frac{1.96}{0.02}\right)^2 0.5 * 0.5$$
$$= 2401$$

If p is unknown and there is no possibility of estimating it, use 0.5 as the value of p because it will generate the greatest possible sample size as compared with other values.

Determining Sample Size When Estimating $\mu_1 - \mu_2$

When taking two random samples and using the difference in sample means to estimate the difference in population means, a researcher should have an idea of how large the sample sizes need to be solving for n form the formula

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

does not look promising, because the equation has nine variables including two different values of n. However making some assumptions can generate a workable sample size formula.

1. Variances of the two populations are the same: $\sigma_1^2 = \sigma_2^2 = \sigma^2$
2. The sample size for each sample is the same: $n_1 = n_2 = n$

The difference between $\overline{x}_1 - \overline{x}_2$ and $\mu_1 - \mu_2$ is the error of estimation. Or $e = (\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)$.

Incorporating these assumptions into the z-formula yields

$$Z = \frac{e}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}} = \frac{e}{\sqrt{\frac{2\sigma^2}{n}}} = \frac{e}{\sqrt{\frac{2}{n}\sigma^2}} = \frac{e}{\sqrt{\frac{2}{n}}\ \sigma}$$

Solving for n produces the sample size:

$$\boxed{n = \frac{2 * Z_{\alpha/2}^2 \sigma^2}{e^2} = 2\left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2}$$

The above formula suggests that the necessary sample sizes for comparing two sample means are each twice as large as the required sample size for estimating single sample means. It is clear that the larger the sample, the more it costs. Thus sample size formulas can be effective aids in ensuring that a research project's goals are met and that the cost of sampling is minimized.

Examples:

1. A college admissions officer wants to estimate the difference in the average Statistics scores of men and women. She plans to take a random sample of men and women who have taken the Statistics at the same time. She wants to be within 10 points of the true difference in the mean scores of men and women and 95% confident of her results. Past Statistics test results indicate that the standard deviation of Statistics test scores is about 105 points. How large the sample sizes be?

   Solution:

   e = 10 points σ = 105 pointsC = 0.95 $Z_{\alpha/2} = Z_{0.025} = 1.96$ n=?

$$n = 2\left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2$$

$$n = 2\left(\frac{1.96 * 105}{10}\right)^2$$

$$= 2(421.54)$$

$$= 847.10 \approx 848$$

2. A researcher wants to estimate the difference between the average price of a 21-inch black and white TV and the average price of a 21-inch color TV set. He believes that the standard deviation of the price of a 21-inch TV set is about Birr 100. He wants to be 99% confident of his results and within Birr 20 of the true difference. How large a sample should he take for each type of television set?

   Solution:

   e = Birr 20 σ = Birr 100 C = 0.99 $Z_{\alpha/2} = Z_{0.005} = 2.57$ n=?

$$n = 2\left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2$$

$$n = 2\left(\frac{2.57 * 100}{20}\right)^2$$

$$= 2(165.12)$$

$$= 330.24.10 \approx 331$$

4. HYPOTHESES TESTING

4.1. Introduction

Hypothesis testing is a method of analysis that makes inferences about population parameters or relationship of variables from sample data. The procedure is that we make an assumption about the population parameter or the relationship variables under study and then test the assumptions on the basis of sample information or data. It is one the major techniques of statistical inference, which is used to solve many practical problems in business and in any field in general.

Objectives

Upon completion of this unit you will be able to

- Define Hypothesis And Statistical Tests
- Identify the types of hypothesis and types of errors
- Know level of significance and power of a test
- Test about means
- Test about proportions
- Make analysis of variance

4.2. Definition of hypothesis

A hypothesis is a statement or assumption made about the population parameters, which may or may not be true. It also takes the form of a statement or assumption about relationship between two variables, an independent variable (cause) and a dependent variable (effect). A hypothesis is not necessarily true, it can be either right or wrong, and we use a sample data to help us decide. If we know everything, there is no need for statistical hypothesis testing. When there is uncertainty, statistical hypothesis testing will help us learn as much possible from the information available to us. In broader view, hypothesis testing is one component of the decision- making process.

As an example consider the following.

1. The proportion of educated people in Ethiopia is not less than 30%.
2. The average monthly income of families of AA is less or equal to Birr 800.
3. The cause of high number of Aids patients in Africa is poverty.

The first two examples are assumptions about population parameters, proportion of educated people in Ethiopia and average family monthly income respectively. The third example shows assumption about the relationship of two variables, poverty and number of aids patients. In all cases of the above examples, we need to draw conclusions or make decisions about the specified population parameters or relationship of variables on the basis of sample information. These decisions are called **statistical decisions**. These assumptions about the population parameters or relationship of variables are called **hypotheses**. Such main statements (assumptions) are also called **null hypotheses** which are usually denoted by Ho. A hypothesis accepted when Ho is rejected is called an **alternative hypothesis** which is also denoted by Ha or $H_{1.}$ The alternative hypothesis, which is also called research hypothesis, is to be accepted only if there is convincing statistical evidence that would rule out the null hypothesis as a reasonable possibility. Hence, hypotheses are composed of both the null and alternative hypotheses and are always formulated as opposite, so that when one is true the other is false. Therefore, the above assumptions are not full hypotheses, for they represent only the null hypothesis, which is the most likely assumption. Let us now try to rewrite the above assumptions into full flagged hypotheses.

1. Ho: The proportion of educated people in Ethiopia is not less than 30%.

Ha: The proportion of educated people in Ethiopia is less than 30%.

Or Ho: $P \geq 0.3$

Ha: $P < 0.3$

2. Ho: The average monthly income of families of AA is less or equal to Birr 800.

   Ha: The average monthly income of families of AA is more than Birr 800.

   Or Ho: $\mu \leq 800$

   Ha: $\mu > 800$

3. Ho: The cause of high number of Aids patients in Africa is poverty.

   Ha: The cause of high number of Aids patients in Africa is not poverty.

Based on sample values we either accept or reject Ho. If Ho is rejected then the alternative hypothesis, Ha, is accepted. Rejecting Ho could either be the assumption is truly wrong or due to chance that the sample data is wrong. If the decision cannot be explained as being probably due to solely to chance, the difference is said to be statistically significant. Tests devised to check whether this is so are called significant tests. The set of outcomes or values of the random sample for which we reject Ho is known as the rejection or critical region, while the set of values for which Ho is accepted is known as the acceptance region. Thus, a test of hypotheses is the partitioning of the set of outcomes into the critical and the acceptance region.

In testing hypothesis we always consider the following assumptions

1. The data set is a random sample from the population of interest, and
2. Either the quantity being measured is approximately normal, or else the sample size is large enough that the central limit theorem ensures that the sample average is approximately normally distributed.

Let us now consider the hypothesis about thee average monthly family income of AA given above

Ho: μ ≤ 800

Ha: μ > 800

The objective here is to check if the claim or assumption made about the average family income is correct or not.

The decision as to whether or not to approve the assumption will depend on the test results from the sample data. Therefore, the decision will be based on the level of the sample mean, $\overline{X}$, which is referred to as the test statistic.

To make test suppose we take a random sample of 500 families and decide to reject Ho if $\overline{X}$ > 800 and accept Ho if $\overline{X}$ ≤ 800.

The value 800 is commonly referred to as critical value. The test statistic is then used to formulate a decision rule to translate sample evidence into a course of action. Therefore, the decision is accept Ho $\overline{X}$ ≤ 800 and reject Ho if $\overline{X}$ > 800.

Therefore, the set of values {$\overline{X}$ > 800} is called critical or rejection region and the set of values {$\overline{X}$ ≤ 800} ia called acceptance region.

When testing such hypotheses we always make errors since the test depends on sample data. There are two types of such errors, type I and type II errors. Type I error is made when Ho is wrongly rejected, while Type II error is made when Ho is wrongly accepted. These two types of errors can be summarized as follow

	Accept Ho	Reject Ho
Ho True	No error	Type I error
Ho False	Type II error	No error

Notation

The probability of type I error is denoted by α and the probability of type II error is denoted by β.

P (Type 1 error) = P (rejecting Ho / Ho is true) = α

P (Type 11 error) = P (accepting Ho / Ho is false) = β

Level of significance. The probability of rejecting H_o when it is infact true is denoted by α and is known as a level of significance. That is,

Level of significance = α = P (reject H_o/ H_o is true)

= P (type I error)

α is usually taken as 0.05 or 0.01.

Power of a test. The power of a test is defined as the probability of rejecting the null hypothesis H_o when it is actually false or the probability of accepting Ho when it is actually false. That is,

Power of a test = P (rejecting H_o/ H_o is false)

= 1 – P (accept H_o/ H_0 is false)

= 1 – P (type II error)

= 1 - β

where β is the probability of type II error.

In a similar manner we show that,

Power of a test = 1 - α This shows that the smaller the two types of errors, α and β, the stronger the power of the test. Therefore, the objective of any test is to minimize the two types of errors. However, it can be shown that we cannot simultaneously minimize the two types of errors. Hence when testing hypothesis, We determine the critical values by first specifying alpha (α), the probability of committing type I error. If the cost of committing type I error is high, the decision makers should specify a small alpha. A small alpha results in a small rejection region. If the rejection region is small, the sample mean is less likely to fall there, and the chances of rejecting a true null hypothesis are small. If the acceptance region is large, the probability of committing a type II error (β) will also tend to be large. A decrease in α will increase in β. However, the increase in β will not equal the decrease in α. Decision makers must examine carefully the costs of committing type I and type II errors and, in light of these costs, attempt to establish acceptable values of α and β.

The goodness of the choice of the critical value 9 depends on the attitude of the decision towards the two types of errors, Type I and Type II errors. The smaller the two types of errors, the better the test. These errors can be computed for any particular test. It can be shown that when one of the probabilities is decreased the other tends to increase. Therefore, the general agreement is that we fix the probability of type one error at α and try to minimize type two error β.

Hypotheses in general are of two forms

Tests of the form,

Ho: $\mu = \mu_o$ or Ho: $\mu = \mu_o$

Ha: μ N μ_o Ha: $\mu < \mu_{o,}$

in which the critical region includes either large or small values of the test statistic are called one- sided or one –tailed tests.

A test statistic is a formula which is used for computing from the data in testing hypothesis. The value of the test statistic is used in determining whether or not we may reject the null hypothesis.

And those of the form,

Ho: $\mu = \mu_o$

Ha: $\mu \neq \mu_o$

Including both large and small values of the test statistic are called two – tailed tests.

Steps in testing hypotheses

In general statistical testing of hypothesis follows certain steps that are more or less taken by most researchers.

1. Formulate the hypothesis. That is, rewrite the problem to a testable form.
2. Determine alpha (α), which is usually given
3. Determine the appropriate statistical distribution used.
4. Determine the test statistic.
5. Establish the critical region of the test statistic or give the decision rule. The

   decision rule of a statistical hypothesis test is a rule that specifies the conditions under which the null hypothesis may be rejected.
6. Gather data.
7. Calculate the value of the test statistic and find the tabulated value.
8. Make a decision, that is, accept or reject the null hypothesis, Ho.

4.3. Tests *about* a single mean

Hypotheses tests about means, just like confidence interval estimation about means considered in the previous unit, are dependent on whether the population is normal or not, samples large or small and population standard deviation known or unknown. Let us see the different possible combinations of these cases and make tests about mean.

4.3.1. Normal population, variance known, large or small sample

Suppose we take a random sample of size n (large or small) from an N (μ, σ^2) population when σ^2 is known. We want to test

H_o: $\mu = \mu_o$

Ha: $\mu > \mu_o$, where μ_o is a specific value.

Since the best estimator of μ is $\overline{X}$, the test statistic must depend on $\overline{X}$.

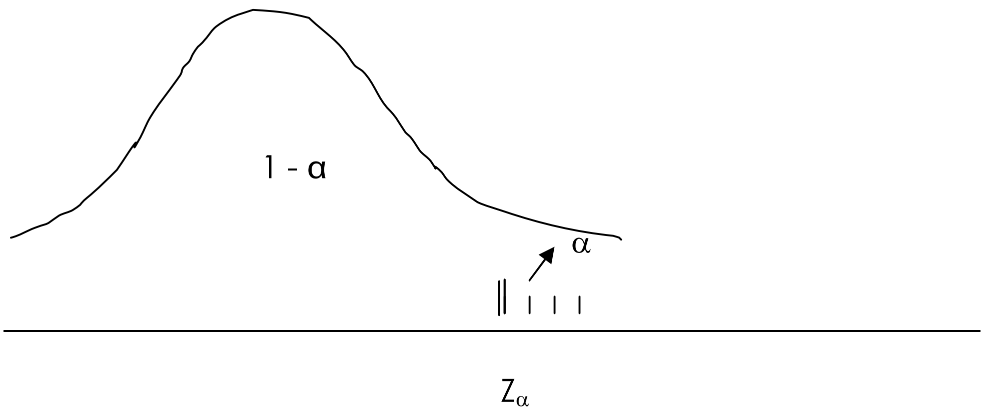
But in this case, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ and $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

When Ho is true, $Z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} \sim N(0, 1)$

Note that $\frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}}$ is the test statistic for testing single mean.

If μ is really greater than μ_o, $\overline{X}$ would most probably be greater than μ_o and therefore, $Z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} > C$

where, the constant C = $Z\alpha$, the value of which is found using the level of significance (α) from the normal table.



$$P\left(\frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} > C\right) = \alpha$$

Critical region for testing the above hypothesis is

$$\frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} > Z_\alpha$$

If Ho: $\mu = \mu_o$

Ha: $\mu < \mu_o$

In a similar argument as above, the critical region is,

$$\frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} < -Z_\alpha$$

To test the two- tailed tests,

Ho: $\mu = \mu_o$

Ha: $\mu \neq \mu_{o,}$ there is no difference between the believed population mean and the specified value.

Critical region can be shown to be,

$$\left| \frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} \right| > Z_{\alpha/2}$$

Example 1

According to a car manufacturing company, their cars averaged at least 32 miles per gallon in the city. From past records it is known that mileage is normally distributed with standard deviation of 2.5 miles per gallon. Tests on 16 cars showed that mean mileage in the city is 31.5 miles per gallon. Do the data support the claim of the company at 1% level of significance?

Solution

The appropriate hypothesis is,

Ho: $\mu \geq 32$

Ha: $\mu < 32$

Here, population is normal, sample size, n = 16, is small and population standard deviation, σ = 2.5, is known, and hence we use the normal distribution for the test.

Let $\alpha = 0.01$

Here, we have to reject Ho if

$$\frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} < -Z_\alpha$$

$$Z_{cal} = \left| \frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}} \right| = \left| \frac{0.985-1}{0.105/5} \right| = 0.728$$

$Z_{tab} = -Z\alpha = -Z_{0.01} = -2.33$

Since $Z_{cal} > Z_{tab}$, accept Ho.

Therefore, the claim of the company is correct or the data support the claim of the company.

Self Test Activity 4.1

A producer of an electric bulb claims that the average life length of its product is more than 1800hrs. Consumer's protection agency on the other hand doubts the claim. Therefore, it takes a random sample of 400 bulbs and found out that the mean life length is 1780hrs with standard deviation of 200hrs. Assuming that life length of bulbs is normal, should the agency support the producer's claim at 5% level of significance?

4.3.2. Non-normal population, large sample, σ known or unknown

We wish to test,

$H_o: \mu = \mu_o$

$H_a: \mu > \mu_o$

Here, since n is large by the central limit theorem, $\overline{X}$ is approximately normally distributed. Therefore,

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N\ (0, 1)$$

Thus the test statistic and critical region are the same as the above case,

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} > Z_\alpha$$

The critical regions for Ha: $\mu < \mu_o$ and Ha: $\mu \neq \mu_o$ could similarly be shown to be the same as above.

Note that if σ is unknown it could be estimated by S, the sample standard deviation given by

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

Example 2

Workers of a given industry complain that their average monthly salary is at most birr 600. Workers union wanted to test the claim of the workers and takes a random sample of 100 workers of the factory that produce a mean monthly salary of Birr 610 with a standard deviation of Birr 50. What should the workers union conclude about the claim of the workers at 5% level of significance?

Solution

The hypothesis is

Ho: $\mu \leq 600$

Ha: $\mu > 600$

Here, σ is unknown and n is large. Therefore, by central limit theorem, the test distribution is Z-distribution.

Critical region is $\frac{\overline{X}-\mu}{s/\sqrt{n}} >$ Z$_\alpha$

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} = \frac{610-600}{50/\sqrt{100}} = 2$$

For σ = 0.05, $Z_{0.05}$ = 1.645

Since $Z_{cal} > Z_{tab}$, then Ho is rejected. Therefore, the workers union should reject the claim of the workers.

4.3.3. Normal population, small sample, σ unknown

In some cases of testing hypotheses, because of reasons such as time, money, convenience or availability, is able to gather only a small random sample of data. We usually say n is small if n < 30. In such cases, if the data are normally distributed in the population and σ is known, the Z- test can be used.

However, if population standard deviation, σ, is unknown the Z test cannot be used but rather the t - test based on the t - distribution is applicable. Here again we replace σ by S, sample standard deviation, where

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

Suppose we wish to test Ho: $\mu = \mu_o$

Ha: $\mu > \mu_o$

We have seen earlier that in this case, $\overline{X}$ has a t- distribution with (n-1) degree of freedom, $\overline{X} \sim$ t (n – 1).

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t\ (n-1)$$

which gives the critical region as,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} > t_\alpha \text{ (n-1)}$$

If Ho: $\mu = \mu_o$

Ha: $\mu < \mu_o$

Critical region is $\frac{\overline{X}-\mu}{S/\sqrt{n}} < - t_\alpha$

And also if Ho: $\mu = \mu_o$

Ha: $\mu \neq \mu_o$

Then critical region similarly is given by

$$\left| \frac{\overline{X}-\mu}{S/\sqrt{n}} \right| > t_{\alpha/2}$$

Example 3

It is known that the average body temperature of human beings is a normal variable with mean 37^{0}c. A medical person took the temperatures of ten patients as 39, 37, 41, 35, 37, 37, 34, 38, 40, and 39 in 0 c. Test at the 5% level of significance whether these sample data are consistent with the specified human body temperature.

Solution

The hypothesis is

Ho: $\mu = 37$

Ha: $\mu \neq 37$

$\overline{X} = \frac{\sum Xi}{n}$ =, 37.7   S = $\sqrt{\frac{\sum(Xi-\overline{X})^2}{n-1}}$ = 1.84

Critical rejoin is $\left| \frac{\overline{X} - \mu_o}{S/\sqrt{n}} \right| > t_{\alpha/2}$

$t_{cal} = \left| \frac{\overline{X} - \mu_o}{S/\sqrt{n}} \right| = \left| \frac{37.7 - 37}{1.84/\sqrt{10}} \right| = 1.2$

At $\alpha = 0.05$, $t_{tab} = t_{\alpha/2}(n - 1) = t_{0.025}(9) = 2.26$

Since $t_{cal} < t_{tab,}$ accept Ho.

Therefore, the sample data are consistent with the specified body temperature.

Self test Activity 4.2

In a particular school District the I.Q. is known to have a normal distribution with mean of 110. From one school a sample of 25 students were taken and found that their average I.Q. is 115 with variance 100. Is the average in this school different from the district average?

4.4. Test about a single population proportion

As shown in the previous sections, proportion is a value between 0 and 1 that shows the number of items possessing a certain characteristics in the population or sample. There are many situations in business in which we must test the validity of statements about population proportions or percentages. A business man can claim that 90% of population of consumers like his product. A random sample of consumers could be taken test this claim in a similar manner as above.

The method here is to compare the sample proportion with the specified population value or to analyze the qualitative data where we test the presence or absence of a certain characteristics bases on sample values.

By the central limit theorem, if nP > 5 and nQ > 5, it can be shown that the sample proportion, $\hat{p}$, values are approximately normally distributed as shown in the previous section.

That is if nP > 5 and nQ > 5, then $\hat{p}$ is approximately normal with mean P and standard deviation $\sqrt{\frac{PQ}{n}}$

where P = population proportion, $\hat{p}$ = sample proportion and Q = 1 – P.

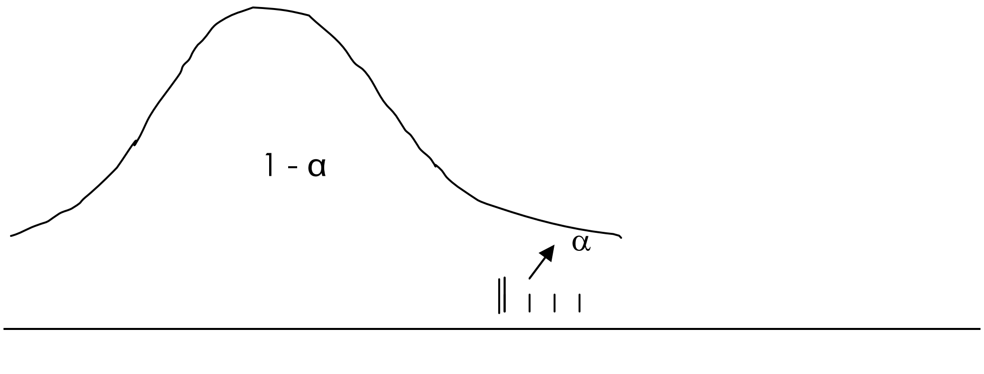
Hence $Z = \frac{\hat{p}-p}{\sqrt{PQ/n}} \sim N(0, 1)$

Again $\frac{\hat{p}-p}{\sqrt{PQ/n}}$ is the test statistic for testing single proportion.

The test procedure is the same as that of the mean, shown in the previous sections. That is the computed value of Z is compared with the table value of Z, which is also known as the critical value of Z.

If the hypothesis is $H_o: P = P_o$

$H_a: P > P_o$, where P_o is a specific value



Z_α

$P(\frac{\hat{p}-p}{\sqrt{PQ/n}} > Z_\alpha) = \alpha$, where α is probability of type I error

Therefore, the critical region is $\frac{\hat{p}-p}{\sqrt{PQ/n}} > Z_\alpha$

Similarly, if Ho: P = P_o

Ha: P < P_o

Critical region is $\frac{\hat{p}-p}{\sqrt{PQ/n}} < -Z_\alpha$

For the two – tailed hypothesis

Ho: P = P_o

Ha: P ≠ P_o

The critical region is $\left|\frac{\hat{P}-P}{\sqrt{PQ/n}}\right| > Z_{\alpha/2}$

Example 4

A survey on the workers of an industry indicates that 240 workers out of a randomly selected sample of 500 workers are not professionals. But workers union believes that less than or equal to half of the workers are professionals. Test at a = 0.05 the claim of workers union.

Solution

The appropriate hypothesis is

Ho: P ≤ 0.5

Ha: P > 0.5

Here, P = 0.5, Q = 1 – P = 0.05, Sample proportion = $\hat{p} = \frac{500-240}{500}$ = 0.52

Critical region is $\frac{\hat{P}-P}{\sqrt{PQ/n}} > Z_\alpha$

The calculated value of Z is

$$Z_{cal} = \frac{\hat{p}-p}{\sqrt{PQ/n}} = \frac{0.52-0.5}{\sqrt{\frac{(0.5)(0.5)}{500}}} = 0.89$$

For $\alpha = 0.05$, the tabulated value is

$Z_{tab} = Z_{\alpha} = Z_{0.05} = 1.645$.

Since $Z_{cal} < Z_{tab}$, accept the Ho.

Therefore, the claim of workers union is correct.

Self Test Activity 4.3

A medical expert claims that at least 60% of diseases of a given community are related to lack of sanitation. A researcher who wanted to test the claim takes a random sample of 500 patients from the community and found out that 350 of the cases are related to sanitation. Test at 1% level of significance if the claim of the medical expert is correct or not.

4.5. Tests involving finite populations

The same way as in the construction of confidence intervals with finite population containing N elements and the sample size, n, constitute at least 5% of the population, we need to use the finite population correction factor, $\sqrt{\frac{N-n}{N-1}}$, in tests of hypotheses too. In this case also the standard error of the respective estimators must be multiplied by the finite population correction factor.

In the case of hypothesis tests about the population mean, μ, we incorporate a finite population correction factor by using the standard error, $\sigma_{\bar{x}} = \sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}$. When the population standard deviation is unknown, we substitute the sample standard deviation, s, for it.

Similarly, for testing hypotheses about population proportion, P, with finite population and the sample constitutes at least 5% of the population, we incorporate the finite population correction factor to get a standard error of the sample proportion, $\hat{p}$, as $\sigma_{\hat{p}}$ = $\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\sqrt{\frac{N-n}{N-1}}$

Example 5

Suppose that a random sample of 200 electric bulbs from a total of 2000 bulbs produces average life length of 12,000 hrs with a standard deviation of Birr 1500 hrs. Test at 5% level of significance that the average life length of the 2000 electric bulbs is at most 11,800 hrs.

Solution

For N = 2000 and n = 200, the sampling fraction, n/N = 200/2000 = 0.1, which is more than 0.05. Therefore, the finite population correction factor must be used in the test.

The hypothesis is

Ho: μ ≤ 11,800

Ha: μ > 11,800

The test statistic is $\frac{\bar{x}-\mu_o}{s/\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$

The critical region is $\frac{\bar{x}-\mu_o}{s/\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$ > Zα

$$\frac{\bar{x}-\mu_o}{s/\sqrt{n}\sqrt{\frac{N-n}{N-1}}} = \frac{12{,}000-11{,}800}{1500/\sqrt{200}\sqrt{\frac{2000-200}{2000-1}}} = 1.99$$

$Z_\alpha = Z_{0.05} = 1.645$

Reject Ho. Therefore, the average life length of the electric bulbs is more than 11,800 hrs.

4.6. Tests about difference of two means

The comparison of two population means is another important and frequently used statistical technique in business. In testing hypotheses about difference of two means the assumptions we saw in testing hypotheses on single mean are also applicable.

4.6.1. Normal populations, variances known

Suppose that we have a random sample of size n_1 from first population and another independent sample of size of n_2 from the second population. The problem here is to see whether there is significant difference between the two population means based on the sample mean difference, $\overline{X}_1$ - $\overline{X}_2$. Where $\overline{X}_1$ and $\overline{X}_2$ are sample means from the first and the second population respectively.

Here, it is obvious that the difference in the sample means is normally distributed with mean, μ_1- μ_2 and variance $\frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2}$

That is, $(\overline{X}_1 - \overline{X}_2) \sim N\ (\mu_1 - \mu_2, \frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2})$

Thus, $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^{\ 2}/n_1 + \sigma_2^{\ 2}/n_2}} \sim N\ (0, 1)$

Suppose we wish to test,

$Ho: \mu_1 = \mu_2$ or $Ho: \mu_1 - \mu_2 = 0$

$Ha: \mu_1 > \mu_2$ $\quad Ha: \mu_1 - \mu_2 > 0$

Here under Ho, the test statistic becomes $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ ~ N (0, 1)

This gives the critical region as

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > Z\alpha$$

If the test is of the form,

Ho: $\mu_1 = \mu_2$ or Ho: $\mu_1 - \mu_2 = 0$

Ha: $\mu_1 < \mu_2$ Ha: $\mu_1 - \mu_2 < 0$

Critical region becomes

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < - Z\alpha$$

Similarly if the test is a two- tailed test

Ho: $\mu_1 = \mu_2$ or Ho: $\mu_1 - \mu_2 = 0$

Ha: $\mu_1 \neq \mu_2$ Ha: $\mu_1 - \mu_2 \neq 0$

The critical region is

$$\left| \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right| > Z\alpha/2$$

Example 6

A sample of 200 students graduated from college in 1990 produce an average age of 25 years and another sample of 250 students graduated from college in 2005 produce an average age of 23.5. Test at 5% level of significance that the average age of graduation decreases in 2005, assuming normal population and standard deviation of ages of graduation of both years are equal to be 5 years.

Solution

Let μ_1 = the average age of all students graduated in 1990

μ_2 = the average age of all students graduated in 2005

The required test is

Ho: $\mu_1 = \mu_2$

Ha: $\mu_1 > \mu_2$

Here, populations are normal, $\sigma_1 = \sigma_2 = 5$, then the test distribution is normal. Therefore, critical region is

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > Z_a$$

$$Z_{cal} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{25 - 23.5}{\sqrt{\frac{25}{200} + \frac{25}{250}}} = 3.16$$

$Z_{tab} = Z_a = Z_{0.05} = 1.645$ Therefore, we reject Ho, which means that the average age of graduation decreases from that of 1990 significantly.

4.6.2. Non- normal population, large samples

Here, since samples are large, by central limit theorem the sample mean difference is approximately normal, and hence the test statistic and critical regions for the various tests in this case is the same as above. It must be noted that in this case results are only approximate. If σ_1^2 and σ_2^2 are unknown they can be replaced with sample variances S_1^2 and S_2^2 respectively, where

$S_1^2 = \frac{1}{n_1 - 1}\Sigma(X_{1_i} - \overline{X}_1)^2$, $S_2^2 = \frac{1}{n_2 - 1}\Sigma(X_{2\,i} - \overline{X}_2)^2$ and $\overline{X}_1$ and $\overline{X}_2$ are sample means of the first and the second populations respectively.

4.6.3. Normal population, variances unknown, small samples

The objective again is to test hypotheses about difference of means of two populations in this case. We have two cases here.

Case1: When population variances are equal, which is, $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$. Because the population variances are equal, σ^2 is logically estimated by the pooled sample variances, $S_p{}^2$, given as

$$S_p{}^2 = \frac{(n_1-1)S_1{}^2 + (n_2-1)S_2{}^2}{n_1+n_2-1}$$

And $(\overline{X}_1 - \overline{X}_2) \sim t(\mu_1 - \mu_2, \frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2})$. Then,

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t\ (n_1 + n_2 - 2)$$

Suppose we wish to test

Ho: $\mu_1 = \mu_2$

Ha: $\mu_1 > \mu_2$

Under Ho, $\frac{\overline{X}_1 - \overline{X}_2}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t\ (n_1+n_2-2)$

Thus the critical region for the above test is

$$\frac{\overline{X}_1 - \overline{X}_2}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_\alpha\ (n_1+n_2-2)$$

If Ho: $\mu_1 = \mu_2$

Ha: $\mu_1 < \mu_2$

Critical region is

$$\frac{\overline{X}_1 - \overline{X}_2}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > -\ t_\alpha\ (n_1+n_2-2)$$

Similarly for the test

Ho: $\mu_1 = \mu_2$

Ha: $\mu_1 \neq \mu_2$

Critical region is

$$\left| \frac{\overline{X}_1 - \overline{X}_2}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > - t_{\alpha/2} \ (n_1+n_2 - 2)$$

Example 7

A transport company wants to compare the fuel efficiency of two types of lorries it operates. It obtains data from samples of the two types of lorries, A and B with the following result.

	Lorry type	
	A	B
Average mileage per liter	10.9	10.0
Standard deviation	2.3	2.0
Sample size	20	25

The population variances, although unknown, are assumed to be equal. Assuming normal population, test at 1% level of significance if there is significance difference in the fuel efficiency between the two lorries.

Solution

The test could be put as

Ho: $\mu_A = \mu_B$

Ha: $\mu_A \neq \mu_B$

The critical region is $\left| \frac{\overline{X}_A - \overline{X}_B}{Sp\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} \right| > t_{\alpha/2}\,(n_1+n_2-2)$

$S_p^2 = \frac{(19)(5.29)+(24)(4)}{20+25-2} = 4.69$

$t_{cal} = \left| \frac{\overline{X}_A - \overline{X}_B}{Sp\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} \right| = 1.39$

At α = 0.05, t_{tab} = $t_{0.025}(43)$ = 2.01 Accept Ho. Therefore, there is no significance difference in fuel efficiency between the two lorries.

Case2: When population variances are not equal, which is, $\sigma_1^2 \neq \sigma_2^2$. Here, it can be shown that, $\frac{(\overline{X}_1 - \overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ ~(approx.) t (ν)

where, the degree of freedom = ν = $\frac{\left[\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right]^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1}+\frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$

Suppose that Ho: $\mu_1 = \mu_2$

Ha: $\mu_1 > \mu_2$

Under Ho, $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ ~(Approx.) t (ν)

Critical region is, $\dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{{s_1}^2/n_1 + {s_2}^2/n_2}} > t_\alpha (\nu)$

Similarly the critical region for the tests with alternative hypotheses Ha: $\mu_1 > \mu_2$ and Ha: $\mu_1 \neq \mu_2$ respectively are

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{{s_1}^2/n_1 + {s_2}^2/n_2}} <-t_\alpha(\nu)$$

And $\left| \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{{s_1}^2/n_1 + {s_2}^2/n_2}} \right| > t_{\alpha/2} (\nu)$

Self test Activity 4.4

Repeat problem 7, above, if population variances are assumed not equal, that is, ${\sigma_1}^2 \neq {\sigma_2}^2$.

4.7. Tests about several means.

We often want to make comparisons among three, four, five or more samples or groups. For instance, we may seek to decide on the basis of sample data whether there really is a difference in the effectiveness of three teaching methods or we may seek to compare the average yields per acre of several varieties of maize, and so on. The method we shall introduce for this purpose is a powerful statistical tool called analysis of variance, ANOVA. It is in general involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the difference among various population means.

Suppose that μ_1, μ_2, --------, μ_k are the means of the k populations from which the samples are drawn, the hypothesis to test if there is significant difference among them is given by the form,

Ho: $\mu_1 = \mu_2 =$ ---------- $= \mu_k$

Ha: these means are not equal (at least two means are equal)

This null hypothesis would be supported if the difference among the sample means are small, and the alternative hypothesis would be supported if at least some of the differences among the sample are large.

Analysis of variance is a statistical methodology based upon the following assumptions.

i. The populations are normally distributed, so that sample statistics tends to reflect the characteristics of the population.
ii. Each sample is independent of the other samples and each sample of size n is drawn randomly.
iii. The populations from which the samples are drawn have equal variances.

The technique in conducting analysis of variance is to split up the **total variation** into two meaningful components, the distance of the raw scores from their group means known as **variation within groups** and the distance of group means from one another referred to as **variation between groups.**

A good criterion for testing the above hypothesis, that is, the hypothesis that whether the population means are the same or not in all cases is simply the computation of a quantity known as the **variance ratio** or **F- ratio**. This quantity is based on the ratio of the two component variances, which is,

$$F = \frac{Beweengroups \operatorname{var} iances}{Withingroups \operatorname{var} iances} = \frac{Meansquarebetweengroups}{Meansquarewithingroups}$$

Now let us see the computations of the F ratio. Even if samples are randomly taken from the same population, there are always variations both between samples and within samples, which comprises the total variation. This total variation is also known as “total sum of squares" or SST, and is the sum of squared differences between each observation and the overall mean, that is

$$\text{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji} - \overline{X})^2$$

where, X_{ji} represents individual observations for all samples and $\overline{X}$ is the grand mean of all sample means. The computational formula for SST is

$$\text{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}X^2{}_{ji} - \frac{T^2}{N}$$

where,

N = n_1+ n_2 + ----------+ n_k = total number observations in all samples.

T = the total sun of all observations

This total sum of squares is contributed by the two component variations,

variance between samples (groups) and variance within samples (groups).

The variation between samples (groups),σ^2 between, is due to the effect of differences in treatments, that is, inter- sample (group) variability. It is known as sum of squares between samples (groups), which is denoted by SSB. It is also given by the formula,

$$\text{SSB} = \sum_{i=1}^{k} n_j(\overline{X}_j - \overline{X})^2$$

where, n_j = the number of observations in the jth group (sample)

$\overline{X}_j$ = the mean of the jth sample (group)

The computational formula for SSB is,

SSB = $\sum_{j=1}^{k}\frac{T_j^{\ 2}}{n_j}-\frac{T^2}{N}$

Where, T_j = total of the j^{th} sample (group)

The variance within samples (groups), σ^2 within, which is also known as sum of squares within samples (groups), denoted by SSW, could be due to sampling error or other natural causes. Some chance variations could still occur to samples that come from the same population.

The sum of squares within samples (groups) is usually obtained from the relationship,

SSW = SST – SSB

Or it can be computed as,

SSW = $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji}-\overline{X}_j)^2$

Or by the computational formula,

SSW = $\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ji}^{\ 2}-\sum_{j=1}^{n_j}\frac{T_j^{\ 2}}{n_j}$

The mean of the sum of squares could therefore be calculated as follows

Mean of squares between samples (groups) is given by,

MSB = $\frac{SSB}{d.f_B}$

And mean squares within samples (groups) is also given by,

MSW = $\frac{SSW}{d.f_W}$

Where, $d.f_B$ = between samples (groups) degree of freedom = k-1. These degrees of freedom are also known as numerator d.f.

$d.f_W$ = within sample (groups) degree of freedom = N-k. These are also called denominator d.f.

k = the number of samples (groups)

The difference between variance between and variance within could be expressed as a ratio to be designated as F-ratio or variance ratio, which is given as

$$F = \frac{\sigma^2 between}{\sigma^2 within} = \frac{SSB / d.f_B}{SSW / d.f_W} = \frac{MSB}{MSW}$$

If population means are exactly the same or equal, then σ^2 between will be equal to σ^2 within and the value of F will be equal to 1.

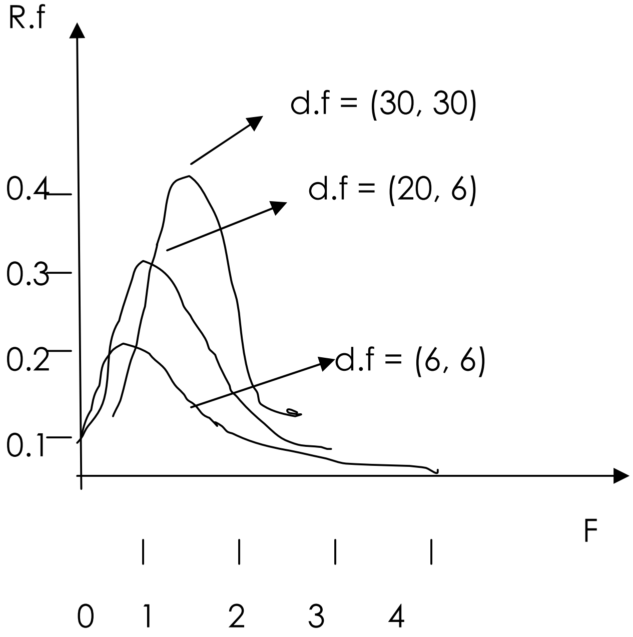
The sampling distribution of the variance ratio is the continuous distribution called F-distribution. The values of F are given on the table called F-distribution table for different α levels and the two degree of freedoms, $d.f_B$ and $d.f_W$, that can be found in any standard statistics book just like the other distributions. That is,

$F_{tabulated} = F_{\alpha}$ (k-1, N-k)

4.8. Properties of the F- distribution

The F- distribution is a continuous distribution characterized by degrees of freedom. F- Statistic is defined as the ratio of two chi-square distributions, and a specific F- distribution is described by the degree of freedom (d.f) for the numerator chi-square and the degree of freedom for the denominator chi-square. Therefore, a pair of degrees of freedom, numerator d.f and denominator d.f describes this distribution. This means that this distribution is identified by the two parameters, the numerator and denominator degrees of freedom.

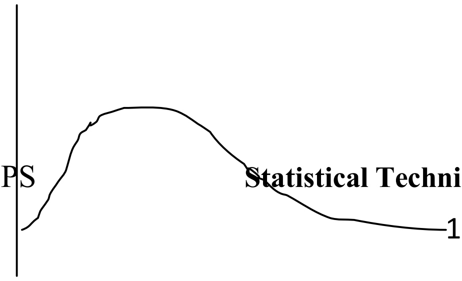
The curves for the F- distribution when the degrees of freedom for the numerator and denominator varies are shown below

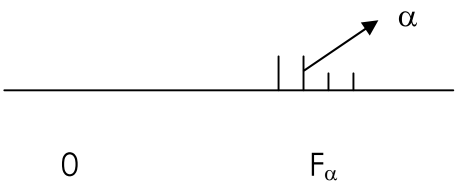


Since this distribution is the ratio of two sum of squares, it is always positive and is bounded by zero from below. Therefore, it is one- tailed test. The F- distribution is a right- skewed or positively skewed distribution. The F- distribution is most commonly used in Analysis of variance (ANOVA).

The F-distribution is a right- skewed distribution as shown in the curve below.

The F- distribution





We reject the null hypothesis at the given level of significance, α, and accept the alternative hypothesis when the observed value or calculated value of F exceeds F_{α}. Which means the critical region is $F > F_{\alpha}\,(k-1, N-k)$

Analysis of variance or ANOVA could be summarized by a table called ANOVA table. The following ANOVA table shows the general form of the table for analysis of variance.

ANOVA table

Sources of Variation	Sum of squares	Degree of freedom	Mean square	F
Treatment	SSB	k-1	MSB = $\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$
Within	SSW	N-k	MSW = $\frac{SSW}{N-k}$	
Total	SST			
				F= $\frac{MSB}{MSW}$

Example 8

The following table shows 4 samples from each of 4 different populations. Check if there is any significant difference between the means of the four populations at α = 0.05 and 0.01.

Solution

The required hypothesis is,

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Ha: Ho is not true or at least one of the means is different.

x_1	x	x	x
	2	3	4
1	1	1	3
2	3	2	2
1	2	2	1
2	2	2	1
Σ	Σ	Σ	Σ
x_1= 6	x_2= 8	x_3= 7	x_4= 7

Means $\overline{X}_1 = 3/2 \quad \overline{X}_2 = 2 \quad \overline{X}_3 = 7/4 \quad \overline{X}_4 = 7/4$

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ji} - \frac{T^2}{N} = (10+18+13+15) - \frac{(6+8+7+7)^2}{4+4+4+4} = 7$$

$$SSB = \sum_{j=1}^{k} \frac{T_j^{\ 2}}{n_j} - \frac{T^2}{N} = (\frac{6^2}{4} + \frac{8^2}{4} + \frac{7^2}{4} + \frac{7^2}{4}) - \frac{(6+8+7+7)^2}{4+4+4+4} = 0.50$$

SSW = SST – SSB = 7- 0.50 = 6.50

$d.f_B$ = k-1 =4-1 = 3

$d.f_W$ = N-k =16-4 =12

$$MSB = \frac{0.50}{3} = 0.17$$

$$\text{MSW} = \frac{6.50}{12} = 0.54$$

$$\text{F} = \frac{MSB}{MSW} = \frac{0.17}{0.54} = 0.31$$

For $\alpha = 0.05$, $F_{0.05}$ (3, 12) =3.49

For $\alpha = 0.01$, $F_{0.01}$ (3, 12) =5.95

Since $F_{\text{calculated}}$ is less than $F_{\text{tabulated}}$, we accept the null hypothesis at both α levels. Therefore, we can conclude that there is no significant difference between the four populations.

The analysis of variance done above can be summarized by the ANOVA table as follows.

ANOVA table

Sources of Variation	Sum of squares	Degree of freedom	Mean square	F
Treatment	SSB =0.5	k-1= 3	MSB = 0.17	$\frac{MSB}{MSW}$ =0.31
Within	SSW = 6.5	N-k= 12	MSW = 0.54	
Total	SST= 7			

5. CHI-SQUARE DISTRIBUTIONS

A Chi-square (x^2) distribution is a continuous distribution ordinarily derived as the sampling distribution of a sum of squares of independent standard normal variables.

Characteristics of the square distributions

1. It is a continuous distribution
2. The X^2 dist has a single parameter; the degree of freedom, v
3. The mean of the chi-square distribution is v
4. The variance of the chi-square distribution is 2v. Thus, the mean and Variance depend on the degree of freedom.
5. It is based on a comparison of the sample of observed data (results) with the expected results under the assumption that the null hypothesis is true.
6. It is a skewed distribution and only non negative values of the variable X^2 are possible. The skewness decreases as v increases; and when V increases without limit it approaches a normal distribution. It extends indefinitely in the positive direction
7. The area under the curve is 1.0

Having the above characteristics, X^2 dist has the following areas of application:

1. Testing for the equality of several proportions
2. Test for independence between two variables
3. Goodness of fit tests (Binomial, Normal, and Poisson)

TEST FOR THE INDEPENDENCE BETWEEN TWO VARIABLES

A X^2 test of independence is used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent. That is, the Chi-square distribution involves using sample data to test for the independence of two variables. The sample data are given in to a two way table called a contingency table. Because the X^2 test of independence uses a contingency table, the test is sometimes referred to as CONTINGENCY ANALYSIS (Contingency table test). The X^2 test is used to analyze, for example, the following cases:

- Whether employee absenteeism is independent of job classification
- Whether beer preference is independent of sex (gender)
- Whether favorite sport is independent of nationality.

- Whether type of financial investment is independent of geographic region.

The steps and procedures are similar with hypothesis testing.

Example:

1. A company planning a TV advertising campaign wants to determine which TV shows its target audience watches and thereby to know whether the choice of TV program an individual watches is independent of the individuals income. The table supporting this is shown below. Use a 5% level of significance and the null hypothesis.

Income	Type of Show			
	Basketball	**Movie**	**News**	**Total**
Low	143	70	37	**250**
Medium	90	67	43	**200**
High	17	13	20	**50**
Total	**250**	**150**	**100**	**500**

Solution

1. Ho: Choice of TV program an individual watches is independent of the individuals income

   Ha: Income and Choice of TV program are not independent

2. Decision rule

   $\alpha = 0.05$

   $v = (R-1)\ (C-1)$[7]*

   $= (3-1)\ (3-1)$

   $= 4$

   $X^2_{\alpha,\, v} = X^2_{0.05,\, 4} = 9.49$

   Reject Ho if sample X^2 is greater than 9.49

[7] For the RxC contingency table, the degrees of freedom are calculated as (R-1) (C-1). The degrees of freedom refers to the number of expected frequencies that can be chosen freely provided the row and column totals of expected frequencies are identical to the row and column totals of the observed frequency table.

3. Compute the test statistic

   In computing the test statistic our first task is to estimate the expected frequencies ($e_{ij} = r_i o_j/n$); where

   r_i = Observed freq total for row i.

   C_j = observed freq total for column j

   n = sample size

e_{11} = 250x250/500 = 125	e_{21} = 200 x 250/500 = 100	e_{31} = 50 x 250/500 =25
e_{12}= 250x150/500 = 75	e_{22} = 200 x 150/500=60	e_{32} = 50 x 150/500=15
e_{13} = 250x100/500 =50	e_{23} = 200x100/500 =40	e_{33} = 50x100/500 =10

A test of the null hypothesis that variables are independent of one another is based on the magnitudes of the differences between the observed frequencies and the expected frequencies. Large differences between o_{ij} and e_{ij} provide evidence that the null hypothesis is false.

The test is based on the following Chi-square test statistic.

$$\chi^2 = \sum \frac{(O_{ij} - eij)^2}{e_{ij}} \textbf{ Or } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where:

O_{ij} (f_o) = observed frequency for contingency table category in row i and column j.

E_{ij} (f_e) = expected frequency for contingency table in row i and column j.

$$\chi^2 = \frac{(143-125)^2}{125} + \frac{(70-75)^2}{75} + \frac{(37-50)^2}{50} + \frac{(90-100)^2}{100} + \frac{(67-60)^2}{60} + \frac{(17-25)^2}{25} + \frac{(13-15)^2}{15}$$
$$+ \frac{(20-10)^2}{10} + \frac{(43-40)^2}{40} = 21.174$$

4. Reject the null hypothesis that choice of TV program is independent from income level.

2. A human resource manager at EAGLE Inc. was interested in knowing whether the voluntary absence behavior of the firm's employees was independent of marital status. The employee files contained data on marital status and on voluntary absenteeism behavior for a sample of 500 employees is shown below.

	Marital Status				
Absence behavior	Married	Divorced	Widowed	Single	Total
Often absent	36	16	14	34	100
Seldom absent	64	34	20	82	200
Never absent	50	50	16	84	200
Total	150	100	50	200	500

Test the hypothesis that absence behavior is independent of marital status at a significance level of 1%.

Solution

1. Ho: Voluntary absence behavior is independent of marital status

   Ha: Voluntary absence behavior and marital status are dependent

2. $\alpha = 0.01$

   V = (R-1) (C-1)

   = (3-1) (4-1) = 6

   $X^2{}_{\alpha,v} = X^2{}_{0.01,6} = 16.81$

   Reject Ho if sample $X^2 > 16.81$

3. Sample X^2

Observed freq (f_o)	Expected Freq (f_e)	$(f_o-f_e)^2$	$\frac{(f_o-f_e)^2}{f_e}$
36	30	36	1.200
64	60	16	0.267
50	60	100	1.667
16	20	16	0.800
34	40	36	0.900
50	40	100	2.500
14	10	16	1.600
20	20	0	0.000
16	20	16	0.800
34	40	36	0.900
82	80	4	0.050
84	80	16	0.200
		$\sum\frac{(f_o-f_e)^2}{f_e}$	**10.883**

4. Do not reject Ho; because 10.883 is less than 16.81.

   Voluntary absence and marital status are independent.

3. The personnel administrator of XYZ Company provided the following data as an example of selection among 40 male and 40 female applicants for 12 open positions.

Applicant	Status		
	Selected	Not selected	Total
Male	7	33	40
Female	5	35	40
Total	12	68	80

a. The X^2 test of independence was suggested as a way of determining if the decision to hire 7 malls and females should be interpreted as having a selection bias in favor of males. Conduct the test of independence using α= 0.10. What is your conclusion?

b. Using the same test, would the decision to hire 8 malls and 4 females suggest concern for a selection bias?

c. How many males could be hired for the 12 open positions before the procedure would concern for a selection bias?

Solution

a.

1. Ho: There is no selection bias in favor of males. (Selection status and gender of the applicant are independent).

   Ha: There is selection bias in favor of males. (Selection status and gender of the applicant are not independent).

2. α = 0.1

   V = (R-1) (C-1)

   = (2-1) (2-1) = 1

   $X^2_{\alpha,v} = X^2_{0.1,1} = 2.71$

   Reject Ho if sample $X^2 > 2.71$

3. Sample X^2

Observed freq (f_o)	Expected Freq (f_e)	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
7	6	1	0.1667
33	34	1	0.0294
5	6	1	0.1667
35	34	1	0.0294
$\sum \frac{(f_o - f_e)^2}{f_e}$			**0.3922**

4. Do not reject Ho; because 0.392 is less than 2.71.

   There is no selection bias in favor of male applicants.

b.

1. Ho: There is no selection bias in favor of males. (Selection status and gender of the applicant are independent).

   Ha: There is selection bias in favor of males. (Selection status and gender of the applicant are not independent).

2. $\alpha = 0.1$

   V = (R-1) (C-1)

   = (2-1) (2-1) = 1

   $X^2{}_{\alpha,v} = X^2{}_{0.1,1} = 2.71$

   Reject Ho if sample $X^2 > 2.71$

3. Sample X^2

Observed freq (f_o)	Expected Freq (f_e)	(f_o-f_e)^2	$\frac{(f_o - f_e)^2}{f_e}$
8	6	4	0.6667
32	34	4	0.1176
4	6	4	0.6667
36	34	4	0.1176
		$\sum \frac{(f_o - f_e)^2}{f_e}$	**1.5686**

4. Do not reject Ho; because 1.569 is less than 2.71.

   There is no selection bias in favor of male applicants.

c. There is no shortcut method to answer this question. Therefore, lets try by increasing the number of male applicants who are accepted and decreasing the number of female applicants who are females.

1. Ho: There is no selection bias in favor of males. (Selection status and gender of the applicant are independent).

   Ha: There is selection bias in favor of males. (Selection status and gender of the applicant are not independent).

2. $\alpha = 0.1$

V = (R-1) (C-1)

= (2-1) (2-1) = 1

$X^2_{\alpha,v} = X^2_{0.1,1} = 2.71$

Reject Ho if sample $X^2 > 2.71$

3. Sample X^2

Observed freq (f_o)	Expected Freq (f_e)	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
9	6	9	1.5000
31	34	9	0.2647
3	6	9	1.5000
37	34	9	0.2647
		$\sum \frac{(f_o - f_e)^2}{f_e}$	**3.5294**

4. Reject Ho; because 3.5294 is less than 2.71.

Therefore, 8 male and 4 female applicants must be hired for the 12 open positions so as to avoid selection bias in favor of males.

The Chi-square test for independence is useful in helping to determine whether a relationship exists between two variables, but it does not enable us to estimate or predict the values of one variable based on the value of the other. If it is determined that a dependence does exist between two quantitative variables, then the techniques of regression analysis are useful in helping to find a mathematical formula that expresses the nature of mathematical relationship.

Small expected frequencies can lead to inordinately large chi-square values with the chi-square test of independence. Hence contingency tables should not be used with expected cell values of less than 5. One way to avoid small expected values is to combine columns or rows whenever possible and whenever doing so makes sense.

TESTING FOR THE EQUALITY OF SEVERAL PROPORTIONS

Testing for the equality of several proportions emphasizes on whether several proportions are equal or not; and hence the null hypothesis takes the following form:

Ho: $P_1 = P_{10}$; $P_2 = P_{2O}$; P3 = P_{30}; --- $P_k = P_{kO}$; and the alternative hypothesis take the following form:

Ha: The population proportions are not equal to the hypothesized values

The degree of freedom is determined as V= K-1; where K refers to the number of proportions and all expected cell values must be greater than or equal to 5.

Example:

1. In the business credit institution industry the accounts receivable for companies are classified as being "current," "moderately late," "very late," and "uncollectible." Industry figure show that the ratio of these four classes is 9: 3: 3: 1. ENDURANCE firm has 800 accounts receivable, with 439, 168, 133, and 60 falling in each class. Are these proportions in agreement with the industry ratio? Let α=0.05.

Solution

1. Ho: $P_1 = 9/16$; $P_2 = 3/16$; $P_3 = 3/16$; $P_4 = 1/16$

   Ha: One or more of the proportions are not equal to the proportions given in the null hypothesis.

2. $\alpha = 0.05$

   v =K - 1 = 4-1 = 3

   $X^2_{\alpha, v} = X^2_{0.05, 3} = 7.81$

   Reject Ho if sample $X^2 > 7.81$

3. Test Statistics (Sample χ^2)

Class	Observed freq (f_o)	Expected Freq ($f_e = np_i$)	$(f_o\text{-}f_e)^2$	$\frac{(f_o - f_e}{f_e}$
Current	439	450	121	0.269
Moderately late	168	150	324	2.160
Very late	133	150	289	1.927
Uncollectible	60	50	100	2.000
			$\sum \frac{(f_o - f_e)^2}{f_e}$	**6.356**

4. Do net reject Ho.

2. ETHIO Plastic Factory sells its products in three primary colors: Red, blue, and yellow. The marketing manager feels that customers have no color preference for the product. To test this hypothesis the manager set up a test in which 120 purchases were given equal opportunity to buy the product in each of the three colors. The results were that 60 bought red, 20 bought blue, and 40 bought yellow. Test the marketing manager's null hypothesis, using α=0.05.

Solution

1. Ho: People have no color preference with this product; $P_1 = P_2 = P_3 = 1/3$

   Ha: People have color preference with this product

2. $\alpha = 0.05$

   V= K-1 = 3 -1=2

   $X^2_{\alpha,v} = X^2_{0.05,2} = 5.99$

   Reject Ho if sample X^2 is greater than 5.99.

3. Sample χ^2

Class	Observed freq (f_o)	Expected Freq ($f_e = np_i$); p_i = 1/3[8]	$(f_o\text{-}f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Red	60	40	400	10.00
Blue	20	40	400	10.00
Yellow	40	40	0	0.00
			$\sum \frac{(f_o - f_e)^2}{f_e}$	**20.00**

[8] Since the null hypothesis states that there is no color preference, each of the three colors is preferred by one third of the purchases.

4. Reject Ho; because 20 > 5.99. This means that customers do have color preference. It appears that red is the most popular color and blue is the least popular.

3. Rating sciences, Inc., a TV program – rating service, surveyed 600 families where the television was turned on during the prime time on week nights. They found the following numbers of people turned to the various networks.

Name of the network	**Type**	**Number of viewers**
NBC	Commercial	**210**
CBS		**170**
ABC		**165**
PBS	Noncommercial	**55**
		600

a) Test the hypothesis that all four networks have the same proportion of viewers during this prime time period. Use $\alpha = 0.05$

b) Eliminate the results for PBS and repeat the test of hypothesis for the three commercial networks, using $\alpha = 0.05$

c) Test the hypothesis that each of the three major networks has 30% of the weeknight prime time market and PBS has 10% using $\alpha = 0.005$

Solution

a.

1. Ho: All of the four networks do have equal number of viewers; $P_1 = P_2 = P_3 = P_4 = 1/4$.

   Ha: All of the four networks do not have equal number of viewers.

2. $\alpha = 0.05$

   V= K-1 = 4 -1= 3

   $X^2_{\alpha,v} = X^2_{0.05,3} = 7.81$

   Reject Ho if sample X^2 is greater than 7.81

3. Sample χ^2

Class	Observed freq (f_o)	Expected Freq ($f_e = np_i$); p_i = 1/4[9]	$(f_o\text{-}f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
NBC	210	150	3,600	24.0000
CBS	170	150	400	2.6667
ABC	165	150	225	1.5000
PBS	55	150	9,025	60.1667
			$\sum\frac{(f_o - f_e)^2}{f_e}$	**88.3334**

4. Reject Ho; because 88.34 > 7.81.

b.

1. Ho: All of the three commercial networks do have equal number of viewers; $P_1 = P_2 = P_3 = 1/3$.

   Ha: All of the three commercial networks do not have equal number of viewers.

2. $\alpha = 0.05$
   V= K-1 = 3 -1= 2
   $X^2_{\alpha,v} = X^2_{0.05,2} = 5.99$
   Reject Ho if sample X^2 is greater than 5.99.
3. Sample χ^2

Class	Observed freq (f_o)	Expected Freq ($f_e = np_i$); p_i = 1/3	$(f_o\text{-}f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
NBC	210	181.67	802.60	4.4179
CBS	170	181.67	136.20	0.7497
ABC	165	181.67	277.90	1.5270
			$\sum\frac{(f_o - f_e)^2}{f_e}$	**6.6946**

[9] Since equal numbers of viewers are expected to watch each network, each of the four networks is watched by one fourth of the viewers.

4. Reject Ho; because 6.70 > 5.99

c.

1. Ho: $P_1 = P_2 = P_3 = 0.30$; $P_4 = 0.10$

   Ha: One or more of the proportions are not equal to the proportions given in the null hypothesis.

2. $\alpha = 0.005$ $X^2_{\alpha,v} = X^2_{0.05,3} = 7.81$

   V= K-1 = 4 -1= 3

   Reject Ho if sample X^2 is greater than 7.81

3. Sample χ^2

Class	Observed freq (f_o)	Expected Freq (f_e = np_i)	$(f_o-f_e)^2$	$\frac{(f_o-f_e)^2}{f_e}$
NBC	210	180	900	5.00
CBS	170	180	100	0.55
ABC	165	180	225	1.25
PBS	55	60	25	0.42
			$\sum\frac{(f_o-f_e)^2}{f_e}$	**7.22**

4. Do not Reject Ho; because 7.22 < 7.81.

4. Suppose that three companies, A, B, and C, have recently conducted aggressive advertising campaigns in order to maintain and possibly increase their respective shares of the market for a particular product. The market shares prior to the campaigns were P_1=0.45 for company A, P_2 = 0.40 for company B, P_3= 0.13 for company C, and P_4 = 0.02 for other competitors. To determine if these market shares changed after the advertising campaigns, a marketing analyst solicited the preferences of a random sample of 200 customers of this product. Of these 200 customers 95 indicated a preference for company A's product, 85 preferred company B's precut, 18 preferred company C's product, and the remainder preferred one or another of the products distributed by the competitors conduct a test, at the 5% level of significance. If the market shares have changed from the levels what they were at before the advertising campaigns.

   Solution

1. Ho: $P_1 = 0.45$; $P_2 = 0.40$; $P_3 = 0.13$; $P_4 = 0.02$

   Ha: At least one Pi is not equal to its specified value

2. $\alpha = 0.05$

   V= K-1 = 3 – 1 = 2[10]

   $\chi^2_{\alpha,v} = \chi^2_{0.05,2} = 5.99$

   Reject Ho if sample $X^2 > 5.99$.

   Because of the above change Ho is restated as:

   Ho: $P_1 = 0.45$, $P_2=0.40$, $P_3 = 0.15$

   Ha: At least one pi is not equal to its specified value

3. Sample X^2

Company	Observed freq (f_o)	Expected freq (f_e)	$(f_o\text{-}f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
A	95	90	25	0.2778
B	85	80	25	0.125
Others	20	30	100	3.3333
			$\sum\frac{(f_o - f_e)^2}{f_e}$	**3.9236**

4. Do not reject Ho. There is no sufficient evidence at the 5% level of significance to conclude that the market shares have changed from the levels they were at before the advertising campaign.

[10] Expected frequency value for P_4 is less than 5 (200*0.02 = 4). So, we have to combine P_4 with one of other expected frequencies, say P_3, to obtain a combined expected frequency of 30 (200*0.15). it can also be combined with other expected frequency values.

6. SIMPLE LINEAR REGRESSION AND CORRELATION

INTRODUCTION

Regression analysis is a very powerful tool in the field of statistical analysis in predicting the value of one variable, given the value of another variable, when these two variables are related to each other.

There are many statistical investigations in which the main objective is to determine whether a relationship exists between two or more variables. If such a relationship can be expressed by a mathematical formula, we will then be able to use it for the purpose of making predictions. For example, measurements from meteorological data are used extensively to predict impact areas for missiles fired under various atmospheric conditions, agronomists predict the yields of farm crops based on the various concentrations of nitrogen, potassium, and phosphorus contained in the fertilizer, admission directors require various tests for entering freshmen in order to predict success in college, and so forth. The reliability of any predictions will, of course, depend on the strength of the relationship between the variables included in the formula, which is measured by correlation analysis.

At the end of this chapter you will be able to:

- State what simile linear regression and correlation are
- Draw the scatter diagram of a given pair of data
- Determine the least squares regression line
- Compute the sample and rank correlation coefficients and interpret them
- Understand what SPURIOUS correlation means
- Understand and appreciate the application of regression and correlation coefficients in business problems

Scientists, economists, population experts, businessmen, and others have always been concerned with the problems of prediction. A mathematical equation that allows us to predict values of one dependent variable from known values of one or more independent variables is called a regression equation. This term is derived from the original heredity studies made by Sir Francis Galton (1822-1911) in which he compared the heights of sons to the heights of fathers. Galton showed that the heights of sons of tall fathers over successive generations *regressed* toward the mean height of the population. In other words, sons of unusually tall fathers tend to be shorter than their fathers and sons of unusually short fathers tend to be taller than their fathers. Today the term *regression* is applied to all types of prediction problems and does not necessarily imply a regression toward the population mean.

In all such cases as, we are interested in:

i) knowing the nature of relationship between any two characteristics of one's interest, and

ii) getting a definite idea about the degree of that relationship.

using appropriate statistical methods, which the present chapter seeks to develop and elucidate, we can develop ways of getting and knowing the nature of relationship obtained by the two variables, say, X and Y.

This means we want to:

i) Bring out the nature and degree of relationship between any two variables.

ii) Measure the rate of change in one (the dependent) variable associated with a given change in the other (the independent)/variable.

iii) Evaluate the predictive strength of the relationship that obtains, and assessing the reliability of an estimate derived from that relationship.

Since these three issues are interrelated, regression and correlation, as two sides of a single basic process, consist of methods of examining the relationship between two variables. Regression is mainly concerned with bringing out the nature of relationship and using it to know the best approximate value of one variable corresponding to a known value of the other variable. Correlation, on the other hand, is concerned with quantifying the closeness of such relationship.

a. TYPES OF RELATIONSHIP

The relationship between any two variables may be *linear* or *non-linear.* It may be described by means of a straight line or a curve. If the relationship is best explained by a *straight line,* it is said to be linear. On the contrary, if it is described more appropriately by *a curve,* the relationship is said to be non-linear.

A linear relationship implies a constant absolute change in the dependent variable in response to unit changes in the independent variable. It is extensively used for its better predictive strength and greater reliability of an estimate based thereon. A non-linear relationship implies varying absolute change in the dependent variable with respect to changes in the independent variable. Its predictive value is limited and, consequently, the reliability of the estimate more doubtful.

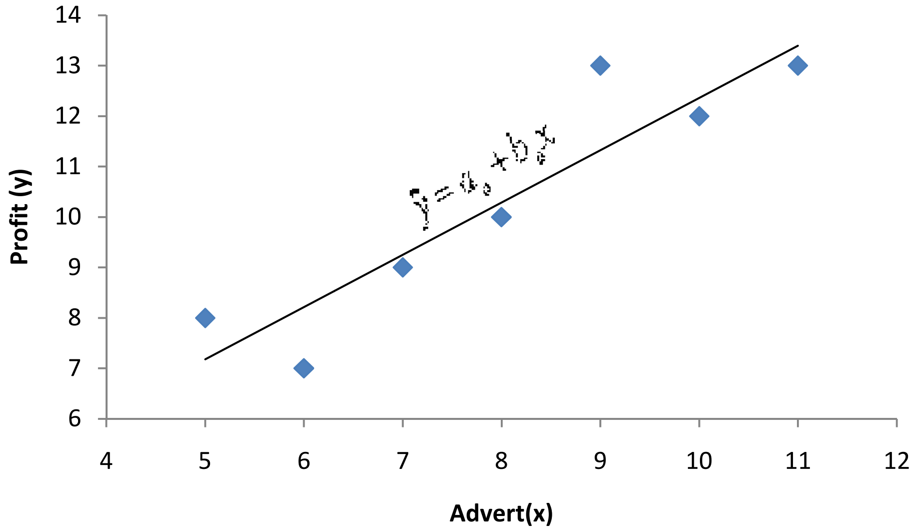
Here the discussion is limited to linear relationship between two variables only. Thus, we begin by developing methods of *linear regression,* which presupposes existence of paired data on two variables. This will provide answer to the first two issues listed above. Methods of measuring the degree of correlation in answer to the third will follow thereafter.

Since a linear relationship is a straight line relationship, linear regression deals with methods of fitting a straight line, often called the *regression line,* on a set of sample paired data on two variables.

Illustration: Suppose the following data, given in the table below, are on advertising of a product and the respective profits earned from each advertising period for the given product.

Advertising(x)	Profit(y)
5	8
6	7
7	9
8	10
9	13
10	12
11	13

The data of the table have been plotted in Figure to give a **scatter diagram**.



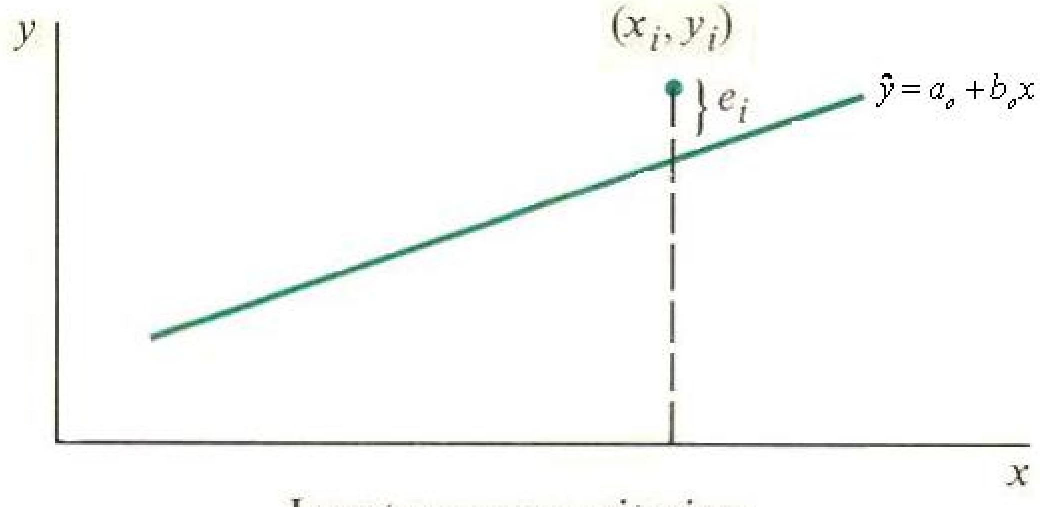
From an inspection of this scatter diagram, it is seen that the points follow closely a straight line, indicating that the two variables are to some extent linearly related. Once a reasonable linear relationship has been ascertained, we usually try to express this mathematically by a straight-line equation called the linear regression line. From elementary analytical geometry or high school algebra, we know that the slope-intercept form of a straight line can be written in the form

$$\hat{y} = a_o + b_o x \, .$$

Where the constants a_o and b_o represent the y intercept and slope, respectively. The symbol $\hat{y}$ is used to distinguish between the predicted or estimated value given by the regression line and an actual observed value y for some value of x. Such a regression line has been drawn on the above scatter diagram.

Once the point estimates a_o and b_o are determined from the sample data, the linear regression line can then be used to predict the value y corresponding to any given value x. Of course, the predicted value of y is a point estimate of y and as such is unlikely to hit right on. We hope that it will be close.

Once we have decided to use a linear regression equation, we face the problem of deriving computational formulas for determining the point estimates a_o and b_o from the available sample points. A procedure known as the method of least squares will be used. Of all possible lines that one might draw freehand on a scatter diagram, the least-squares procedure selects that particular line for which the sum of the squares of the vertical distances from the observed points to the line is as small as possible. Therefore, if we let e_i represent the vertical deviation from the i^{th} point to the regression line, as indicated in Figure below, the method of least squares yields formulas for calculating a_o and b_o so that the sum of the squares of these deviations is a minimum.



Least-squares criterion.

This sum of the squares of the deviations is often called the sum of squares of the errors about the regression line and is denoted by SSE. Thus if we are given a set of paired data *{(x*$_i$*, y*$_i$*); i* = 1, 2, ...,, *n}*, we shall find a_o and b_o so as to minimize;

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - a_o - b_o x_i)^2$$

The determination of a_o and b_o so as to minimize SSE is most easily accomplished by means of the method of least squares, which uses differential calculus. We omit the details here and state the final formulas.

Estimation of Parameters*:* Given the sample {(x_i, y_i); i = 1, 2, ..., n} the least-squares estimates of the parameters in the regression line of y on x

$$\hat{y} = a_o + b_o x$$

are obtained from the formulas

$$b_o = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad and \quad a_o = \bar{y} - b_o\bar{x}$$

Example: A garment manufacturer was interested in predicting the annual maintenance cost of sewing machines based upon the age of the machine. A sample of 12 machines revealed the following ages and maintenance costs during the previous year.

Machine	1	2	3	4	5	6	7	8	9	10	11	12
Age (years), x	6	5	4	7	5	7	8	9	6	8	4	3
Cost (inBirr), y	85	74	76	90	85	87	94	98	81	91	76	74

a) Determine the regression line of Cost (y) on Age (x)
b) Interpret the meaning of the regression coefficient (that is the slope) in this example.
c) Predict the maintenance cost of a machine that is 10 years old.
d) Draw a scatter diagram together with the linear regression line.

Solution: The regression coefficients b₀ can be estimated by the following formulae.

$$b_o = \frac{n\sum_{1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

And the y intercept a₀ can also be calculated by

$$a_o = \bar{y} - b_o\bar{x}$$

To get these values in the formulae, let us calculate them using a working table and then substitute into the formula as follows:

Machine	*Age(years), x*	*Cos(inBirr) y*	*xy*	x^2
1	6	85	510	36
2	5	74	370	25
3	4	76	304	16
4	7	90	630	49
5	5	85	425	25
6	7	87	609	49
7	8	94	752	64
8	9	98	882	81
9	6	81	486	36
10	8	91	728	64
11	4	65	260	16
12	3	40	120	9
Totals	72	966	6076	470

Therefore, $b_o = \dfrac{(12)(6076) - (72)(966)}{(12)(470) - (72)^2} = \dfrac{72912 - 69552}{5640 - 5184} = \dfrac{3360}{456} = 7.3684$

And $a_o = \bar{y} - b_o\bar{x} = \dfrac{966}{12} - 7.3684\left(\dfrac{72}{12}\right) = 36.2895$

a) The linear regression line is given by

$$\hat{y} = 36.2895 + 7.3684x$$

b) The regression coefficient 7.3684 is the amount of increase in maintenance cost in Birr per unit of increase in machine age.

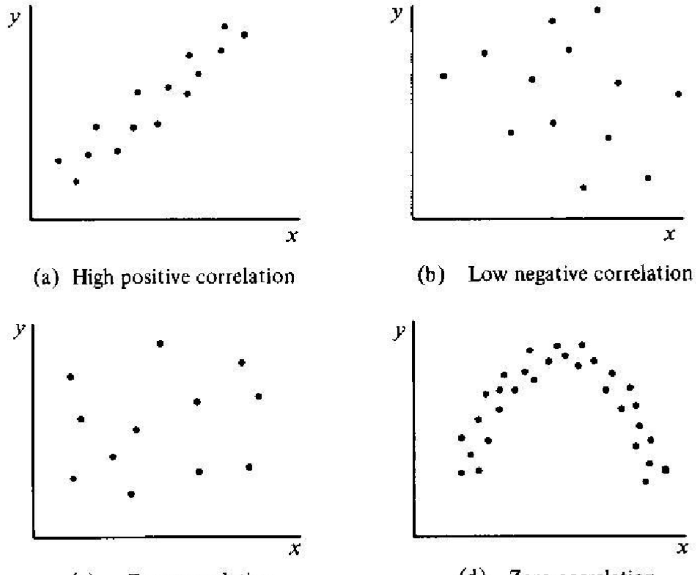
c) The cost of a machine that has age x=110 is Birr 36.2895 + 7.3684(10) = 109.74; found by substituting the age x=10 in the regression line.

b. Linear Correlation

In this section we shall consider the problem of measuring the relationship between two variables X and *Y* rather than predicting a value of *Y* from knowledge of the independent variable X, as in our study of linear regression. For example, if X represents the amount of money spent yearly on advertising by a retail merchandising firm and *Y* represents their total yearly sales, we might ask ourselves whether a decrease in the advertising budget is likely to be accompanied by a decrease in the yearly sales. On the other hand, if X represents the age of a used automobile and *Y* represents the retail book value of the automobile, we would expect large values of X to correspond to small values of *Y* and small values of X to correspond to large values of *Y*. Correlation analysis attempts to measure the strength of such relationships between two variables by means of a single number called a correlation coefficient.

We define a linear correlation coefficient to be a measure of the *linear* relationship between the two random variables X and *Y,* and denote it by *r.* That is, *r* measures the extent to which the points cluster about a straight line. Therefore, by constructing a scatter diagram for the *n* pairs of measurements *{(*x_i, y_i*):* i = 1, 2, ..., *n}* in our random sample (see Figure below), we are able to draw certain conclusions concerning *r.*

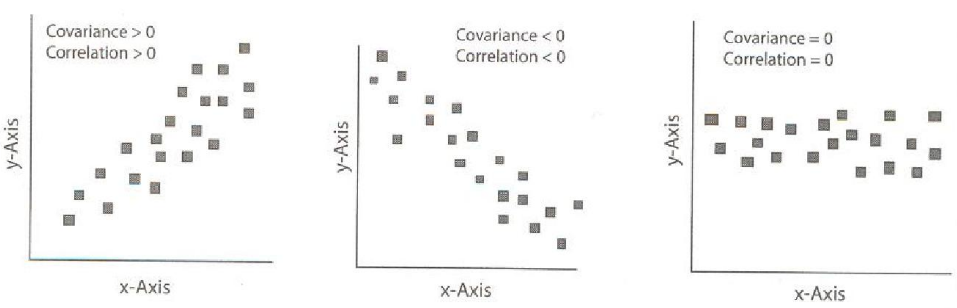
Scatter Diagrams showing various degrees of correlation.

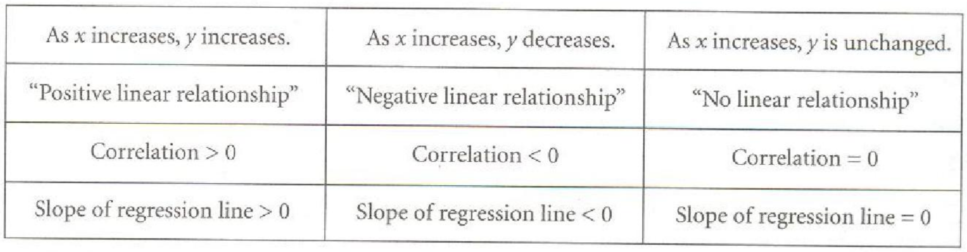


(a) High positive correlation (b) Low negative correlation

(c) Zero correlation (d) Zero correlation

Should the points follow closely a straight line of positive slope, we have a high positive correlation between the two variables. On the other hand, if the points follow closely a straight line of negative slope, we have a high negative correlation between the two variables. The correlation between the two variables decreases numerically as the scattering of points from a straight line increases. If the points follow a strictly random pattern as in the Figure below, we have zero correlation and conclude that no linear relationship exists between X and *Y.*





As *x* increases, *y* increases.	As *x* increases, *y* decreases.	As *x* increases, *y* is unchanged.
"Positive linear relationship"	"Negative linear relationship"	"No linear relationship"
Correlation > 0	Correlation < 0	Correlation = 0
Slope of regression line > 0	Slope of regression line < 0	Slope of regression line = 0

It is important to remember that the correlation coefficient between two variables is a measure of their linear relationship and a value of *r* = 0 implies a lack of linearity and not a lack of association. Hence, if a strong quadratic relationship exists between X and *Y as* indicated in Figure d above, we shall still obtain a zero correlation even though there is a strong nonlinear relationship.

The most widely used measure of linear correlation between two variables is called the Pearson product-moment correlation coefficient or simply the sample correlation coefficient. We now present a formula for computing the correlation coefficient in terms of the original measurements that is applicable even when the variables are measured in different units. Thus, if X and Y represent height and weight of an individual, respectively, the following formula will still provide a measure of the linear relationship between these two variables.

CORRELATION COEFFICIENT*:* The measure of linear relationship between two variables X and *Y* is estimated by the *sample correlation coefficient r,* where

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = b_o \frac{s_x}{s_y}$$

Where s_x and s_y are the sample standard deviations of X and Y.

The value of *r* ranges from - 1 to + 1. A value of r = - 1 will occur when all the sample points lie exactly on a straight line having a negative slope. If the points lie in a straight line having a positive slope we obtain a value of *r*=+1. Hence a perfect linear relationship exists between the values of X and *Y* in our sample when *r* $=\pm 1$. If *r* is close to + 1 or - 1, the linear relationship between the two variables is strong and we say that we have high correlation. However, if *r* is close to zero, the linear relationship between X and *Y* is weak or perhaps nonexistent.

One must be careful in interpreting *r* beyond what has been stated above. For example, values of *r* equal 0.4 and 0.8 only mean that we have two positive correlations, one somewhat stronger than the other. It is wrong to conclude that *r* = 0.8 indicates a linear relationship twice as strong as that indicated by the value *r* = 0.4. On the other hand, if we consider r^2, which is usually referred to as the sample **coefficient of determination**, we have a number that expresses the proportion of the total variation in the values of the variable *Y* that can be accounted for or explained by the linear relationship with the values of the variable X. Thus a correlation of *r* = 0.8 means that 0.64 or 64% of the total variation of the values of *Y* in our sample is accounted for by a linear relationship with the values of X.

Example:

In economics, the demand function for a product is often estimated by the price charged for such a product. The quantity of new *crying baby dolls* sold and the corresponding price charged at 10 stores of a large toy store chain for a one week period is shown in the following table.

Store	1	2	3	4	5	6	7	8	9	10
Quantity	225	250	280	290	310	340	350	350	360	380
Price	25	22	20	19	17	16	15	13	13	12

a) Compute and interpret the correlation coefficient.
b) Calculate the explained and total variation.
c) Calculate the coefficient of determination and define the relationship between price and quantity.

Solution: From the data we can obtain the following computations for calculating r.

Store	Quantity(y)	Price(x)	xy	x2	y2
1	225	25	5625	625	50625
2	250	22	5500	484	62500
3	280	20	5600	400	78400
4	290	19	5510	361	84100
5	310	17	5270	289	96100
6	340	16	5440	256	115600
7	350	15	5250	225	122500
8	350	13	4550	169	122500
9	360	13	4680	169	129600
10	380	12	4560	144	144400
Totals	**3135**	**172**	**51985**	**3122**	**1006325**

Based on the computations from the table, we find

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = \frac{(10)(51985) - (172)(3135)}{\sqrt{[(10)(3122) - (172^2)][(10)(1006325) - (3135^2)]}}$$

$$= -0.98783$$

A correlation coefficient of -0.98783 indicates a very good inverse linear relationship between X and *Y*. Since r^2 = 0.975802, we can say that 97.6% of the variation in the values of *Y* is accounted for by a linear relationship with X.

SUMMARY PROPERTIES OF r

1. The value of r is always between -1and 1(i.e. $-1 \leq r \leq 1$).
2. r >0 indicates positive association (positive linear relationship) between x and y.
3. r<0 indicates negative association (negative linear relationship) between x and y.

4. r= ± 1occurs only when perfect linear relationship exists between the two variables x and y, i.e., when the points in a scatter plot lie exactly along a straight line. Therefore, values of r close to 1 or –1 indicate that the points lie close to a straight line.
5. r is free of any unit of measurement. It is a dimensionless number and hence it is a coefficient.
6. r=0 indicates **no linear** relationship between x and y.
7. The linear association increases in strength as r moves away from 0 toward either -1 or 1.

c. RANK CORRELATION

Often in correlation analysis, information is not available in the form of numerical value like those we used in the examples. But if we can assign rankings of the items in each of the two variables we are studying, a rank correlation coefficient can be calculated. This is a measure of the correlation that exists between the two sets of ranks, a measure of the degree of association between the variables that we would not have been able to calculate otherwise.

A second reason for learning the method of rank correlation is to be able to simplify the process of computing a correlation coefficient from a very large set of data for each of two variables.

RANK CORRELATION COEFFICIENT

Correlation of ranks is applied either when quantification of some information is not possible or where exact magnitudes are not ascertainable. A possible answer in any such situation is to do ranking with reference to a particular characteristic. Ranks may be assigned either by two persons to a single characteristic or by a single person to two different characteristics.

RANK CORRELATION COEFFICIENT: A nonparametric measure of association between two variables X and Y is given by the rank correlation coefficient

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Where d_i is the difference between the ranks assigned to x_i and y_i , and is the number of pairs of data.

As ranks are assigned in any first *N* natural numbers, ranking done either way offers two rank series. This allows obtaining a measure of correlation between ranks, known as Spearman rank correlation coefficient. *Denoted by r_s* it measures the degree of relationship between two rank series.

When there are no ties among either set of measurements, the formula for r_s reduces to a much simpler equation, which is given as follows:

The value of r_s will usually be close to the value obtained by finding *r* based on **numerical** measurements and is interpreted in much the same way. As in r, the values of r_s range from - 1 to + 1. A value of + 1 or- 1 indicates perfect association between X and *Y,* the plus sign occurring for identical rankings and the minus sign occurring for reverse rankings. When r_s are close to zero, we would conclude that the variables are uncorrelated.

There are some advantages to using r_s rather than r:

a) We no longer assume the underlying relationship between X and *Y* to be linear and therefore, when the data possess a distinct curvilinear relationship, the rank correlation coefficient will likely be *more* reliable than the conventional measure.
b) A second advantage in using the rank correlation coefficient is the fact that no assumptions of normality are made concerning the distributions of X and *Y.*
c) The greatest advantage occurs when one is unable to make meaningful numerical measurements but nevertheless can establish rankings. Such is the case, for example, when different judges rank a group of individuals according to some attribute. The rank correlation coefficient can be used in this situation as a measure of the consistency of the two judges.

Remark: Where ranks are not given, i.e. when we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must consistently follow the same method in case of both variables.

<u>Example 1</u>

Two judges gave the following ranks (from highest to lowest) to eleven girls who contested in a beauty competition.

Girl A:	1	2	3	4	5	6	7	8	9	10	11
Rank by Judge A:	3	4	1	2	5	10	11	7	9	8	6
Rank by Judge B:	2	4	3	1	7	9	6	11	10	5	8

Is there an agreement between the independent rankings of the two judges?

Solution

Whether or not there is an agreement between the independent rankings of the two judges can be ascertained only by finding out the rank correlation between the ranks awarded by the two judges. The differences between the ranks and their squares are obtained as follows:

Girl A:	1	2	3	4	5	6	7	8	9	10	11	
Rank by Judge A:	3	4	1	2	5	10	11	7	9	8	6	
Rank by Judge B:	2	4	3	1	7	9	6	11	10	5	8	
Difference di	1	0	-2	1	-2	1	5	4	1	3	2	Total
Di^2	1	0	4	1	4	1	25	16	1	9	4	66

Since $\sum d_i^2 = 66$ and n=11

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 66}{11(11^2-1)}$$
$$= 1 - 0.30$$
$$= 0.70$$

The value of rank correlation r_s=0.70, which is quite high Hence it can be concluded that there is an agreement between judges with regard to the beauty of the girls.

❖ SELF ACTIVITY TEST 6-1

The ranks of 12 students according to their marks in Mathematics and statistics were as follows:

Student No.:	1	2	3	4	5	6	7	8	9	10	11	12
Mathematics:	5	2	1	6	8	11	12	4	3	9	7	10
Statistics:	4	3	2	7	6	9	10	5	1	11	8	12

Do students who are good in Mathematics also excel in statistics and vice versa?

Rank Correlation in the Case of Tied Ranks

There may be situations when the ranks assigned in two or more cases are the same. For example, suppose the ranking of 6 beauty candidates contestants in an interview conducted by an expert are 1, 3, 2, 3, 5, 6. Here out of the 6 ranks, two ranks, second and fourth, are the same (3, 3). This being a situation of tie of ranks, such ranks may be called tied ranks.

When tied ranks are obtained, we modify the tied ranks to the average of ranks that would have been assigned to them otherwise. Following this, the modified ranks will be 1, 3.5, 2, 3, 5, 6.

The above modification for tied ranks is needed to ensure that the sum of ranks remains unaffected whether or not tied ranks exist. Since the suggested modification affects the standard deviation of the concerned set of ranks, the factor $\sum d_i^2$ in the equation of r$_s$ is to be corrected as follows:

$$\sum d_c^2 = \sum d_i^2 + \frac{t^3 - t}{12}$$

In which $\sum d_c^2$ denotes the corrected $\sum d_i^2$, and $\frac{t^3 - t}{12}$ is the correction factor and t represents the number of tied ranks in set of ranks.

It may be noted that in obtaining $\sum d_c^2$, the correction factor $\frac{t^3 - t}{12}$ is to be added to $\sum d_i^2$ for every set of tied ranks in any paired rank data.

Example: In case of two sets of tied ranks, $\sum d_c^2 = \sum d_i^2 + \frac{t^3 - t}{12} + \frac{t^3 - t}{12}$

Accordingly, the coefficient of rank correlation r$_s$ in the case of tied ranks is computed as

$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{t^3 - t}{12}\right]}{n(n^2 - 1)}$$

Where the correction factor $\frac{t^3 - t}{12}$ is to be added for each set of tied ranks as in the example above.

Example: A selection board consisting of two experts for the post of gen manager in a company interviewed 10 candidates whom the two experts assign ranks as in Cols. (I) and (2) of Table. Since there are tied ranks occurring in two railings, the coefficient of rank correlation r_s is obtained by using the formula given above, as under

Computation of Coefficient of Rank Correlation r_s

	Original Ranks		Adjusted Ranks			
Sr. No	Expert I	Expert II	Expert I (Xi)	Expert II (Yi)	*di= (Xi -* Yi,)	di^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	7	8	7	8	-1	1
2	9	10	9	10	-1	1
3	2	4	2	4	-2	4
4	4	.6	4	6	-2	4
5	5	4	5.5	4	1.5	2.25
6	5	4	5.5	4	1.5	2.25
7	8	7	8	7	1	1
8	10	9	10	9	1	1
9	3	1	3	1	2	4
10	1	2	1	2	1	1
	54	55	55	55	$\sum di^2 = 21.5$	

Since rank 5 occurs twice and rank 4 occur thrice, the coefficient of correlation r_s is

$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{t^3 - t}{12}\right]}{n(n^2 - 1)}$$

$$r_s = 1 - \frac{6\left[21.5 + \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12}\right]}{10(10^2 - 1)} = 0.864$$

d. SPURIOUS CORRELATION

Correlation coefficient is not always meaningful unless the two variables are properly chosen. For example, if we find out the correlation between the total imports and the number of shoes produced per year, it would not lead to any conclusion. Because there is no cause and effect relationship between the total exports and number of persons migrate per year. Such a correlation is known as spurious or non-sense correlation between two variables. Another example could be the correlation between the number of customers who drink Coca Cola each year and the number of TV sets purchased each year. From this discussion, we come to the conclusion that it is more important to see the nature of the variables. If they have cause and effect relationship, and if one variable can be explained with the help of the other, then the correlation will have sense, otherwise it is spurious.

❖ SELF ACTIVITY TEST 6-2

Answer the questions below:

1. Linear correlation coefficients must lie between what two values? What value indicates "no linear correlation"? Does this mean no correlation at all?
2. Give one example of directly correlated variables
3. Give the purposes of linear regression
4. The input (x) and output (y) of a given business in millions of Birr is given below

Input (x)	2	4	3	1	6
Output (y)	5	6	4	3	7

a) Find the appropriate regression equation and estimate the amount of output for an input of 8 million Birr

b) Identify the coefficient of regression (the slope) you obtained in (a) and interpret it in line with the type of relation.

c) Compute the sample correlation coefficient , r, and interpret its value

d) Compute rank correlation coefficient for the same data and compare it with sample correlation coefficient.

5. The following sample data show the demand for a product in 000"s units and its price in birr in 6 different areas.

Price in birr	19	24	22	16	17	12
Demand in 000's unit	30	7	5	75	70	76

a) Fit a least squares line to predict the demand for the product in terms of its price
b) Estimate the demand if the product is priced Birr 22.
c) Calculate the sample and rank correlation coefficients and comment on the results.

6. Shown below are the scores of 20 male college professors on a measure of satisfaction with their work (y) and a measure of satisfaction with their life in general (x). Also shown are the means, S values, and r

Work (y)	Life (x)	Work (y)	Life (x)
44	57	49	61
53	64	39	48
47	60	60	69
50	59	46	58
36	49	48	61
49	57	43	55
54	62	57	68
40	53	47	56
52	66	51	62
48	55	45	55

$\bar{y} = 47.90 \qquad S_y = 5.90 \qquad \bar{x} = 58.75 \qquad S_x = 5.64 \qquad n = 20 \qquad$ r=0.94

a) Find the linear regression equation for predicting work satisfaction (y) from life satisfaction (x)

b) Predict work satisfaction (y) score from the x scores of 50 and 70?

7. Find out the Spearman's Rank Correlation Coefficient (r_s) and Pearson's product moment correlation coefficient (r) for the following data and give your comments regarding their results.

Maths score	29	32	53	47	45	32	70	45	70	53
English score	56	60	72	48	72	35	67	67	75	31

8. The management of a large furniture store would like to determine if there is any relationship the number of people entering the store on a given day and sales (in thousands Birr) for that day. The records were kept and a random sample of ten days was selected for the study. The summary results were as follows:

$$\sum X = 580, \quad \sum X^2 = 41,658, \quad \sum Y = 370, \quad \sum Y^2 = 17,206, \quad \sum XY = 11,494$$

Assuming a linear relationship:

a. Calculate linear regression coefficient.

b. Interpret the meaning of the regression coefficient.

c. Calculate the linear correlation coefficient and interpret it.

7. INTRODUCTION TO SPSS

SPSS: stands for Statistical Package for the Social Sciences. SPSS enables you to perform intense numerical calculations in a fraction of time. SPSS is frequently used in both academia and business environments. Much of SPSS's popularity within academia and various industries can be attributed to its capacity for managing data sets, a functionality that represents the bulk of the work done by professional statisticians. In addition, SPSS allows you to create, with great ease, beautiful graphics and tabular outputs. However, despite these significant conveniences, it is important to remember that no statistical software will relieve you of the need to think critically about the results any software package produces.

There are three basic tasks associated with data analysis:

A. type data into a computer, and organize and format the data so both SPSS and you can identify easily .
B. Tell SPSS what type of analysis you wish to conduct.
C. Be able to interpret what the SPSS output means.

1.1 Overview of SPSS for windows

SPSS for Windows consists of five different windows, each of which is associated with a particular SPSS file type. This document discusses the two windows most frequently used in analyzing data in SPSS, the *Data Editor* and the *Output Viewer* windows. In addition, the *Syntax Editor* and the use of SPSS command syntax is discussed briefly. The Data Editor is the window that is open at start-up and is used to enter and store data in a spreadsheet format. The Output Viewer opens automatically when you execute an analysis or create a graph using dialog box or command syntax to execute a procedure. The Output Viewer contains the results of all statistical analyses and graphical displays of data. The Syntax Editor is a text editor where you compose SPSS commands and submit them to the SPSS processor. All output from these commands will appear in the Output Viewer.

1.2 Starting SPSS

To start SPSS:

From the Windows Start menu choose:

Programs
SPSS for Windows

A small window will appear. This window has several choices with the following questions and options. What would you like to do?

- Run tutorial

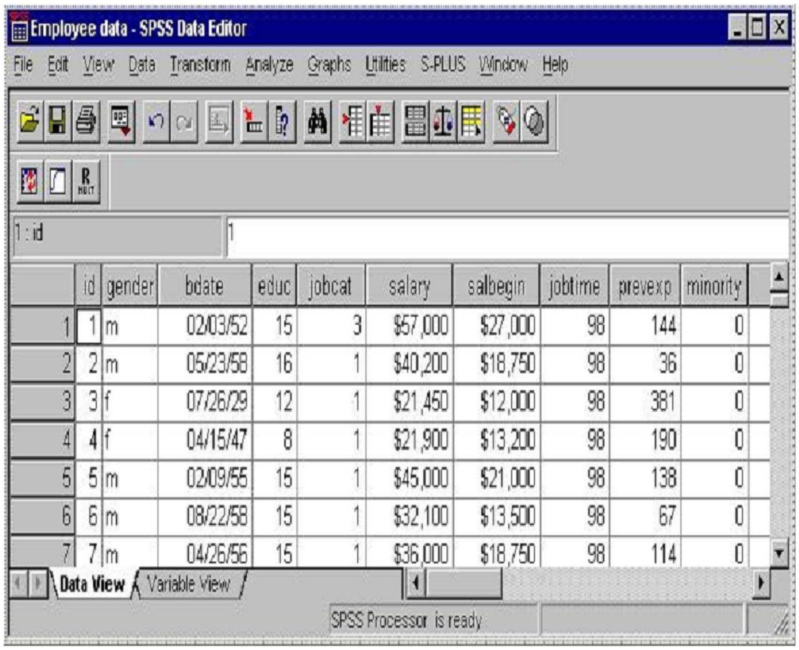
- Type in Data
- Run an existing query
- Create new query using an existing data base
- Open an existing data source

If you choose type in data, you will get **Data Editor Window.**

DATA EDITOR

SPSS data files are organized by cases **(rows)** and variables **(columns).**

The Data Editor displays the contents of the active data file. The information in the Data Editor consists of **variables** and **cases.** The employee data is located under the directory **C:\program files: \ SPSSEVAL.** If we open this data set**, t**he data editor window looks like the following**.**


Employee data - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities S-PLUS Window Help
1 : id
1
Data View
Variable View
SPSS Processor is ready

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
1	1	m	02/03/52	15	3	$57,000	$27,000	98	144	0
2	2	m	05/23/58	16	1	$40,200	$18,750	98	36	0
3	3	f	07/26/29	12	1	$21,450	$12,000	98	381	0
4	4	f	04/15/47	8	1	$21,900	$13,200	98	190	0
5	5	m	02/09/55	15	1	$45,000	$21,000	98	138	0
6	6	m	08/22/58	15	1	$32,100	$13,500	98	67	0
7	7	m	04/26/56	15	1	$36,000	$18,750	98	114	0

- **In Data View**, columns represent variables and rows represent cases (observations).
- **In Variable View**, each row is a variable, and each column is an attribute associated with that variable.

Variables are used to represent the different types of data that you have compiled. A common analogy is that of a survey. The response to each question on a survey is equivalent to a variable. Variables come in many different types, including numbers, strings, currency, and dates. The SPSS variable naming convention requires the following:

- The variable names should be up to eight characters or fewer.
- The variable name should not begin with any special characters such as numerals, comma, inequality symbols etc.
- The latest versions of SPSS can accept variable names with length greater than 8 characters.

Entering Data

In Data View, you enter your data just as you would in a spreadsheet program. You can move from cell to cell with the arrow keys on your keyboard or by clicking on the cell with the mouse. Once one case (row) is complete, begin entering another case at the beginning of the next row. You can delete a row of data by clicking on the row number at the far left and pushing the delete key on your keyboard. In a similar fashion, you delete a variable (column) by clicking on the variable name so that the entire column is highlighted and pushing the delete key.

In the steps that follow, we would see how to type in data by defining different variable types. Click the Variable View tab at the bottom of the Data Editor window.

Define the variables that are going to be used. In our case, let us consider three variables: namely age, marital status, and income.

In the first row of the first column, **type age.** In the second row, **type marital.**

In the third row, **type income**.

New variables are automatically given a numeric data type.

If you don't enter variable names, unique names are automatically created. However, these names are not descriptive and are not recommended for large data files.

Click the Data View tab to continue entering the data.

The names that you entered in Variable View are now the headings for the first three columns in Data View.

Begin entering data in the first row, starting at the first column.

In the age column, type 55.

In the marital column, type 1.

In the income column, type 72000.

Move the cursor to the first column of the second row to add the next subject's data.

In the age column, type 53.

In the marital column, type 0.

In the income column, type 153000.

Currently, the age and marital columns display decimal points, even though their values are intended to be integers. To hide the decimal points in these variables:

Click the Variable View tab at the bottom of the Data Editor window.

Select the Decimals column in the age row and type 0 to hide the decimal.

Select the Decimals column in the marital row and type 0 to hide the decimal.

Non-numeric data, such as strings of text, can also be entered into the Data Editor.

Click the Variable View tab at the bottom of the Data Editor window.

In the first cell of the first empty row, type sex for the variable name.

Click the Type cell.

Click the button in the Type cell to open the Variable Type dialog box.

Select String to specify the variable type.

Click OK to save your changes and return to the Data Editor.

In addition to defining data types, you can also define descriptive variable and value labels for variable names and data values. These descriptive labels are used in statistical reports and charts. Labels can be up to **256 characters long**. These labels are used in your output to identify the different variables.

Click the Variable View tab at the bottom of the **Data Editor window**.

In the Label column of the **age row,** type Respondent's Age.

In the Label column of the marital row, type Marital Status.

In the Label column of the income row, type Household Income.

In the Label column of the sex row, type Gender.

Adding a **Variable Label:** Click the Variable View tab at the bottom of the Data Editor window. In the Label column of the age row, type Respondent's Age. In the Label

column of the **marital row,** type Marital Status. In the Label column of the income row, type Household Income. In the Label column of the sex row, type Gender.
The Type column displays the current data type for each variable. The most common are numeric and string, but many other formats are supported.

In the current data file, the income variable is defined as a numeric type.
Click the Type cell for the income row, and then click the button to open the Variable Type dialog box.

Select **Dollar in the Variable Type** dialog box. The formatting options for the currently selected data type are displayed. Select the format of this currency. For this example, select $###,###,###.
Click OK to save your changes.
Value labels provide a method for mapping your variable values to a string label. In the case of this example, there are two acceptable values for the **marital variable**.
A value of "0" means that the subject is single and a value of "1 "means that he or she is married.

Click the values cell for the marital row, and then click the button to open the Value Labels dialog box.

The **value** is the actual numeric value.

The **value label** is the string label applied to the specified numeric value.

Type "0" in the value field.

Type "Single" in the Value Label field.

Click Add to add this label to the list.

Repeat the process, this time typing 1 in the value field and Married in the Value Label field. Click Add, and then click OK to save your changes and return to the Data Editor.

These labels can also be displayed in **Data View**, which can help to make your data more readable.

Click the Data View tab at the bottom of the Data Editor window.

From the menus choose:

View
Value Labels

The labels are now displayed in a list when you enter values in the **Data Editor**. This has the benefit of suggesting a valid response and providing a more descriptive answer.

Adding Value Labels for String Variables.

String variables may require value labels as well. For example, your data may use single letters, M or F, to identify the sex of the subject.

Value labels can be used to specify that M stands for Male and F stands for Female.
Click the Variable View tab at the bottom of the Data Editor window.
Click the Values cell in the sex row, and then click the button to open the Value Labels dialog box.

Type F in the **value field**, and then type Female in the **Value Label field**.

Click Add to add this label to your data file.

Repeat the process, this time typing M in the Value field and Male in the Value Label field. Click Add, and then click OK to save your changes and return to the Data Editor.

Because string values are **case sensitive**, you should make sure that you are consistent. A lowercase m is not the same as an uppercase M.

In a previous example, we choose to have value labels displayed rather than the actual data by selecting Value Labels from the View menu. You can use these values for data entry.

Click the **Data View tab** at the bottom of the **Data Editor window**. In the first row, select the cell for sex and select Male from the **drop-down list.**

In the second row, select the cell for sex and select Female from the drop-down list.

Only defined values are listed, which helps to ensure that the data entered are in a format that you expect.

Handling Missing Data

Missing or invalid data: are generally too common to ignore. Survey respondents may refuse to answer certain questions, may not know the answer, or may answer in an unexpected format.

If you don't take steps to filter or identify these data, your analysis may not provide accurate results.

For numeric data, empty data fields or fields containing invalid entries are handled by converting the fields to **system missing**, which is identifiable by a **single period**.

The reason a value is missing may be important to your analysis. For example, you may find it useful to distinguish between those who refused to answer a question and those who didn't answer a question because it was not applicable.

Click the Variable View tab at the bottom of the **Data Editor window**. Click the **Missing cell** in the age row, and then click the button to open the Missing Values dialog box. In this dialog box, you can specify up to **three distinct missing values**, or a range of values plus one additional **discrete value**.

Select Discrete missing values. Type 999 in the first text box and leave the other two empty.

Click OK to save your changes and return to the Data Editor. Now that the missing data value has been added, a label can be applied to that value. Click the Values cell in the age row, and then click the button to open the Value Labels dialog box.

Type 999 in the Value field. Type No Response in the Value Label field.

Click Add to add this label to your data file. Click OK to save your changes and return to the Data Editor.

Missing values for string variables are handled similarly to those for numeric values.

Unlike numeric values, empty fields in string variables are not designated as system missing. Rather, they are interpreted as an empty string.

Click the Variable View tab at the bottom of the Data Editor window.
Click the Missing cell in the sex row, and then click the button to open the Missing Values dialog box. Select Discrete missing values. Type NR in the first text box.

Missing values for string variables are **Case sensitive**. So, a value of “nr” is not treated as a missing value.

Click OK to save your changes and return to the Data Editor. Now you can add a label for the missing value. Click the Values cell in the sex row, and then click the button to open the Value Labels dialog box.

Type NR in the Value field. Type “No Response” in the Value Label field.

Click Add to add this label to your project. Click OK to save your changes and return to the Data Editor.

Once you've defined variable attributes for a variable, you can copy these attributes and apply them to other variables.

In Variable View, type **agewed** in the first cell of the first empty row. In the Label column, type Age Married. Click the Values cell in the age row.

From the menus choose:

Edit
Copy

Click the Values cell in the **agewed** row

From the menus choose:

Edit
Paste

The defined values from the age variable are now applied to the agewed variable. To apply the attribute to multiple variables, simply select multiple target cells (click and drag down the column).

When you paste the attribute, it is applied to all of the selected cells. New variables are automatically created if you paste the values into empty rows.

You can also copy all of the attributes from one variable to another. Click the row number in the marital row.

From the menus choose:

Edit
Copy

Click the row number of the first empty row.

From the menus choose:

Edit
Paste

All of the attributes of the marital variable are applied to the new variable.

For categorical (nominal, ordinal) data, **Define Variable Properties** can help you define value labels and other variable properties. Define Variable Properties:

- Scans the actual data values and lists all unique data values for each selected variable.
- Identifies unlabeled values and provides an "auto-label" feature.

- Provides the ability to copy defined value labels from another variable to the selected variable or from the selected variable to multiple additional variables.

This example uses the data file demo.sav . This data file already has defined value labels; so before we start, let's enter a value for which there **is no defined value label**:

In Data View of the Data Editor, click the first data cell for the variable ownpc (you may have to scroll to the right) and enter the value 99.

From the menus choose:

Data
Define Variable Properties...

In the initial Define Variable Properties dialog box, you select the nominal or ordinal variables for which you want to define value labels and/or other properties.

Since Define Variable Properties relies on actual values in the data file to help you make good choices, it needs to read the data file first. This can take some time if your data file contains a very large number of cases, so this dialog box also allows you to limit the number of cases to read, or scan.
Limiting the number of cases is not necessary for our sample data file. Even though it contains over 6,000 cases, it doesn't take very long to scan that many cases.

Drag and drop Owns computer [ownpc] through Owns VCR [ownvcr] into the Variables to Scan list.

You might notice that the measurement level icons for all of the selected variables indicate that they are scale variables, not categorical variables.
By default, all numeric variables are assigned the scale measurement level, even if the numeric values are actually just codes that represent categories.

All of the selected variables in this example are really categorical variables that use the numeric values 0 and 1 to stand for No and Yes, respectively--and one of the variable properties that we'll change with Define Variable Properties is the measurement level.
Click Continue

In the Scanned Variable List, select ownpc. The current level of measurement for the selected variable is scale. You can change the measurement level by selecting one from the drop-down list or you can let Define Variable Properties suggest a measurement level.

Click Suggest

Since the variable doesn't have very many different values and all of the scanned cases contain integer values, the proper measurement level is probably ordinal or nominal.

Select Ordinal and then click Continue.

The measurement level for the selected variable is now ordinal.

The Value Labels Grid displays all of the unique data values for the selected variable, any defined value labels for these values, and the number of times (count) each value occurs in the scanned cases.

The value that we entered, 99, is displayed in the grid. The count is only 1 because we changed the value for only one case, and the Label column is empty because we haven't defined a value label for 99 yet.
An X in the first column of the Scanned Variable List also indicates that the selected variable has at least one observed value without a defined value label.

In the Label column for the value of 99, enter No answer.

Then click (check) the box in the Missing column. This identifies the value 99 as user missing. Data values specified as user missing are flagged for special treatment and are excluded from most calculations.

Before we complete the job of modifying the variable properties for ownpc, let's apply the same measurement level, value labels, and missing values definitions to the other variables in the list. In the Copy Properties group, click To Other Variables.

In the Apply Labels and Level to dialog box, select **all of the variables** in the list, and then click Copy.

If you select any other variable in the list in the Define Variable Properties main dialog box now, you'll see that they are **all now ordinal variables**, with a value of 99 defined as user missing and a value label of No answer. Click OK to save all of the variable properties that you have defined. By doing so, we copied the property of the **ownpc** variable to the other five selected variables.

Exercise 1:

1. The following small data set consists of four variables namely, Agecat, gender, accid and pop.

Where agecat is a categorical variable created for age.
1= ‘ Under 21 ‘ 2= ‘ 21-25’ , 3. = ‘ 26-30’
Gender : 0 = ‘Male’ and 1= ‘Female’
Accid and Pop are numeric.
After defining these variable in a data editor window, enter the following data for the variables agecat, Gender, Accid and Pop respectively. Your data should appear as given below. Save the data set as trial1.sav.

1	1	57997	198522
2	1	57113	203200
3	1	54123	200744
1	0	63936	187791
2	0	64835	195714
3	0	66804	208239

2. Create a data set called Trial2.sav from the following data. The data set has the following variables :

I. Subject numeric width =2, right aligned and columns = 8
II. anxiety : numeric width =2, right aligned and columns = 8
III. tension : numeric width =2, right aligned and columns = 8
IV. Score : numeric width =2, right aligned and columns = 8
V. Trial : numeric width =2, right aligned and columns = 8

In addition there is no value levels for each of the above variables.

After completing the definition of the above variables, type in the following data into your data editor window so that your data appears as given below.

1	1	1	18	1
1	1	1	14	2
1	1	1	12	3
1	1	1	6	4
2	1	1	19	1
2	1	1	12	2
2	1	1	8	3
2	1	1	4	4
3	1	1	14	1
3	1	1	10	2
3	1	1	6	3
3	1	1	2	4
4	1	2	16	1
4	1	2	12	2
4	1	2	10	3
4	1	2	4	4
5	1	2	12	1
5	1	2	8	2
5	1	2	6	3
5	1	2	2	4
6	1	2	18	1
6	1	2	10	2
6	1	2	5	3
6	1	2	1	4
7	2	1	16	1
7	2	1	10	2
7	2	1	8	3
7	2	1	4	4
8	2	1	18	1

8	2	1	8	2
8	2	1	4	3
8	2	1	1	4
9	2	1	16	1
9	2	1	12	2
9	2	1	6	3
9	2	1	2	4
10	2	2	19	1
10	2	2	16	2
10	2	2	10	3
10	2	2	8	4
11	2	2	16	1
11	2	2	14	2
11	2	2	10	3
11	2	2	9	4
12	2	2	16	1
12	2	2	12	2
12	2	2	8	3

3. Given below is an example of a questionnaire, suppose you have information on several of such questionnaires. Prepare a data entry format that will help you to enter your data to SPSS.

Examples of questionnaire Design

Name ____
Age ____________ __________Sex ______________________________
City ____

Marital Status ↘ Married ↘ Single

Family Type ↘ Joint ↘ Nuclear

Family Members ↘ Adults ↘ Children

Family Income ↘ less than 10,000 ↘ 10, 000 to 15,000 ↘ 15,000-20,000
↘ More than 20, 000

Date:_____________

Place:_____________

1. What kind of food do you normally eat at home?
↘ North Indian ↘ South Indian ↘ Chinese ↘ Continental

2. How frequently do you eat out?
In a week ↘ once ↘ Twice ↘ Thrice ↘ More than thrice

3. You usually go out with:
↘ Family ↘ Friends ↘ Colleagues ↘ Others _______________

4. Is there any specific day when you go out?
↘ Weekdays ↘ Weekends ↘ Holidays ↘ Special occasions
↘ No specific days

5. You generally go out for
↘ Lunch ↘ Snacks ↘ Dinner ↘ Party/Picnics

6. Where do you usually go?
↘ Restaurant ↘ Chinese Joint ↘ Fast food joint ↘ others __________

7. Who decide on the place to go?

↘ Husband ↘ Wife ↘ Children ↘ Others ______________

8. How much do you spend on eating out (one time)?
↘ Below 200 ↘ 200-500 ↘ 500-800 ↘ More than 800

9. What did you normally order?

↘ Pizza ↘ Burgers ↘ Curries and Breads ↘ Soups ↘ Pasta

10. The price paid by you for the above is

10.1 Pizza: ↘ Very high ↘ A little bit high ↘ Just right

10.2 Burgers: ↘ Very high ↘ A little bit high ↘ Just right

10.3 Curries and Breads: ↘ Very high ↘ A little bit high ↘ Just right

10.4 Soups: ↘ Very high ↘ A little bit high ↘ Just right

10.5 Pasta: ↘ Very high ↘ A little bit high ↘ Just right

1.4. Data Importing From Microsoft Excel and ASCII files

Data can be directly entered in SPSS (as seen above), or a file containing data can be opened in the Data Editor. From the menu in the Data Editor window, choose the following menu options.

File
Open...

I. If the file you want to open is not an SPSS data file, you can often use the *Open* menu item to import that file directly into the Data Editor.

II. If a data file is not in a format that SPSS recognizes, then try using the software package in which the file was originally created to translate it into a format that can be imported into SPSS.

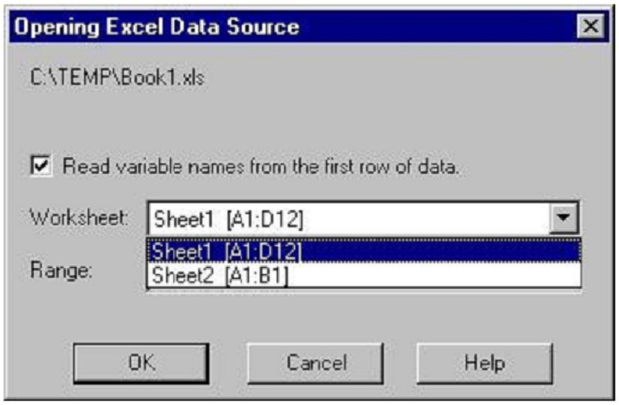
Importing Data From Excel Files

Data can be imported into SPSS from Microsoft Excel with relative ease. If you are working with a spreadsheet in another software package, you may want to save your data as an Excel file, then import it into SPSS.

To open an Excel file, select the following menu options from the menu in the Data Editor window in SPSS.

File
Open...

First, select the desired location on disk using the ***Look in*** **option**. Next, select Excel from the ***Files of type*** drop-down menu. The file you saved should now appear in the main box in the *Open File* dialog box. You can open it by double-clicking on it. You will see one more dialog box which appears as follows.



Opening Excel Data Source

C:\TEMP\Book1.xls

☑ Read variable names from the first row of data.

Worksheet: Sheet1 [A1:D12]

Sheet1 [A1:D12]
Sheet2 [A1:B1]

Range:

This dialog box allows you to select a spreadsheet from within the Excel Workbook.

The drop-down menu in the example shown above offers two sheets from which to choose.

As SPSS only operates on one spreadsheet at a time, you can only select one sheet from this menu.

This box also gives you the option of **reading variable names** from the Excel Workbook directly into SPSS.

Click on the *Read variable names* box to read in the first row of your spreadsheet as the variable names**.**

If the first row of your spreadsheet does indeed contain the names of your variables and you want to import them into SPSS, these variables names should conform to SPSS variable naming conventions (eight characters or fewer, not beginning with any special characters).

You should now see data in the Data Editor window. Check to make sure that all variables and cases were read correctly. Next, save your dataset in SPSS format by choosing the *Save* option in the *File* menu.

Example: Import an excel data set called **book1.xls into SPSS data editor window** from the desktop.

The procedure is as follows:

File
    Open... Data

After you select data you will see a window with the header "opens file". On the same window, select the **desktop** using the ***Look in*** **option**

Then select Excel (*.xls) from the file type drop down menu. Then another small window will appear. In this window you may see that there is only one worksheet. Now if the first row of the Book1.xls data set has variables names, then you select the option "Read **variable names from the first row of the data**". Subsequently, SPSS will consider the elements of the first row as variables. If the first of row of book1.xls is not variable names then leave the option unselected, then SPSS will understand elements of the first row as data values.

Importing data from ASCII files

Data are often stored in an ASCII file format, alternatively known as a text or flat file format. Typically, columns of data in an ASCII file are separated by a space, tab, comma, or some other character. To import text files to SPSS we have two wizards to consider:

I. Read Text Data: If you know that your data file is an ASCII file, then you can open the data file by opening the **Read Text Data** Wizard from the *File* menu. The Text Import Wizard will first prompt you to select a file to import. After you have selected a file, you will go through a series (about six steps) of dialog boxes that will provide you with several options for importing data.

Once we are through with importing of the data, we need to check for its accuracy. It is also necessary to save a copy of the dataset in SPSS format by selecting the *Save* or *Save As* options from the *File* menu.

II . Open Data. The second option to read an ASCII file to SPSS is by using

File open Data option.

File
Open... Data

After you select data you will see a dialogue box with the header "opens file". On the same window, select the desktop using the *Look in* option

Then select Text (*.txt) from the file type drop down menu. Select the file and click on open button. A serious of dialog boxes will follow.

Exercise: Suppose there is a text file named mychap1 on the desktop under the sub-directory training. Import this file to SPSS. Also name the first variable as X and the second as Y.

2. Modifying and organizing data

2.1 **Retrieving data: We can** retrieve **A. sav** data file from any directory in our personal computer or from floppy disk or any other removable disk.

To retrieve a data file from floppy disk, we select from application window menu bar

File

Open

Data

You will see the open data file dialogue box. Assuming the data type you want is on the floppy disk and has been saved previously by SPSS, open the drives drop down list and click on the icon for the drive a: All the files on drive A ending with **. SAV** extension will be listed in the files list. Click on the name of the file you want to retrieve, and it will appear in the file name box. Click on the Ok button on the right hand side of the dialogue box. The file will then be put into the **Data Editor Window**, and its name will be the title of that window.

Assume the data type you want is on the hard disk and has been saved previously by SPSS, under the directory program files. Open the drive's drop down list and click on the icon for program files: All the files under program files ending with". **SAV"** extension will be listed in the files list. Click on the name of the file you want to retrieve, and it will appear in the file name box. Click on the Ok button on the right hand side of the dialogue box. The file will then be put into the **Data Editor Window**, and its name will be the title of that window.

2.2. Inserting cases and variables

You may want to add new variables or cases to an existing dataset. The Data Editor provides menu options that allow you to do that. For example, you may want to add data about participants' ages to an existing dataset.

To insert a new variable, **click on the variable name** to select the column in which the variable is to be inserted.

To insert a case, select the row in which the case is to be added by clicking on the row's number. Clicking on either the row's number or the column's name will result in that row or column is being highlighted. Next, use the insert options available in the *Data* menu in the Data Editor:

Data

Insert Variable

Insert case

If a row has been selected, choose *Insert **Case*** from the *Data* menu; if a column has been selected, choose, ***Insert Variable***. This will produce an empty row or column in the highlighted area of the Data Editor. The existing cases and variables will be shifted down and to the right respectively.

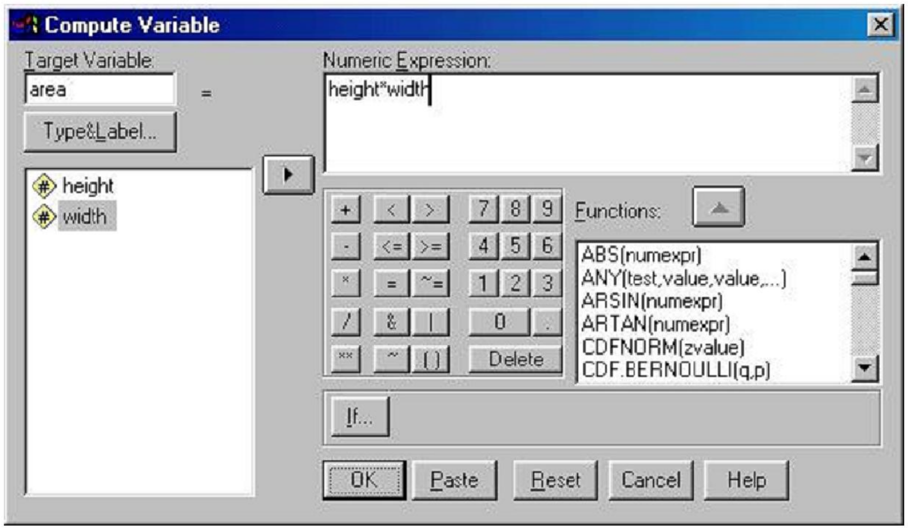
Deleting cases or variables

You may want to delete cases or variables from a dataset. To do that, select a row or column by highlighting as described above. Next, use the **Delete** key to delete the highlighted area. Or you can use the ***Delete*** option in the ***Edit*** menu to do it.

2.3.1 Transforming Variables with the Compute Command

In the Data Editor, you can use the **COMPUTE** or the **RECODE** command to create new variables from existing variables

The COMPUTE option allows you to arithmetically combine or alter variables and place the resulting value under a new variable name. As an example, to calculate the area of shapes based on their height and width, you compute a new variable "area" by multiplying "height" and "width" with one another. See below.



Compute Variable

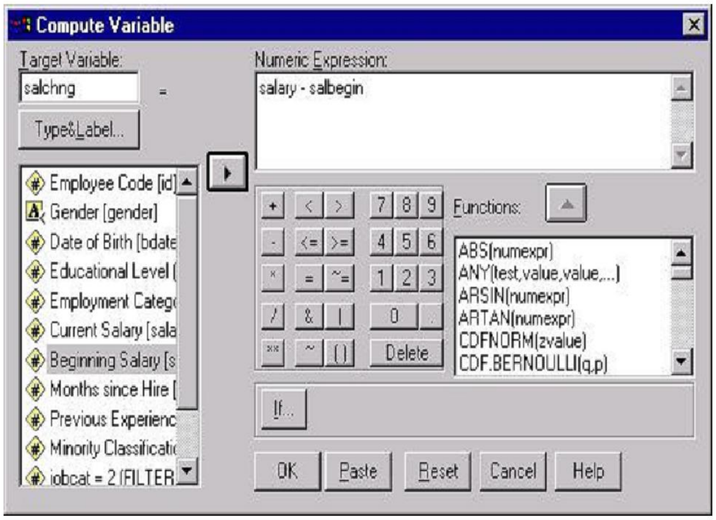
Target Variable: area =

Numeric Expression: height*width

The new variable created is area. This is specified under target variable. This target variable is the product of the two existing variables height and width.

Another example may be a dataset that contained employees' salaries in terms of their beginning and current salaries. Our interest is on the difference between starting salary and present salary. A new variable could be computed by subtracting the starting salary from the present salary. See the dialogue box below.

Transform
  Compute...



Compute Variable

Target Variable: salchng

Numeric Expression: salary - salbegin

In other situations, you may also want to transform an existing variable. For example, if data were entered as months of experience and you wanted to analyze data in terms of years on the job, then you could re-compute that variable to represent experience on the job in numbers of years by dividing number of months on the job by 12.

2.4 Transforming Variables with the Recode Command

The RECODE option allows you to create discrete categories from continuous variables. As an example, you may want to change the height variable where values can range from 0 to over 100 into a variable that only contains the categories tall, medium, and short. We have to pass through the following steps.

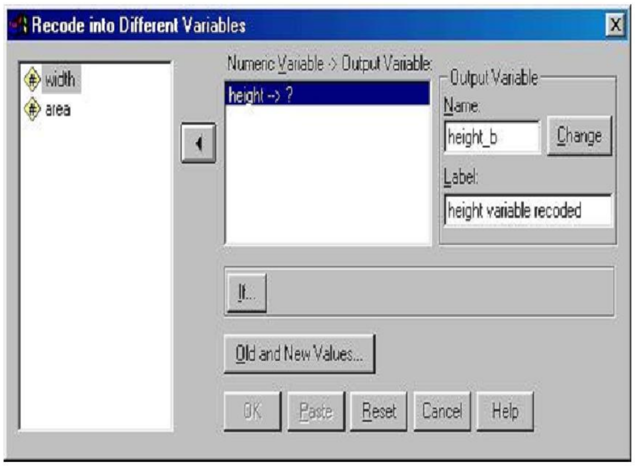
I. Select Transform/Recode/Into Different Variables.

II. A list of variables in the active data set will appear. Select the variable you wish to change by clicking once on the variable name and clicking the arrow button.

III. Click the output box and enter a new variable name (8 characters maximum) and click Change.

See the figure below. The variable to be recoded is the height.

NOTE: In dialog boxes that are used for mathematical or statistical operations, only those variables that you defined as numeric will be displayed. String variables will not be displayed in the variable lists.



width
area

Numeric Variable -> Output Variable:
height --> ?

Name: height_b

Label: height variable recoded

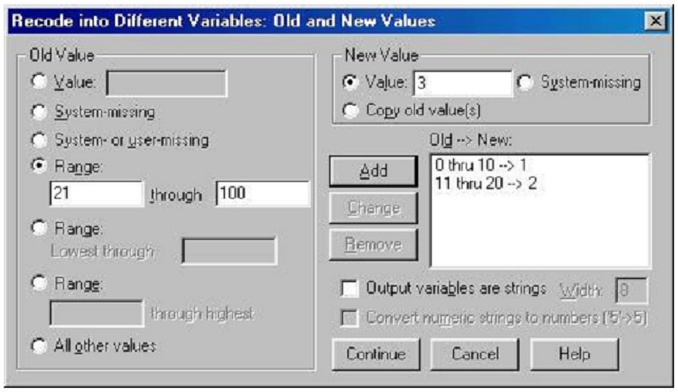
Now the variable height_b is the new variable that will be obtained after recoding. The value label for the new variable is "Height variable recoded".

IV. Select **OLD AND NEW VALUES.** This box presents several recoding options. You identify one value or a range of values from the old variable and indicate how these values will be coded in the new variable.

V. After identifying one value category or range, enter the value for the new variable in the **New Value** box. In our example, the old values might be 0 through 10, and the new value might be 1 (the value label for 1 would be "short", for 2 "medium", for 3 "tall").

VI. Click **ADD** and repeat the process until each value of the new variable is properly defined .

(See Figure Below) . Recode: Old and new values



● Range: 21 through 100

● Value: 3

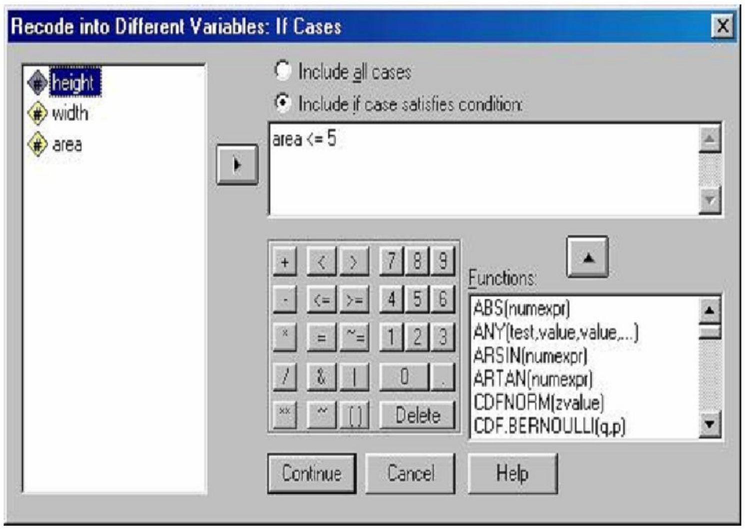
Old --> New:
0 thru 10 --> 1
11 thru 20 --> 2

Caution: You also have the option of recoding a variable into the same name. If you did this in the height example, the working data file would change all height data to the three categories (a value of 1 for "short"), 2 ("for medium", or 3 for "tall"). If you save this file with the same name, you *will lose all of the original height data.* **The best way to avoid this is to always use the recode option that creates a different variable**. Saving the data file keeps the original height data intact while adding the new categorized variable to the data set for future use.

Using if statement in the Data Editor

IF statement is an option to use within the **compute or recode** command. You can choose to only recode values if one of your variables satisfies a condition of your choice. This condition, which is captured by means of the "IF" command, can be simple **(such as "if area=15)**. To create more sophisticated conditions, you can employ logical transformations using AND, OR, NOT. The procedure is as given below.

1. In the Compute and Recode dialog boxes click on the IF button.
2. The **Include If Case Satisfies Condition** dialog pops up (see the Figure below).
3. Select the variable of interest and click the arrow button.
4. Use the key pad provided in the dialog box or type in the appropriate completion of the IF statement.
5. When the IF statement is complete, click **CONTINUE.**



Recode into Different Variables: If Cases

○ Include all cases

● Include if case satisfies condition:

area <= 5

2.4.1 Banding Values

Banding: means taking two or more contiguous values and grouping them into the same category.

The data you start with may not always be organized in the most useful manner for your analysis or reporting needs. For example, you may want to:

- Create a categorical variable from a scale variable.
- Combine several response categories into a single category.
- Create a new variable that is the computed difference between two existing variables.
- Calculate the length of time between two dates.

Once again we use the data file demo.sav.

Several categorical variables in the data file demo.sav are, in fact, derived from scale variables in that data file. For example, the variable inccat is simply income grouped into four categories.

This categorical variable uses the integer values 1–4 to represent the following income categories: less than 25, 25–49, 50–74, and 75 or higher.

To create the categorical variable inccat:

From the menus in the Data Editor window choose:

Transform
Visual Bander...

In the initial Visual Bander dialog box, you select the scale and/or ordinal variables for which you want to create new, banded variables. Banding is taking two or more contiguous values and grouping them into the same category.

Since the Visual Bander relies on actual values in the data file to help you make good banding choices, it needs to read the data file first. Since this can take some time if your data file contains a large number of cases, this initial dialog box also allows you to limit the number of cases to read ("scan").
This is not necessary for our sample data file. Even though it contains more than 6,000 cases, it does not take long to scan that number of cases.

Drag and drop Household income in thousands [income] from the Variables list into the Variables to Band list, and then click Continue.

In the main Visual Bander dialog box, select Household income in thousands [income] in the Scanned Variable List.

A histogram displays the distribution of the selected variable (which in this case is highly skewed). Enter inccat2 for the new banded variable name and Income category (in thousands) for the variable label.

Click Make Cutpoints.

Select Equal Width Intervals.

Enter 25 for the first cut-point location, 3 for the number of cut-points, and 25 for the width.

The number of banded categories is one greater than the number of cut-points. So, in this example, the new banded variable will have four categories, with the first three categories each containing ranges of 25 (thousand) and the last one containing all values above the highest cut-point value of 75 (thousand).

Click Apply.

The values now displayed in the grid represent the defined cut-points, which are the upper endpoints of each category. Vertical lines in the histogram also indicate the locations of the cut-points.

By default, these cut-point values are included in the corresponding categories. For example, the first value of 25 would include all values less than or equal to 25.
But in this example, we want categories that correspond to less than 25, 25–49, 50–74, and 75 or higher.

In the Upper Endpoints group, select Excluded (<).

Then click Make Labels.

This automatically generates descriptive value labels for each category. Since the actual values assigned to the new banded variable are simply sequential integers starting with 1, the value labels can be very useful.

You can also manually enter or change cut-points and labels in the grid, change cut-point locations by dragging and dropping the cut-point lines in the histogram, and delete cut-points by dragging cut-point lines off of the histogram. **Click OK to create the new, banded variable.**

The new variable is displayed in the Data Editor. Since the variable is added to the end of the file, it is displayed in the far right column in Data View and in the last row in Variable View.

But in this example, we want categories that correspond to less than 25, 25–49, 50–74, and 75 or higher.

In the Upper Endpoints group, select Excluded (<).

Then click Make Labels.

This automatically generates descriptive value labels for each category. Since the actual values assigned to the new banded variable are simply sequential integers starting with 1, the value labels can be very useful.

Sorting Cases

Sorting cases allows you to organize rows of data in ascending or descending order on the basis of one or more variable. For instance consider once again the **Employee data set**. Suppose we are interested to sort the data based on the variable **"Jobcat"** which refers to the category of employment. The procedure for sorting will be as follows:

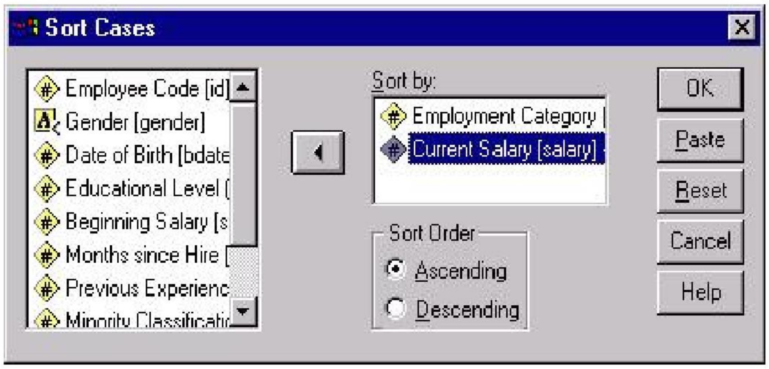
Data
Sort Cases...

A small dialog box with header **Sort Cases** will pop up. This dialogue box has few options. If you choose the ascending option in the dialogue box and click ok, you data will be sorted by Jobcat. All of the cases coded as job category 1 appear first in the dataset, followed by all of the cases that are labeled 2 and 3 respectively.

The data could also be sorted by more than one variable. For example, within job category, cases could be listed in order of their salary. Again we can choose

Data
Sort Cases...

In the small dialogue box select, select the variable **jobcat** followed by **salary**. The dialogue box comes into view as follows.



Sort by:
Employment Category (
Current Salary [salary]

Sort Order
● Ascending
○ Descending

To choose whether the data are sorted in ascending or descending order, select the appropriate button. Let us choose **ascending** so that the data are sorted in ascending order of magnitude with respect to the values of the selected variables. The hierarchy of such a sorting is determined by the order in which variables are entered in the *Sort by* box. Data are sorted by the first variable entered, and then sorting will take place by the next variable within that first variable. In our case *jobcat* was the first variable entered, followed by *salary*, the data would first be sorted by *job category*, and then, within each of the job categories, data would be sorted by *salary*.

Merging Files:

We can merge files into two different ways. The first option is "**add variables**" and the second is "**add cases**".

Add variables: The *Add Variables* adds new variables on the basis of variables that are common to both files. In this case, we need to have two data files. Each case in the one file corresponds to one case in the other file. In both files each case has an identifier, and the identifiers match across cases. We want to match up records by identifiers. First, we must sort the records in each file by the identifier. This can be done by clicking Data, Sort Cases, and then selecting the identifier into the "Sort by" box, OK.

Example, Given below we have a file containing **dads** and we have a file containing **faminc**. We would like to merge the files together so we have the **dads** observation on the same line with the **faminc** observation based on the key variable **famid**. The procedure to merge the two files is as follows:

- First sort both data sets by famid.
- Retrieve the **dads** data set into data editor window.
- Select

Data Merge files … add variables and select the file faminc.

dads

famid	name	inc
2	Art	22000
1	Bill	30000
3	Paul	25000

faminc

famid	faminc96	faminc97	faminc98
3	75000	76000	77000
1	40000	40500	41000
2	45000	45400	45800

After merging the **dads** and **faminc**, the data would look like the following.

famid	name	inc	faminc96	faminc97	faminc98
1	Bill	30000	40000	40500	41000
2	Art	22000	45000	45400	45800
3	Paul	25000	75000	76000	77000

Add variables (one to many)

The next example considers a **one to many** merge where one observation in one file may have multiple matching records in another file. Imagine that we had a file with **dads** like we saw in the previous example, and we had a file with **kids** where a dad could have more than one kid.

It is clear why this is called a **one to many** merge since we are matching **one dad** observation to **one or more (many) kids** observations. Remember that the **dads** file is the file with **one** observation, and the **kids** file is the one with **many** observations. Below, we create the data file for the **dads** and for the **kids**.

Dads data set

Famid	Name	Inc
2	Art	22000
1	Bill	30000
3	Paul	25000

Kids data set

Famid	Kid's name	birth	age	wt	sex
1	Beth	1	9	60	f
1	Bob	2	6	40	m
1	Barb	3	3	20	f
2	Andy	1	8	80	m
2	Al	2	6	50	m
2	Ann	3	2	20	f
3	Pete	1	6	60	m
3	Pam	2	4	40	f
3	Phil	3	2	20	m

To merge the two data sets, we follow the steps indicated below.

1. SORT the data set dads by famid and save that file and call it dads2
2. SORT the data set kids by famid and save that file as kids2
3. Retrieve the data set **kids2** to data editor window.
4. Select **data …merge files… add variables.**
5. From the dialogue box select the file dads2
6. Another dialogue box will appear. In this dialogue box we select the option " **match cases on key variables in sorted files".**
7. Select external file is keyed table and choose **famid** as key variable
8. **Click Ok.**

The Data Editor window will appear as given below.

```
FAMID KIDNAME    BIRTH      AGE       WT SEX NAME       INC
 1.00 Beth        1.00     9.00    60.00 f   Bill 30000.00
 1.00 Bob         2.00     6.00    40.00 m   Bill 30000.00
 1.00 Barb        3.00     3.00    20.00 f   Bill 30000.00
 2.00 Andy        1.00     8.00    80.00 m   Art  22000.00
 2.00 Al          2.00     6.00    50.00 m   Art  22000.00
 2.00 Ann         3.00     2.00    20.00 f   Art  22000.00
 3.00 Pete        1.00     6.00    60.00 m   Paul 25000.00
 3.00 Pam         2.00     4.00    40.00 f   Paul 25000.00
 3.00 Phil        3.00     2.00    20.00 m   Paul 25000.00
```

We can also retrieve the data set **dads2** to data editor window and perform steps 4 to 6 for the file kids2. This time you select **working file is keyed table** and choose **famid** as key variable. The data editor window will appear as given below.

FAMID	NAME	Inc	Kidname	BIRTH	AGE	WT	SEX
1	Bill	30000	Beth	1	9	60	f
1	Bill	30000	Bob	2	6	40	m
1	Bill	30000	Barb	3	3	20	f
2	Art	22000	Andy	1	8	80	m
2	Art	22000	Al	2	6	50	m
2	Art	22000	Ann	3	2	20	f
3	Paul	25000	Pete	1	6	60	m
3	Paul	25000	Pam	2	4	40	f
3	Paul	25000	Phil	3	2	20	m

Here the correct choice of keyed table can give us correct results.

The key difference between a **one to one** merge and a **one to many** merge is that you need to correctly identify the keyed table. That means we have to identify which file plays the role of **one (in one to many).** That file should be chosen as keyed table. In the above example the keyed table file is only dads2 but not kids2.

Merging files (add cases option)

The *Add Cases* option combines two files with different cases that have the same variables. To merge files in this option we should follow the following procedures.

data …merge files… add cases

All variables should be listed under the small window "**new working data file**". Click Ok to complete merging.

Exercise:

1. Merge two files addcas1 and addcas2 in the directory Desktop\Training r. using the add case option.

2. Merge two files addcas3 and addcas4 in the directory Desktop\Training r. and sort the newly merged file by ID.

3. Merge the file employee11 with the file Mean_salary from the directory Desktop\Training **r.** In this case the files should be matched on the basis of *jobcat. (this will be one to many).*

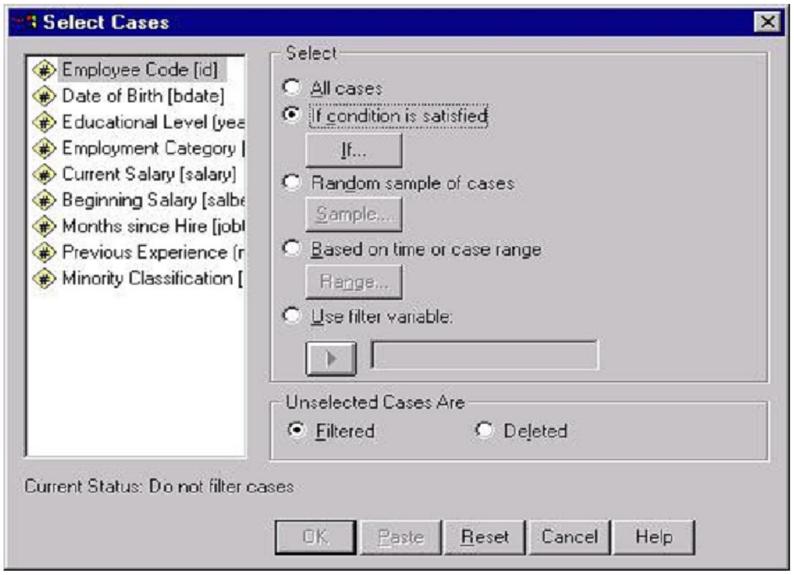
2.6. Keeping and dropping of cases

Selecting Cases

You can analyze a specific subset of your data by selecting only certain cases in which you are interested. For example, you may want to do a particular analysis on employees only if the employees have been with the company for greater than six years. This can be done by using the *Select Cases* menu option, which will either temporarily or permanently remove cases you didn't want from the dataset. The *Select Cases* option (or Alt+D+C) is available under the *Data* menu item:

Data
Select Cases...

Selecting this menu item will produce the following dialog box. This box contains a list of the variables in the active data file on the left and several options for selecting cases on the right.



Select Cases

Employee Code [id]
Date of Birth [bdate]
Educational Level [yea
Employment Category [
Current Salary [salary]
Beginning Salary [salbe
Months since Hire [jobt
Previous Experience [r
Minority Classification [

Select
○ All cases
● If condition is satisfied
○ Random sample of cases
○ Based on time or case range
○ Use filter variable:

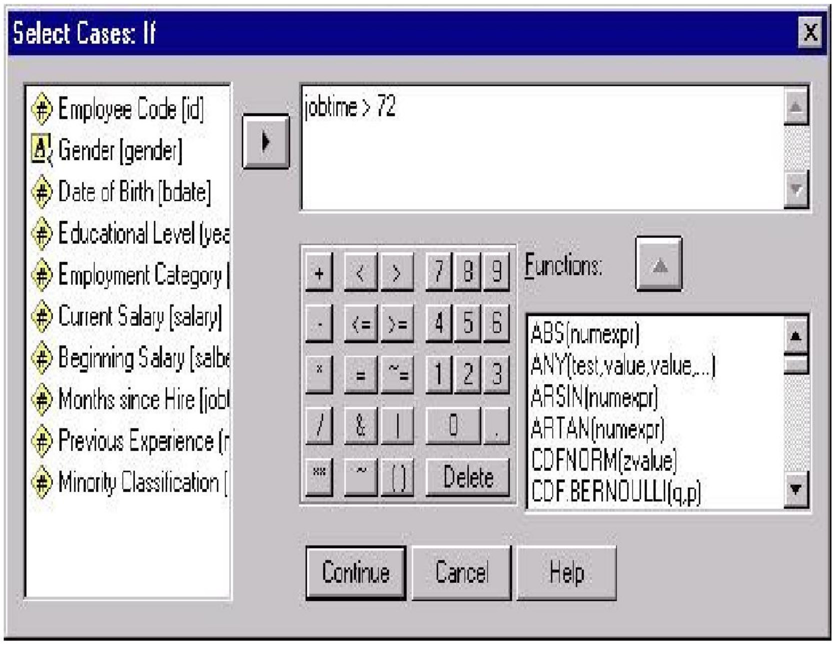
Unselected Cases Are
● Filtered ○ Deleted

Current Status: Do not filter cases

The portion of the dialog box labelled "***Unselected Cases Are*** " gives us the option of temporarily or permanently removing data from the dataset.

- If the "***Filtered***" option is selected, the selected cases will be removed from subsequent analyses until "All Cases" option reset.
- If the "***Deleted***" option is selected, the **unselected cases** will be removed from the working dataset. If the dataset is subsequently saved, these cases will be permanently deleted.

Selecting one of these options will produce a second dialog box that prompts us to a particular specification in which we are interested. For example, if we choose the "***If condition is satisfied***" option and clicking on the **If** button the results in a second dialog box, will appear as shown below.



jobtime > 72

The above example selects all of the cases in the dataset that meet a specific criterion: employees that have worked at the company for greater than six years (72 months) will be selected. After this selection has been made, subsequent analyses will use only this subset of the data. If you have chosen the *Filter* option in the previous dialog box, SPSS will indicate the inactive cases in the Data Editor by placing a slash over the row number. To select the entire dataset again, return to the *Select Cases* dialog box and select the *All Cases* option.

2.7 Collapsing and transposing Data

Collapsing data across observations

At times we might have data files that need to be collapsed to be useful to us. For Instance, you might have student data but we really want classroom data, or we might have weekly data but we are interested on monthly data, etc. Let us see how we can collapse data across kids to make family level data.

Aggregating Files

Aggregating files is one way of data manipulation procedure. The *Aggregate* procedure allows you to condense a dataset by collapsing the data on the basis of one or more variables. For example, to investigate the characteristics of people in the company on the basis of the amount of their education, you could collapse all of the variables you want to analyze into rows defined by the number of years of education. To access the dialog boxes for aggregating data, follow the following steps:

1. Select **Data** and then **AGGREGATE**

2. We will observe a dialogue box. This dialogue box has several options. These are as follows.

Break variable: The top box, labeled *Break Variable(s)*, contains the variable within which other variables are summarized. This is something like classification variable.

Aggregated Variables: contains the variables that will be collapsed.

Number of cases: This option allows us to save the number of cases that were collapsed at each level of the break variable.

Save: This has three different options. I) Add the aggregated variables to working data file II) Create new data file containing aggregated variables. III) Replace working data with aggregated variables only. We may choose one of the above three options depending on our interest.

Options for very large data sets: This has two options

- File is already sorted on break variable(s)
- Sort file before aggregating.

Example: Suppose we have a file containing information about the kids in three families. There is one record per kid. **Birth** is the order of birth (i.e., 1 is first), **age wt** and **sex** are the child's age, weight and sex respectively. This data is saved as kid3.sav file in the directory

desktop:\ training r . We will use this file for showing how to collapse data across observations. If we consider the **aggregate** command under the data menu we can collapse across all of the observations and make a single record with the average age of the kids. To do so we need to create a break variable const=1 using the compute command.

famid	Kidname	birth	Age	Wt	Sex
1	Bekele	1	9	60	f
1	Bogale	2	6	40	m
1	Barbie	3	3	20	f
2	Anteneh	1	8	80	m
2	Alemayehu	2	6	50	m
2	Abush	3	2	20	f
3	Chapie	1	6	60	m
3	Chuchu	2	4	40	f
3	Mamush	3	2	20	m

To collapse the above data, we follow the following steps:

- **Select** Data **and then** AGGREGATE**.** In the observed dialogue box, select const as break variable.
- Choose "**age**" for summaries of variables

- Choose add aggregated variables to working data file

The "**age_mean**" variable will be added to our working data. This is the mean age of all 9 children.

If we follow all of the above steps and change the last option to **"**Create new data file containing aggregated variables only**"**, we will have the following output saved as aggr.sav.

CONST	AVGAGE	N_Break
1.00	5.11	9

If we use "famid" as break variable**, the aggregate option** will the average age of the kids in the family**.** The following output will be obtained**.**

FAMID	AGE1
1.00	6.00
2.00	5.33
3.00	4.00

We can request averages for more than one variable. For instance, if we want to aggregate both age and weight by **famid** we can follow the following steps.

- **Select** Data **and then** AGGREGATE**.** In the observed dialogue box, select **Const** as break variable.
- Choose "**age**" and "**wt**" for summaries of variables
- Choose "Create new data file containing aggregated variables only".

The following output will be produced. The variable N_Break is the count of the number of kids in each family.

Famid	Age_mean	Wt_mean	N_Break
1	6.00	40.00	3
2	5.33	50.00	3
3	4.00	40.00	3

We can variable "**girls**" that counts the number of girls in the family, and "**boys"** that can help us count the number of boys in the family. You can also add a label after the new variable name. If you save the output in SPSS, you can see the labels in SPSS data editor after clicking on the "variable view" tab in the lower left corner of the editor window.

To have summary information which shows the number of boys and girls per family, we will follow the following procedure. We create two dummy variables **Sexdum1** for girls and **Sexdum2** for boys. The sum of **sexdum1** is the number of girls in the family. The sum of **sexdum2** is the number of boys in the family.

I) We recode sex into dumgirl=1 if sex=girl and dumgirl=0 if sex=m

II) We recode sex into dumboy =1 if sex=m, dumboy=0 if sex=f

III) We select **Data … Aggregate** option. At this step a dialogue box will appear. In this dialogue box, we select Break-variable = famid,

IV) Select dumgirl and dumboy for aggregated variables.

V) Below the aggregated variables we have two options 1. Function 2. Name and label. After selecting one of the variables to be aggregated choose the 'function' option. A new dialogue box will pop-up.

VI) From this new dialogue box, we select the function '**sum'** and **click continue** for both variables.

VII) Again below the aggregated variables, select the option Name and label. Change the name "dumgirl**_sum**" to **girl and** "dumboy_sum" to boy. You can also boy as number of boys and girl as number of boys.

VIII) Now click on the number cases box and change the name N_Brake to "NumKids".

Ix) Finally we have to choose the save option. If we choose the option "Create **new data file containing aggregated variables only**". , SPSS will save the file in the directory of our choice.

For instance, if we save our file in the directory **desktop\training r**, our file will be saved as SPSS file. Our results look like the following output.

FamId	Boys	girls	Numkid
1	1.00	2.00	3
2	2.00	1.00	3
3	2.00	1.00	3

Restructure Data : We use Restructure data wizard to restructure our data.

In the first dialog box, we select the type of restructuring that we want to do. Suppose, we the data that are arranged in groups of related columns. Our interest is to restructure these data into groups of rows in the new data file. Then we choose the option restructure selected variables into cases.

Example : Consider a small data set consisting of three variables as given below.

V1	V2	V3
8	63	82
9	62	87
10	64	89
12	66	85
15	67	86

The objective is then to restructure the above data into groups of rows in the new data file. In other words we want to convert the above data into one variable that has all the values of the three variables and one factor variable that indicate group. This procedure is known as the restructuring of variables to cases. The procedure is as follows.

- From the **data** menu select **restructure, the** dialogue box which says "Welcome **to the restructure Data wizard**" will appear.

- Choose the first option "Restructure **selected variables into cases**" and click next.
- Another dialogue box which says "**Variable to cases : Number of variable groups**" will appear. Choose the first option " One" and click next.
- Give the name of target variable call it " **all-inone"**
- Select all three variables (V1, V2 and V3) to variables to be transposed box and Click next.
- Another dialogue box which says **"Variable to cases: Create index variable**" will appear. Choose the first option " **One**" and click next.
- Another new dialogue box will appear here change the variable name "**Index**" to group. Click finish and see your restructured data. The data may appear as shown below.

Id	Group	All_inone
1	1	8
1	2	63
1	3	82
2	1	9
2	2	62
2	3	87
3	1	10
3	2	64
3	3	89
4	1	12
4	2	66
4	3	85
5	1	15
5	2	67
5	3	86

The variable ID stands for the row position in of the data before the data was restructured. We can also restructure the data from cases to variables.

For instance, consider the following small data set on age of Nurses and Doctors. The variable group 1 stands for nurses and 2 stands for doctors.

Id	Age	Group
1	23	1
2	25	1
3	26	1
4	35	1
5	42	1
6	22	1
1	60	2
2	36	2
3	29	2
4	56	2
5	32	2
6	54	2

The objective is to restructure the above age data into a data set having two separate variables for Nurses and Doctors. To do so, we follow the following procedure.

- From the **data** menu we select **restructure**
- From the dialogue box we select " **Restructure selected cases to variables** "
- We select Id for Identifier variable
- We select group for Index variable and click next and respond to the dialogue box that will appear.
- When you observe the dialogue box which says "**Cases to variables**: **Options"** dialogue box select group **by Index** and click next.
- Click finish.

Our data will be restructured as given below.

Id	Age.1	Age.2
1	23	60
2	25	36
3	26	29
4	35	56
5	42	32
6	22	54

Therefore, we have separate variables for ages of nurses and doctors.

Transpose all data. We choose this when we want to transpose our data. All rows will become columns and all columns will become rows in the new data. The procedure is as follows:

- From the **data** menu we select **restructure**
- From the dialogue box we select " **Transpose all data** " and click finish

- Transpose dialogue box will appear. We have to select all variables to transpose. (Note un-selected variables will be lost.) Click Ok.
- The transformed data that change rows to columns and columns to row will appear.

Example: Consider the following data set.

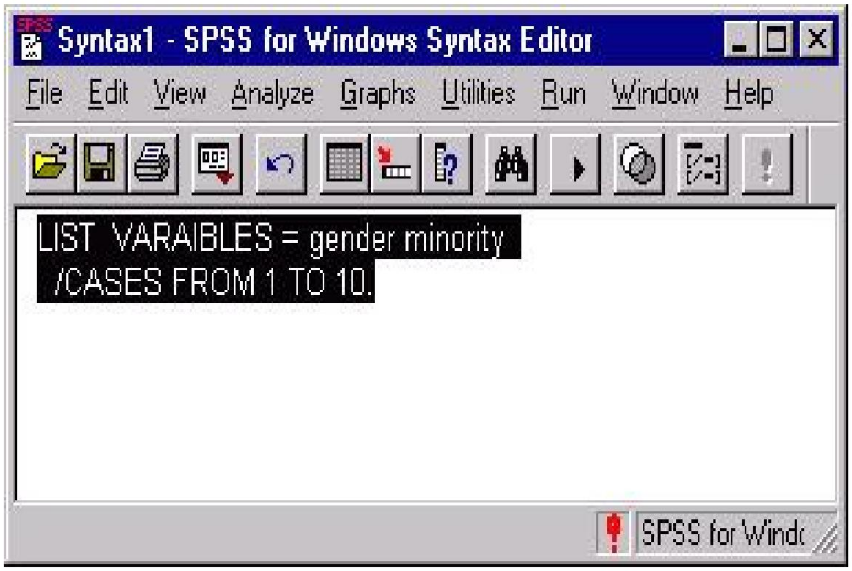
Id	Age	group
1	23	1
2	25	1
3	26	1
4	35	1
5	42	1
6	22	1
7	60	2
8	36	2
9	29	2
10	56	2
11	32	2
12	54	2

Applying the above procedure, the transposed form of this data is as given below.

Case_lbl	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
Id	1	2	3	4	5	6	7	8	9	10	11	12
age	23	25	26	35	42	22	60	36	29	56	32	54
group 1	1	1	1	1	1	1	2	2	2	2	2	2

2.8. Listing Cases

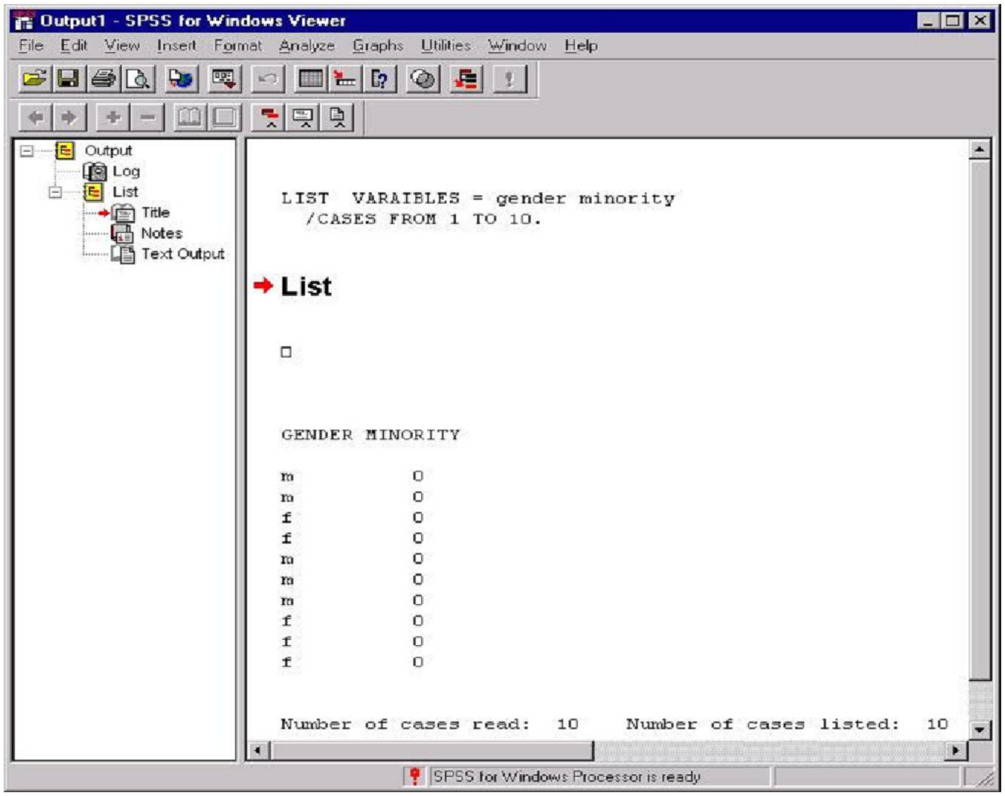
You may sometime want to print a list of your cases and the values of variables associated with each case, or perhaps a list of only some of the cases and variables. For example, if you have two variables that you want to examine visually, but this cannot be done because they are at very different places in your dataset, you could generate a list of only these variables in the Output Viewer. The procedure for doing this cannot be performed using dialog boxes and is available only through command syntax. The syntax for generating a list of cases is shown in the Syntax Editor window below. The variable names shown in lower case below instruct SPSS which variables to list in the output. Or, you can type in the command **ALL** in place of variables names, which will produce a listing of all of the variables in the file. The subcommand **/CASES FROM 1 TO 10**, is an instruction to SPSS to print only the first ten cases. If this instruction were omitted, all cases would be listed in the output.


Syntax1 - SPSS for Windows Syntax Editor
File Edit View Analyze Graphs Utilities Run Window Help
SPSS for Windc

```
LIST  VARAIBLES = gender minority
 /CASES FROM 1 TO 10.
```

To execute this command, first highlight the selection by pressing on your mouse button while dragging the arrow across the command or commands that you want to execute. Next, click on the icon with the black, right-facing arrow on it. Or, you can choose a selection from the ***Run* menu**.

Executing the command will print the list of variables, *gender* and *minority* in the above example, to the Output Viewer.The *Output Viewer* is the window in which all output will be printed. The Output Viewer is shown below, containing the text that would be generated from the above syntax.


Output1 - SPSS for Windows Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Output
Log
List
Title
Notes
Text Output
SPSS for Windows Processor is ready

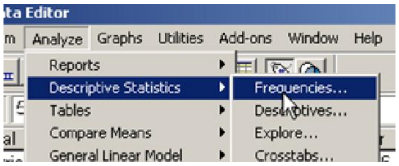
```
LIST  VARAIBLES = gender minority
  /CASES FROM 1 TO 10.
```

List

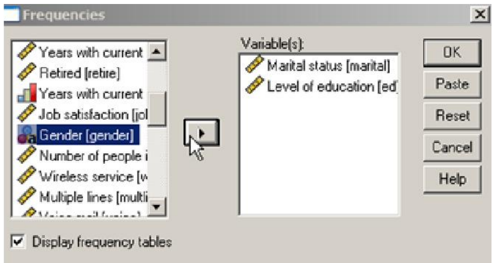
```
GENDER MINORITY

m          0
m          0
f          0
f          0
m          0
m          0
m          0
f          0
f          0
f          0


Number of cases read:  10    Number of cases listed:  10
```

Running Analyses (Frequency)

19. Select Analyze- Descriptive Stats- Frequencies



20. Select the desired variables and click the arrow to move them to the right side

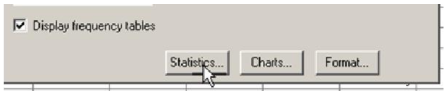


Variable(s):
Marital status [marital]
Level of education [ed

☑ Display frequency tables

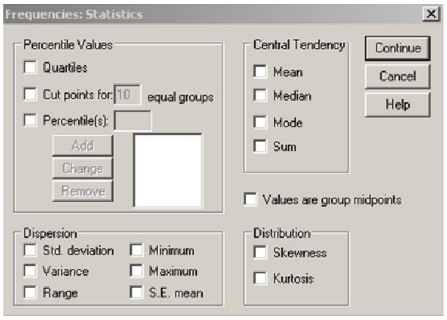
Running Analyses (Frequency)

21. Click Statistics



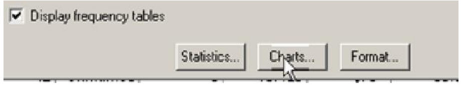
☑ Display frequency tables

22. Select any stats that you want to see, click Continue



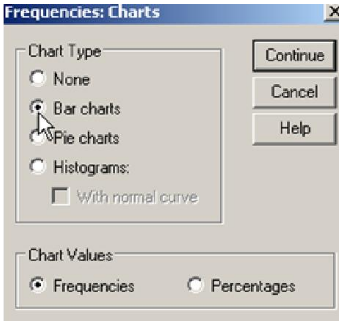
Running Analyses (Frequency)

23. Click Charts



☑ Display frequency tables

24. Select the type of chart you want, click Continue, then OK



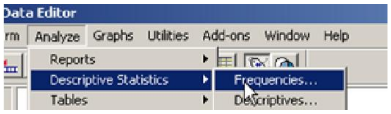
Frequencies: Charts

Chart Type
○ None
● Bar charts
○ Pie charts
○ Histograms:
□ With normal curve

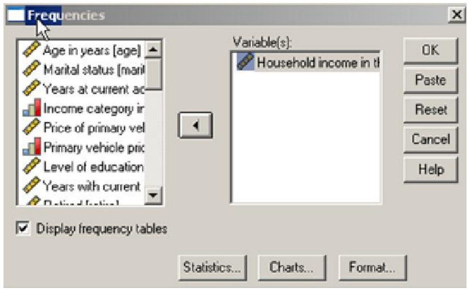
Chart Values
● Frequencies ○ Percentages

Running Analyses (Central Tendency)

25. Select Analyze- Descriptive Stats- Frequencies



26. Select the desired variables (household income) and click the arrow to move them to the right side



Frequencies

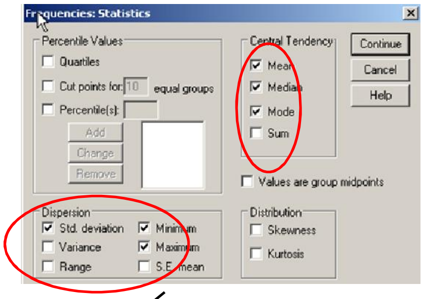
Age in years [age]
Marital status [marit
Years at current ac
Income category ir
Price of primary vel
Primary vehicle pric
Level of education
Years with current

Variable(s):
Household income in tl

☑ Display frequency tables

Running Analyses (Central Tendency)

27. Select some measures of central tendency and dispersion- click Continue then OK



Frequencies: Statistics

Percentile Values
□ Quartiles
□ Cut points for: 10 equal groups
□ Percentile(s):

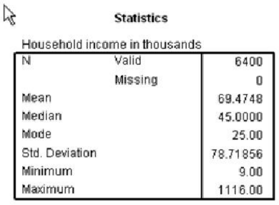
Central Tendency
☑ Mean
☑ Median
☑ Mode
□ Sum

□ Values are group midpoints

Dispersion
☑ Std. deviation
□ Variance
□ Range
☑ Minimum
☑ Maximum
□ S.E. mean

Distribution
□ Skewness
□ Kurtosis

Results will appear



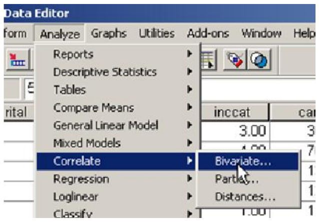
Statistics

Household income in thousands

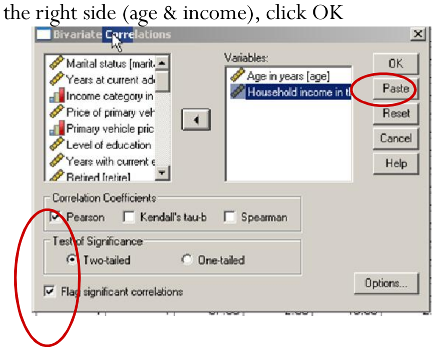
N	Valid	6400
	Missing	0
Mean		69.4748
Median		45.0000
Mode		25.00
Std. Deviation		78.71856
Minimum		9.00
Maximum		1116.00

Running Analyses (Correlation)

28. Click Analyze- Correlate- Bivariate



29. Move the two variables of interest to the right side (age & income), click OK



Bivariate Correlations

Marital status [marit...
Years at current ad...
Income category in...
Price of primary veh...
Primary vehicle pric...
Level of education
Years with current e...
Retired [retire]

Variables:
Age in years [age]
Household income in t...

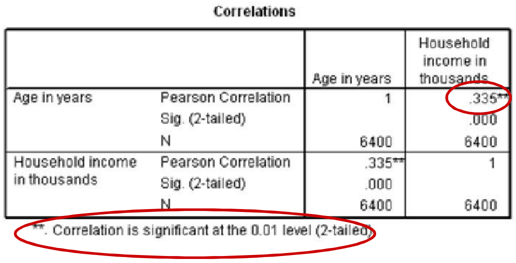
Correlation Coefficients
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Running Analyses (Correlation)

30. Results appear and tell us that the relationship is weak to moderate and results are not due to chance



Correlations

		Age in years	Household income in thousands
Age in years	Pearson Correlation	1	.335**
	Sig. (2-tailed)		.000
	N	6400	6400
Household income in thousands	Pearson Correlation	.335**	1
	Sig. (2-tailed)	.000	
	N	6400	6400

**. Correlation is significant at the 0.01 level (2-tailed).

ANSWERS TO SELF TEST ACTIVITY QUESTIONS

❖ SELF TEST ACTIVITY 1-4

1. The ABC Company drafted a set of wage and benefit demands to be presented to management. To get an idea of worker support for the package. He randomly selects the two largest groups of workers at his plant, the machinists (M) and the inspectors (I). He selects 30 of each group with the following results:

OPINION OF PACKAGE	M	I
Strongly support	8	12
Mildly support	13	5
Undecided	1	3
Mildly oppose	3	3
Strongly oppose	5	7

iv. What is the probability that a machinist randomly selected from the selected group mildly supports the package?

v. What is the probability that an inspector randomly selected from the selected group is undecided about the package?

vi. What is the probability that an employee randomly selected from the group strongly or mildly supports the package?

vii. What types of probability estimates are these?

Solution: Calculate the row and column totals

OPINION OF PACKAGE	M	I	Row Total
Strongly support	8	12	20
Mildly support	13	5	18
Undecided	1	3	4
Mildly oppose	3	3	6
Strongly oppose	5	7	12
Column Total	**30**	**30**	**60**

i. the probability that a machinist randomly selected from the selected group mildly supports the package is equal to $\frac{13}{30}$

ii. the probability that an inspector randomly selected from the selected group is undecided about the package is equal to $\frac{3}{30}$

iii. the probability that an employee randomly selected from the group strongly or mildly supports the package is calculated as follows:
let
S=A randomly selected employee from the group strongly supports the package
M= A randomly selected employee from the group mildly supports the package
Hence, the desired probability is
P (S or M) = P(S∪M) = P(S) + P(M) = $\frac{20}{60}+\frac{18}{60}=\frac{38}{60}$

iv. Relative frequency

❖ SELF TEST ACTIVITY 1-5

1. Find the errors in each of the following statements:

   a. The probabilities that an automobile salesperson will sell 0, 1, 2, or 3 cars on any given day in February are, respectively, 0.19, 0.38, 0.29, and 0.15.

      Solution: The total sum of the given probabilities is greater than one.

   b. The probability that it will rain tomorrow is 0.40 and the probability that it will not rain tomorrow is 0.52.

      Solution: The total probability of two complementary events should be one but here we have a sum of 0.40 + 0.52 = 0.92, which is less than one.

   c. The probabilities that a printer will make 0, 1, 2, 3, or 4 or more mistakes in printing a document are, respectively, 0.19, 0.34, - 0.25, 0.43, and 0.29.

      Solution: Even though the total probability is equal one, there is a negative probability. Every probability should be non-negative

d. On a single draw from a deck of playing cards the probability of selecting a heart is $\frac{1}{4}$, the probability of selecting a black card is $\frac{1}{2}$, and the probability of selecting both a heart and a black card is $\frac{1}{8}$.

Solution: Selecting both a black card and a heart at the same time is an impossible event since cards that are heart are also red as well. It means that black and red cards are mutually exclusive.

2. If *A* and *B* are mutually exclusive events and P(A) = 0.3 and P(B) = 0.5, find
   a. P(A U B);
   b. P(A');
   c. P(A'∩ B).

   Solution: A and B are mutually exclusive means P (A and B) = P(A∩B) = 0

   a. P(A U B) = P(A) + P(B) = 0.3 + 0.5 = 0.8
   b. P(A') = 1- P(A) = 1-0.3 = 0.7
   c. P(A'∩ B) = P(B) – P(A∩B) = 0.5 – 0 = 0.5

❖ Self Test Activity 1-6

1. A random sample of 200 employees is classified below according to sex and the level of education attained.

	Male	*Female*
Elementary	70	35
Secondary	40	25
College	20	10

If an employee is picked at random from this group, find the probability that

2. the person is a male, given that the person has a secondary education;
3. the person does not have a college degree, given that the person is a female.

Solution:

Let M = A randomly picked employee is a male
S = A randomly picked employee has secondary education
C = A randomly picked employee has a college degree

a. The probability that a randomly picked employee is male, given that the person has a secondary education is

$$P(M/S) = \frac{P(M\cap S)}{P(S)} = \frac{n(M\cap S)}{n(S)} = \frac{40}{200} = 0.2$$

b. The probability that a randomly picked employee does not have a college degree, given that the person is a female is

$$P(C'/S) = \frac{P(C'\cap M')}{P(M')} = \frac{n(C'\cap M')}{n(M')} = \frac{50}{60} = \frac{5}{6}$$

❖ SELF TEST ACTIVITY 1-7

1. By investing in a particular stock, a person can make a profit in 1 year of Birr5000 with probability 0.4 or take a loss of Birr1000 with probability 0.6. What is this person's expected gain?

Solution: Let X = Investing in a particular stock for one year
The values of X are x=Birr5,000, -Birr1,000
Thus the expected value of X is

$$E(X) = \sum x\,P(X = x) = (5000)(0.4) + (-1000)(0.6)$$

$$= 200 – 600 = -400$$

Therefore, the person's loses Birr 400 in one year.

❖ SELF TEST ACTIVITY 1-8

1. The Abay car account research team estimated that 20% of all car owners in Addis Ababa prefer Abay. Suppose a random sample of 5 car owners is chosen. Find the probability that;

   c. exactly 4 car owners prefer Abay car

   d. none of the car owners prefer Abay car

Solution

Let X: The number of people who prefer Abay car in a random sample of 5 car owners.

Probability that a randomly selected car owner prefer Abay car is p= 0.2, q= 1–p = 0.8, n= 5 people (Considering the random sample as 5 random trials)

Therefore, X ~ bin (n=5, p=0.2), which means

$$P(X=x)=\binom{5}{x}(0.2)^{x}(0.8)^{5-x}$$

a. $P(X=4)=\binom{5}{4}(0.2)^{4}(0.8)^{1}=5(0.0016)(0.8)=0.0064$

b. $P(X=0)=\binom{5}{0}(0.2)^{0}(0.8)^{5}=0.32768$

2. It is estimated that 4000 of the 10,000 voting residents of a town are against a new sales tax. If 15 eligible voters are selected at random and asked their opinion, what is the probability that at least 2 favor the new tax?

   Solution: P (success) $=p=\frac{4000}{10000}=\frac{2}{5}$, and $q=1-p=1-\frac{2}{5}=\frac{3}{5}$, n = 15

   Let X= the number of eligible voters that favor the new sales tax.

   Then X ~ bin (15, $\frac{2}{5}$), then the required answer for the question is,

$$
\begin{aligned}
P(X\geq 2) &= \sum_{x=2}^{15}\binom{15}{x}\left(\frac{2}{5}\right)^{x}\left(\frac{3^{15-x}}{5}\right)\\
&= 1-P(X\leq 1)\\
&= 1-P(X=1)-P(X=0)\\
&= 1-\binom{15}{1}\left(\frac{2}{5}\right)^{1}\left(\frac{3}{5}\right)^{14}-\binom{15}{0}\left(\frac{2}{5}\right)^{0}\left(\frac{3}{5}\right)^{15}\\
&= 1-0.004702-0.00047\\
&= 0.005172
\end{aligned}
$$

❖ SELF ACTIVITY TEST 1-9

1. If a receptionist's phone rings an average of 4 times an hour, find the probability of

d. no calls in a randomly selected hour,

e. exactly 2 calls per hour,

f. 3 or more calls in a duration of 2 hours.

Solution:

Let X = the number of times the receptionist's phone rings per hour

Then μ =4= average number of times the receptionist phone rings per hour

∴ X has the **Poisson** distribution with parameter μ =4.

P (X = x)= $\frac{e^{-4}4^{x}}{x!}=$ the probability that the phone rings exactly x times/hour

a. P (X=0)= $\frac{e^{-4}4^{0}}{0!}=e^{-4}=0.0183$

b. P (X=2)= $\frac{e^{-4}4^{2}}{2!}=8e^{-4}=0.1465$

c. To answer question c, you have to adjust the average number of rings per the time length given, that is, adjusted mean μ =4x2=8=average number of times the receptionist phone rings per 2 hours. This gives the distribution as follows

P (X = x)= $\frac{e^{-8}8^{x}}{x!}$, and the required you should give is

P(X≥3) =1 – P(X≤2)=1–P(X=0) –P(X=1) –P(X=2)

$$=1-\frac{e^{-8}8^{0}}{0!}-\frac{e^{-8}8^{1}}{1!}-\frac{e^{-8}8^{2}}{2!}$$

$$=1-e^{-8}-8e^{-8}-\frac{64}{2}e^{-8}$$

$$=1-41e^{-8}=0.986$$

2. The number of students entering a library of a certain university college is on the average 90 per hour. Find the probability of 1 to 3 students entering the library in a minute.

Solution:

Let X = the number of students entering the library per minute.

Given μ = $\frac{90}{60}=1.5=$ Average number of students entering the library per minute (note that 1 hour is 60 minutes.)

$$P(X=x)=\frac{e^{-1.5}(1.5)^x}{x!}, \quad \text{for x = 0, 1, 2, } \cdots.$$

Therefore, $\sum_{x=1}^{3} P(X=x) \quad = \sum_{x=1}^{3} \frac{e^{-1.5}(1.5)^x}{x!}$ =

$$\frac{e^{-1.5}(1.5)}{1!}+\frac{e^{-1.5}(1.5)^2}{2!}+\frac{e^{-1.5}(1.5)^3}{3!}$$

$$=e^{-1.5}\left(\frac{1.5}{1!}+\frac{(1.5)^2}{2!}+\frac{(1.5)^3}{3!}\right)$$

= 0.711

❖ SELF TEST ACTIVITY 1.10

2. Training was designed to upgrade the technical skills of factory workers. A study of past participants indicates that the length of time spent on the training is normally distributed with mean 500 hours and standard deviation of 100 hours.

   What is the probability that a randomly selected worker will need

   e) more than 500 hours?

   f) between 600 and 750 hours?

   g) between 450 and 550 hours?

   h) between 350 and 450 hours?

   Solution:

   Let X = the time spent on the training by a randomly selected worker

   Given: X~ N (500,100^2)

   a. P (X>500)=P (Z> $\frac{500-500}{100}$) =P (Z>0)=0.5

   b. P (600<X<750)=P ($\frac{600-500}{100} < Z < \frac{750-500}{100}$) = P (1<Z<2.5)

   $$= P(0<Z<2.5) - P(0<Z<1)$$
   $$= 0.4938 - 0.3413$$
   $$= 0.1525$$

   c. $\text{P}(450 < \text{X} < 550) = \text{P}\left(\frac{450-550}{100} < Z < \frac{550-500}{100}\right)$

$$= P(-0.5 < Z < 0.5)$$
$$= (0 < Z < |-0.5|) + P(0 < Z < 0.5)$$
$$= 2P(0 < Z < 0.5)$$
$$= 2 \times 0.1915$$
$$= 0.3830$$

d. P (350<X<450)=P ($\frac{350-500}{100} < Z < \frac{450-500}{100}$)

$$= (-1.5 < Z < -0.5)$$
$$= P(0.5 < Z < 1.5)$$
$$= P(0 < Z < 1.5) - P(0 < Z < 0.5)$$
$$= 0.4332 - 0.1915$$
$$= 0.2417$$

❖ SELF ACTIVITY TEST 6-1

The ranks of 12 students according to their marks in Mathematics and statistics were as follows:

Student No.:	1	2	3	4	5	6	7	8	9	10	11	12
Mathematics:	5	2	1	6	8	11	12	4	3	9	7	10
Statistics:	4	3	2	7	6	9	10	5	1	11	8	12

Do students who are good in Mathematics also excel in statistics and vice versa?

Solution

In order to calculate r_s we need to get $\sum d_i^2$ as in the following table:

Student No.:	1	2	3	4	5	6	7	8	9	10	11	12
Mathematics:	5	2	1	6	8	11	12	4	3	9	7	10
Statistics:	4	3	2	7	6	9	10	5	1	11	8	12
di	1	-1	-1	-1	2	2	2	-1	2	-2	-1	-2

di2	1	1	1	1	4	4	4	1	4	4	1	4

From the table, we get $\sum d_i^2 = 30$

$$\therefore r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 30}{12 \times 143}$$

$$= 1 - 0.105$$

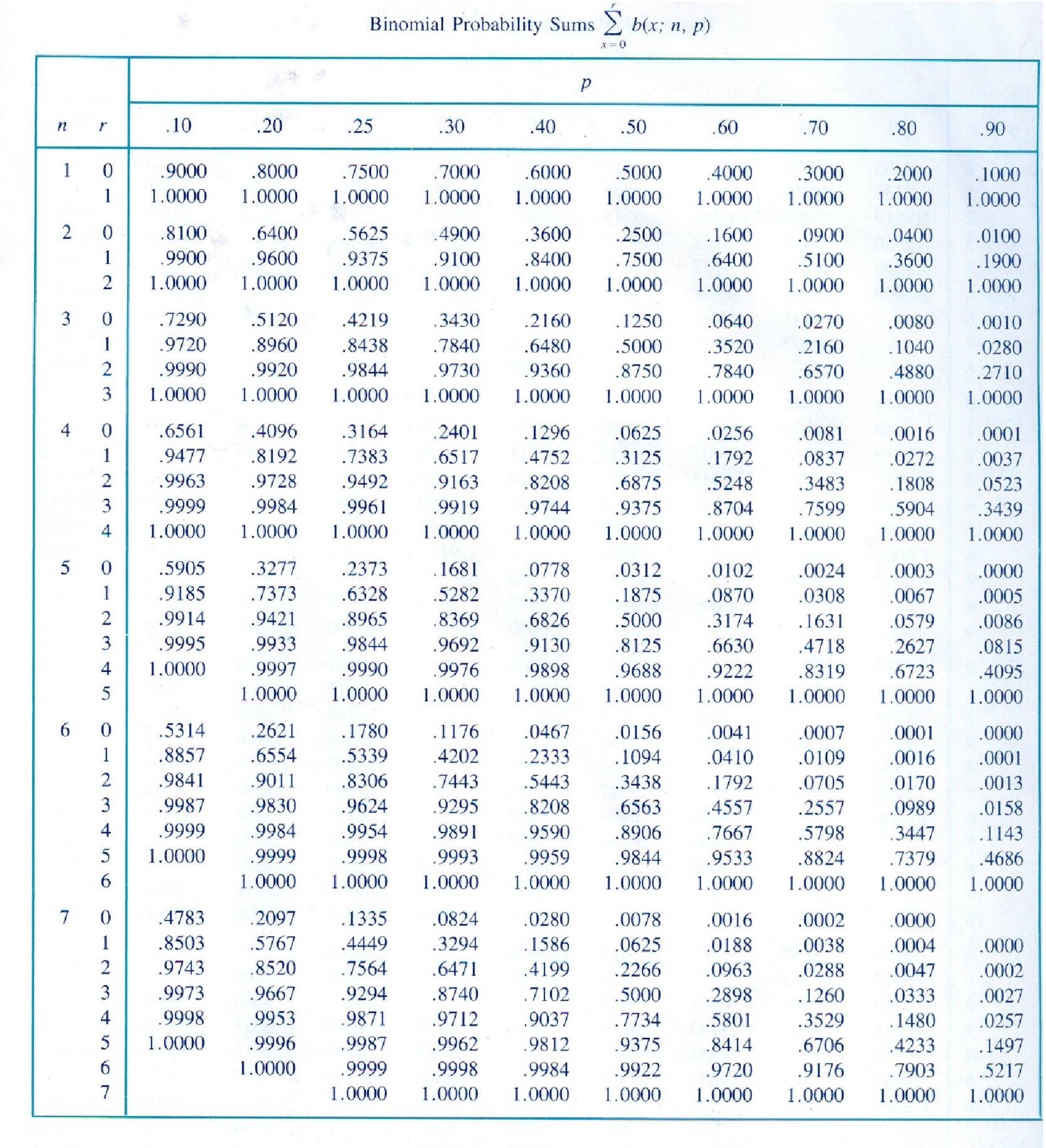
$$= 0.895$$

The correlation between the ranks of marks in the two subjects is very high. From this, it can be concluded that the students who are good in Mathematics are also good in Statistics.

Statistical Tables

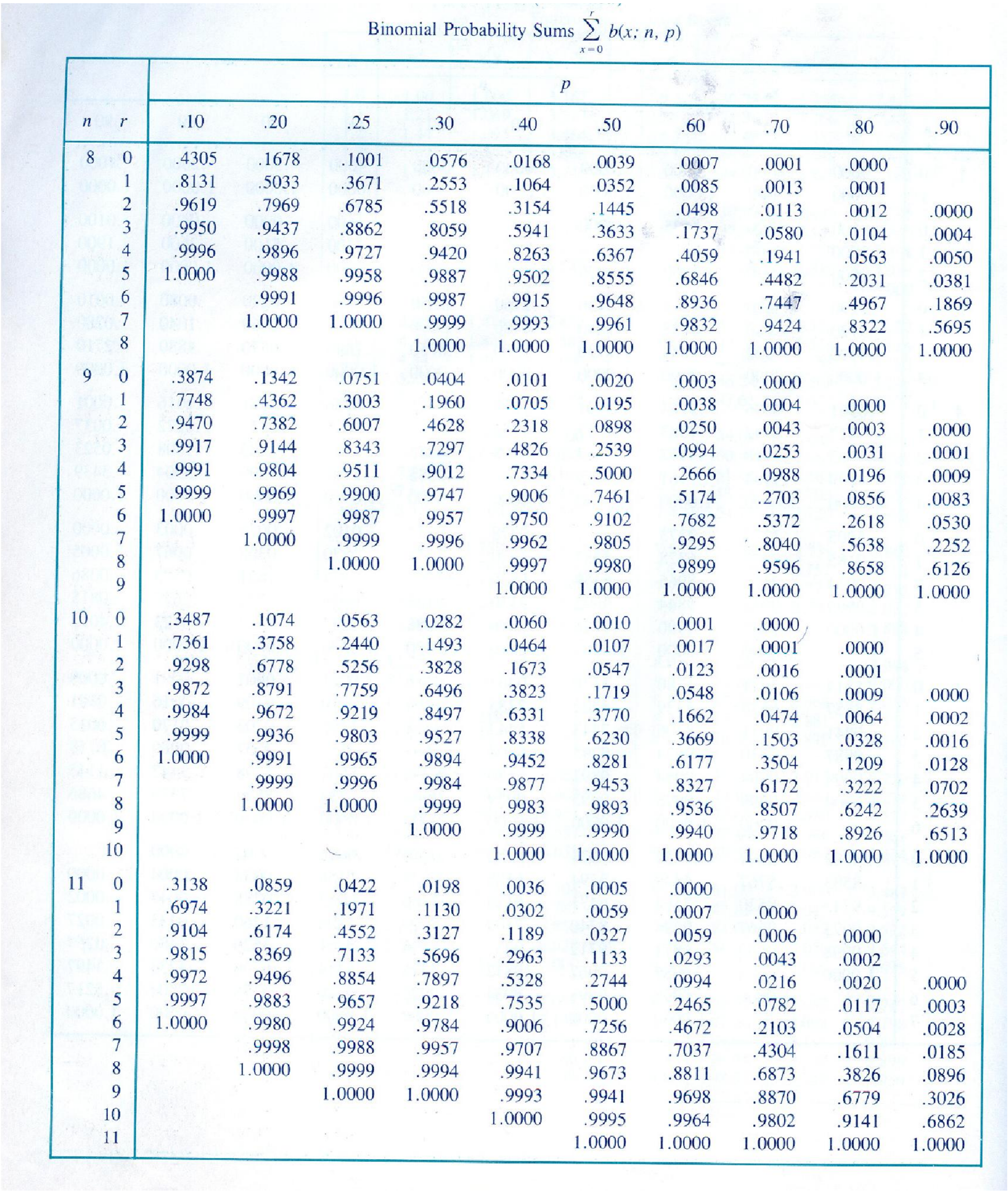
VALUE OF $e^{-\lambda}$

λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$
0.1	0.9048	2.6	0.0743	5.1	0.0061	7.6	0.0005	10.1	0.0000
0.2	0.8187	2.7	0.0672	5.2	0.0055	7.7	0.0005	10.2	0.0000
0.3	0.7408	2.8	0.0608	5.3	0.0050	7.8	0.0004	10.3	0.0000
0.4	0.6703	2.9	0.0550	5.4	0.0045	7.9	0.0004	10.4	0.0000
0.5	0.6065	3.0	0.0498	5.5	0.0041	8.0	0.0003	10.5	0.0000
0.6	0.5488	3.1	0.0450	5.6	0.0037	8.1	0.0003	10.6	0.0000
0.7	0.4966	3.2	0.0408	5.7	0.0033	8.2	0.0003	10.7	0.0000
0.8	0.4493	3.3	0.0369	5.8	0.0030	8.3	0.0002	10.8	0.0000
0.9	0.4066	3.4	0.0334	5.9	0.0027	8.4	0.0002	10.9	0.0000
1.0	0.3679	3.5	0.0302	6.0	0.0025	8.5	0.0002	11.0	0.0000
1.1	0.3329	3.6	0.0273	6.1	0.0022	8.6	0.0002	11.1	0.0000
1.2	0.3012	3.7	0.0247	6.2	0.0020	8.7	0.0002	11.2	0.0000
1.3	0.2725	3.8	0.0224	6.3	0.0018	8.8	0.0002	11.3	0.0000
1.4	0.2466	3.9	0.0202	6.4	0.0017	8.9	0.0001	11.4	0.0000
1.5	0.2231	4.0	0.0183	6.5	0.0015	9.0	0.0001	11.5	0.0000
1.6	0.2019	4.1	0.0166	6.6	0.0014	9.1	0.0001	11.6	0.0000
1.7	0.1827	4.2	0.0150	6.7	0.0012	9.2	0.0001	11.7	0.0000
1.8	0.1653	4.3	0.0136	6.8	0.0011	9.3	0.0001	11.8	0.0000
1.9	0.1496	4.4	0.0123	6.9	0.0010	9.4	0.0001	11.9	0.0000
2.0	0.1353	4.5	0.0111	7.0	0.0009	9.5	0.0001	12.0	0.0000
2.1	0.1225	4.6	0.0101	7.1	0.0008	9.6	0.0001	12.1	0.0000
2.2	0.1108	4.7	0.0091	7.2	0.0007	9.7	0.0001	12.2	0.0000
2.3	0.1003	4.8	0.0082	7.3	0.0007	9.8	0.0001	12.3	0.0000
2.4	0.0907	4.9	0.0074	7.4	0.0006	9.9	0.0001	12.4	0.0000
2.5	0.0821	5.0	0.0067	7.5	0.0006	10.0	0.0000	12.5	0.0000



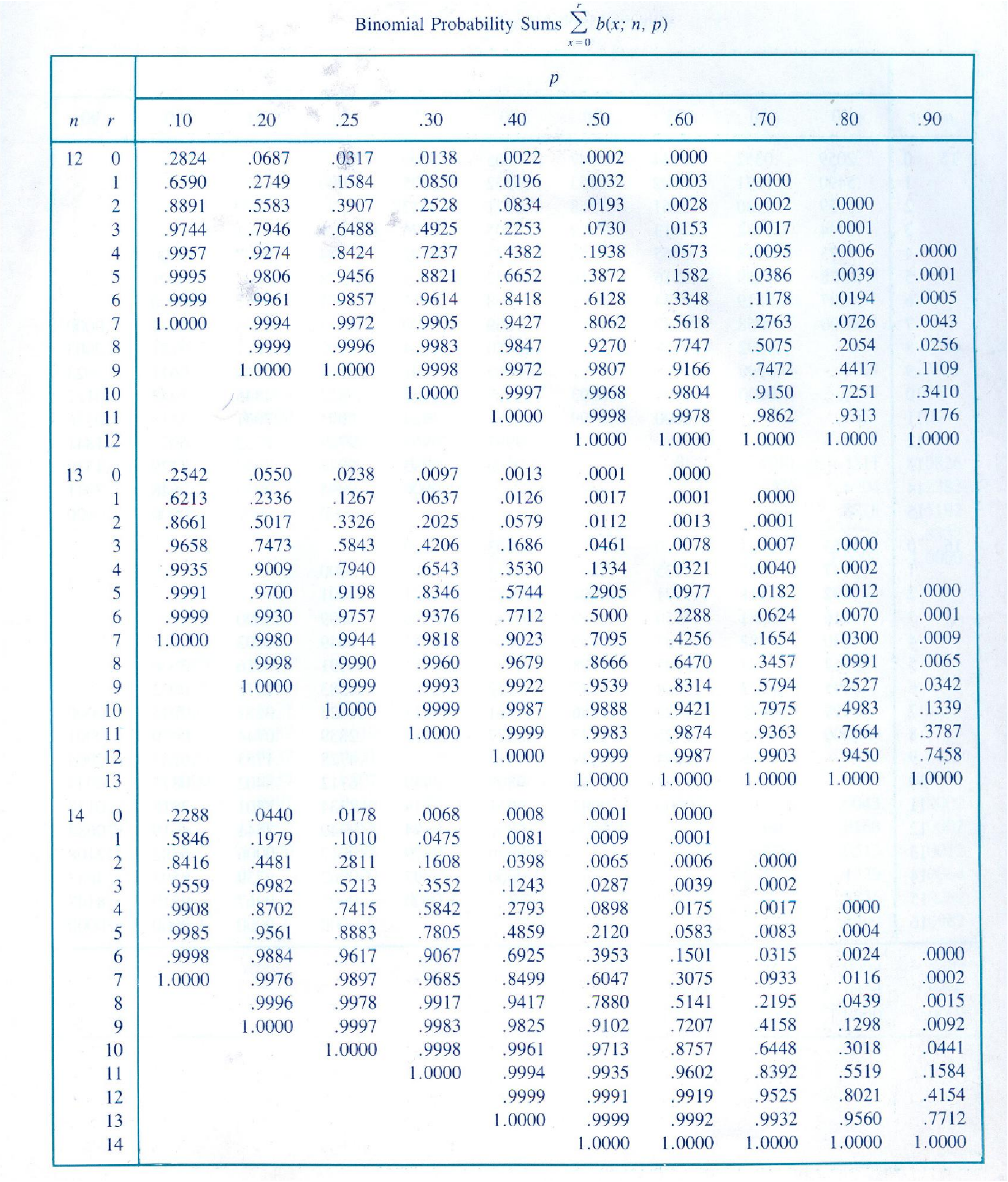
Binomial Probability Sums $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	*r*	.10	.20	.25	.30	.40	.50	.60	.70	.80	.90
1	0	.9000	.8000	.7500	.7000	.6000	.5000	.4000	.3000	.2000	.1000
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0	.8100	.6400	.5625	.4900	.3600	.2500	.1600	.0900	.0400	.0100
	1	.9900	.9600	.9375	.9100	.8400	.7500	.6400	.5100	.3600	.1900
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	.7290	.5120	.4219	.3430	.2160	.1250	.0640	.0270	.0080	.0010
	1	.9720	.8960	.8438	.7840	.6480	.5000	.3520	.2160	.1040	.0280
	2	.9990	.9920	.9844	.9730	.9360	.8750	.7840	.6570	.4880	.2710
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	.6561	.4096	.3164	.2401	.1296	.0625	.0256	.0081	.0016	.0001
	1	.9477	.8192	.7383	.6517	.4752	.3125	.1792	.0837	.0272	.0037
	2	.9963	.9728	.9492	.9163	.8208	.6875	.5248	.3483	.1808	.0523
	3	.9999	.9984	.9961	.9919	.9744	.9375	.8704	.7599	.5904	.3439
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	.5905	.3277	.2373	.1681	.0778	.0312	.0102	.0024	.0003	.0000
	1	.9185	.7373	.6328	.5282	.3370	.1875	.0870	.0308	.0067	.0005
	2	.9914	.9421	.8965	.8369	.6826	.5000	.3174	.1631	.0579	.0086
	3	.9995	.9933	.9844	.9692	.9130	.8125	.6630	.4718	.2627	.0815
	4	1.0000	.9997	.9990	.9976	.9898	.9688	.9222	.8319	.6723	.4095
	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	.5314	.2621	.1780	.1176	.0467	.0156	.0041	.0007	.0001	.0000
	1	.8857	.6554	.5339	.4202	.2333	.1094	.0410	.0109	.0016	.0001
	2	.9841	.9011	.8306	.7443	.5443	.3438	.1792	.0705	.0170	.0013
	3	.9987	.9830	.9624	.9295	.8208	.6563	.4557	.2557	.0989	.0158
	4	.9999	.9984	.9954	.9891	.9590	.8906	.7667	.5798	.3447	.1143
	5	1.0000	.9999	.9998	.9993	.9959	.9844	.9533	.8824	.7379	.4686
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	.4783	.2097	.1335	.0824	.0280	.0078	.0016	.0002	.0000	
	1	.8503	.5767	.4449	.3294	.1586	.0625	.0188	.0038	.0004	.0000
	2	.9743	.8520	.7564	.6471	.4199	.2266	.0963	.0288	.0047	.0002
	3	.9973	.9667	.9294	.8740	.7102	.5000	.2898	.1260	.0333	.0027
	4	.9998	.9953	.9871	.9712	.9037	.7734	.5801	.3529	.1480	.0257
	5	1.0000	.9996	.9987	.9962	.9812	.9375	.8414	.6706	.4233	.1497
	6		1.0000	.9999	.9998	.9984	.9922	.9720	.9176	.7903	.5217
	7			1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000



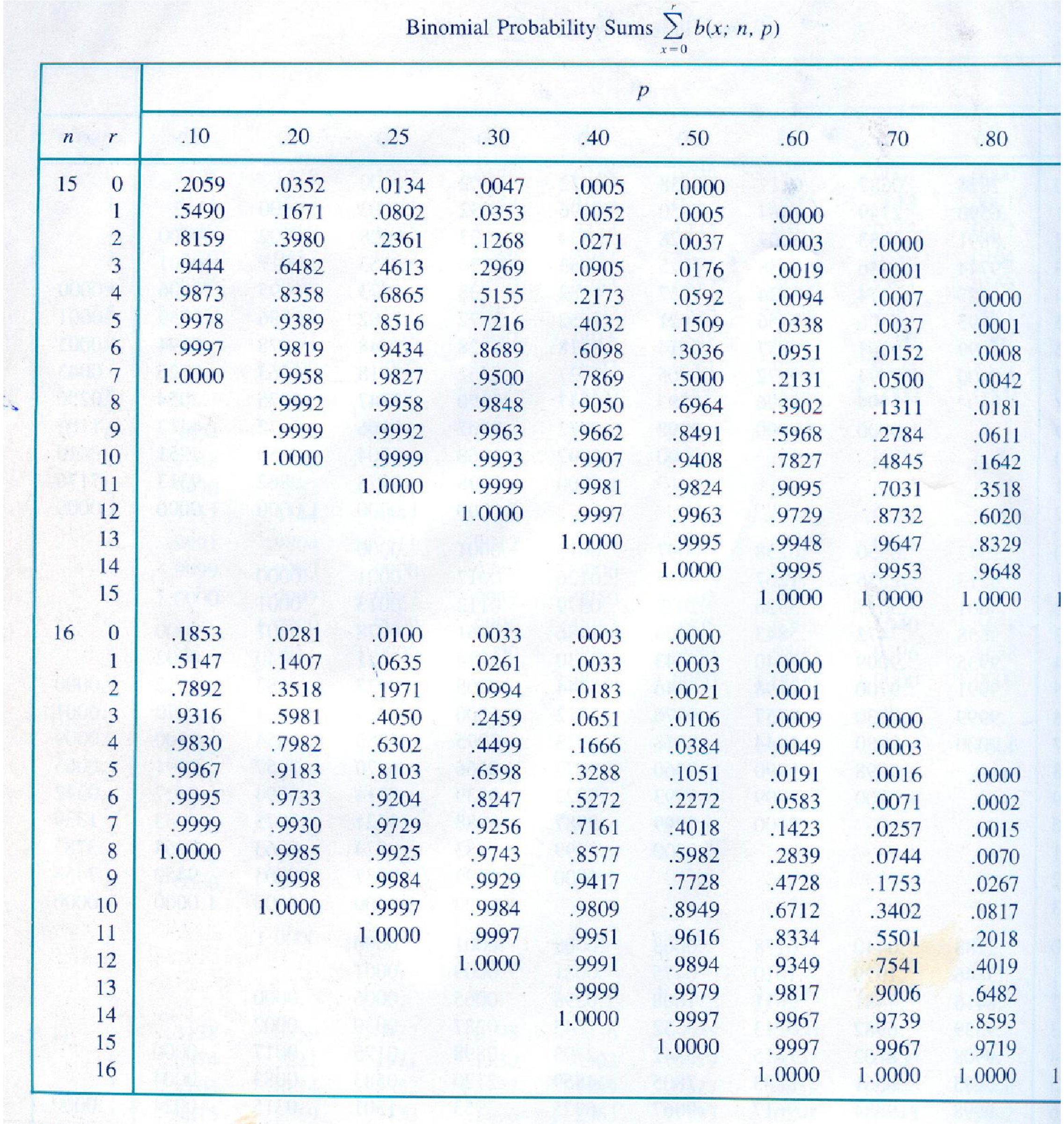
Binomial Probability Sums $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	*r*	.10	.20	.25	.30	.40	.50	.60	.70	.80	.90
8	0	.4305	.1678	.1001	.0576	.0168	.0039	.0007	.0001	.0000	
	1	.8131	.5033	.3671	.2553	.1064	.0352	.0085	.0013	.0001	
	2	.9619	.7969	.6785	.5518	.3154	.1445	.0498	.0113	.0012	.0000
	3	.9950	.9437	.8862	.8059	.5941	.3633	.1737	.0580	.0104	.0004
	4	.9996	.9896	.9727	.9420	.8263	.6367	.4059	.1941	.0563	.0050
	5	1.0000	.9988	.9958	.9887	.9502	.8555	.6846	.4482	.2031	.0381
	6		.9991	.9996	.9987	.9915	.9648	.8936	.7447	.4967	.1869
	7		1.0000	1.0000	.9999	.9993	.9961	.9832	.9424	.8322	.5695
	8				1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	.3874	.1342	.0751	.0404	.0101	.0020	.0003	.0000		
	1	.7748	.4362	.3003	.1960	.0705	.0195	.0038	.0004	.0000	
	2	.9470	.7382	.6007	.4628	.2318	.0898	.0250	.0043	.0003	.0000
	3	.9917	.9144	.8343	.7297	.4826	.2539	.0994	.0253	.0031	.0001
	4	.9991	.9804	.9511	.9012	.7334	.5000	.2666	.0988	.0196	.0009
	5	.9999	.9969	.9900	.9747	.9006	.7461	.5174	.2703	.0856	.0083
	6	1.0000	.9997	.9987	.9957	.9750	.9102	.7682	.5372	.2618	.0530
	7		1.0000	.9999	.9996	.9962	.9805	.9295	.8040	.5638	.2252
	8			1.0000	1.0000	.9997	.9980	.9899	.9596	.8658	.6126
	9					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	.3487	.1074	.0563	.0282	.0060	.0010	.0001	.0000		
	1	.7361	.3758	.2440	.1493	.0464	.0107	.0017	.0001	.0000	
	2	.9298	.6778	.5256	.3828	.1673	.0547	.0123	.0016	.0001	
	3	.9872	.8791	.7759	.6496	.3823	.1719	.0548	.0106	.0009	.0000
	4	.9984	.9672	.9219	.8497	.6331	.3770	.1662	.0474	.0064	.0002
	5	.9999	.9936	.9803	.9527	.8338	.6230	.3669	.1503	.0328	.0016
	6	1.0000	.9991	.9965	.9894	.9452	.8281	.6177	.3504	.1209	.0128
	7		.9999	.9996	.9984	.9877	.9453	.8327	.6172	.3222	.0702
	8		1.0000	1.0000	.9999	.9983	.9893	.9536	.8507	.6242	.2639
	9				1.0000	.9999	.9990	.9940	.9718	.8926	.6513
	10					1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	.3138	.0859	.0422	.0198	.0036	.0005	.0000			
	1	.6974	.3221	.1971	.1130	.0302	.0059	.0007	.0000		
	2	.9104	.6174	.4552	.3127	.1189	.0327	.0059	.0006	.0000	
	3	.9815	.8369	.7133	.5696	.2963	.1133	.0293	.0043	.0002	
	4	.9972	.9496	.8854	.7897	.5328	.2744	.0994	.0216	.0020	.0000
	5	.9997	.9883	.9657	.9218	.7535	.5000	.2465	.0782	.0117	.0003
	6	1.0000	.9980	.9924	.9784	.9006	.7256	.4672	.2103	.0504	.0028
	7		.9998	.9988	.9957	.9707	.8867	.7037	.4304	.1611	.0185
	8		1.0000	.9999	.9994	.9941	.9673	.8811	.6873	.3826	.0896
	9			1.0000	1.0000	.9993	.9941	.9698	.8870	.6779	.3026
	10					1.0000	.9995	.9964	.9802	.9141	.6862
	11						1.0000	1.0000	1.0000	1.0000	1.0000



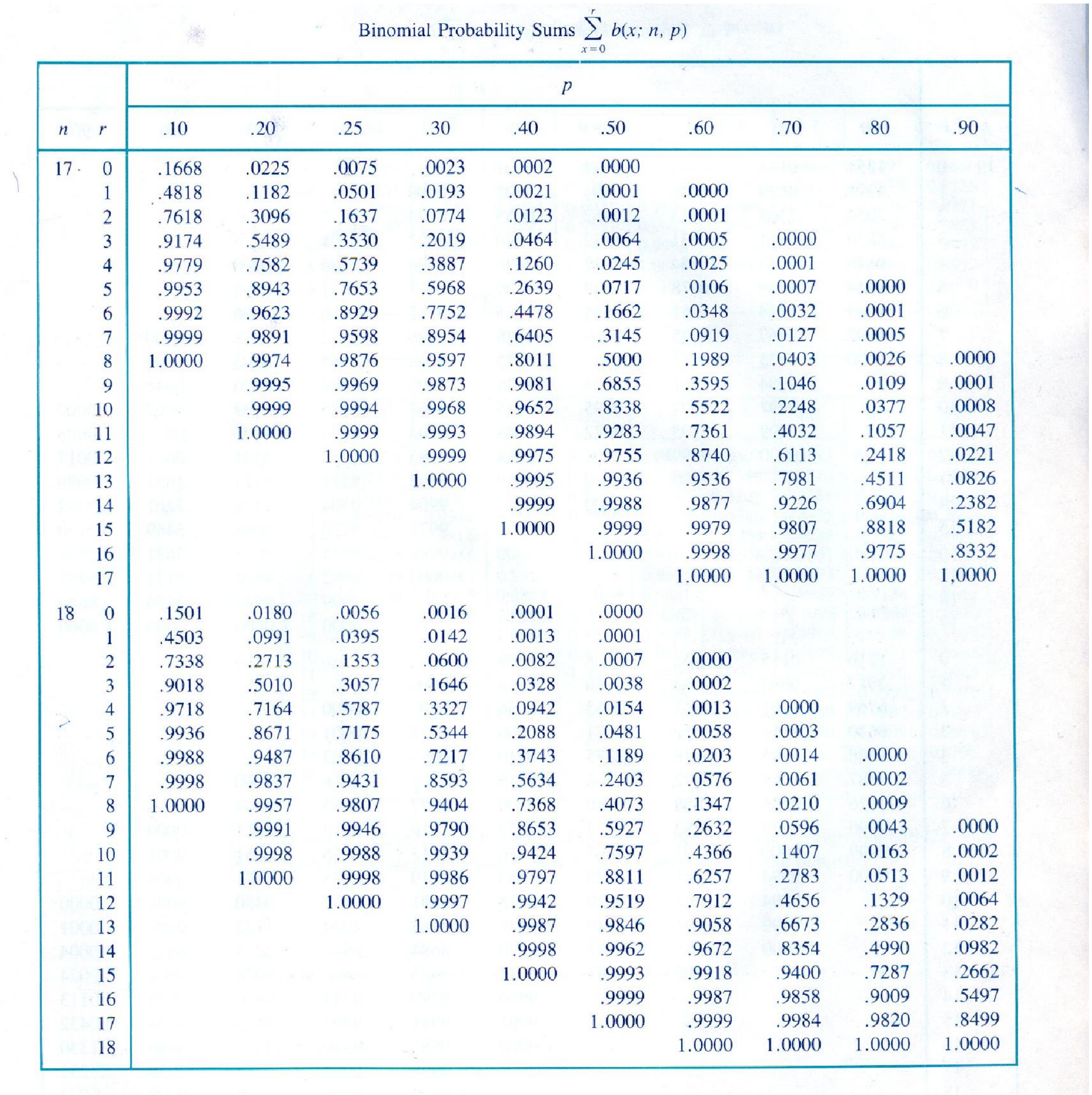
Binomial Probability Sums $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	r	.10	.20	.25	.30	.40	.50	.60	.70	.80	.90
12	0	.2824	.0687	.0317	.0138	.0022	.0002	.0000			
	1	.6590	.2749	.1584	.0850	.0196	.0032	.0003	.0000		
	2	.8891	.5583	.3907	.2528	.0834	.0193	.0028	.0002	.0000	
	3	.9744	.7946	.6488	.4925	.2253	.0730	.0153	.0017	.0001	
	4	.9957	.9274	.8424	.7237	.4382	.1938	.0573	.0095	.0006	.0000
	5	.9995	.9806	.9456	.8821	.6652	.3872	.1582	.0386	.0039	.0001
	6	.9999	.9961	.9857	.9614	.8418	.6128	.3348	.1178	.0194	.0005
	7	1.0000	.9994	.9972	.9905	.9427	.8062	.5618	.2763	.0726	.0043
	8		.9999	.9996	.9983	.9847	.9270	.7747	.5075	.2054	.0256
	9		1.0000	1.0000	.9998	.9972	.9807	.9166	.7472	.4417	.1109
	10				1.0000	.9997	.9968	.9804	.9150	.7251	.3410
	11					1.0000	.9998	.9978	.9862	.9313	.7176
	12						1.0000	1.0000	1.0000	1.0000	1.0000
13	0	.2542	.0550	.0238	.0097	.0013	.0001	.0000			
	1	.6213	.2336	.1267	.0637	.0126	.0017	.0001	.0000		
	2	.8661	.5017	.3326	.2025	.0579	.0112	.0013	.0001		
	3	.9658	.7473	.5843	.4206	.1686	.0461	.0078	.0007	.0000	
	4	.9935	.9009	.7940	.6543	.3530	.1334	.0321	.0040	.0002	
	5	.9991	.9700	.9198	.8346	.5744	.2905	.0977	.0182	.0012	.0000
	6	.9999	.9930	.9757	.9376	.7712	.5000	.2288	.0624	.0070	.0001
	7	1.0000	.9980	.9944	.9818	.9023	.7095	.4256	.1654	.0300	.0009
	8		.9998	.9990	.9960	.9679	.8666	.6470	.3457	.0991	.0065
	9		1.0000	.9999	.9993	.9922	.9539	.8314	.5794	.2527	.0342
	10			1.0000	.9999	.9987	.9888	.9421	.7975	.4983	.1339
	11				1.0000	.9999	.9983	.9874	.9363	.7664	.3787
	12					1.0000	.9999	.9987	.9903	.9450	.7458
	13						1.0000	1.0000	1.0000	1.0000	1.0000
14	0	.2288	.0440	.0178	.0068	.0008	.0001	.0000			
	1	.5846	.1979	.1010	.0475	.0081	.0009	.0001			
	2	.8416	.4481	.2811	.1608	.0398	.0065	.0006	.0000		
	3	.9559	.6982	.5213	.3552	.1243	.0287	.0039	.0002		
	4	.9908	.8702	.7415	.5842	.2793	.0898	.0175	.0017	.0000	
	5	.9985	.9561	.8883	.7805	.4859	.2120	.0583	.0083	.0004	
	6	.9998	.9884	.9617	.9067	.6925	.3953	.1501	.0315	.0024	.0000
	7	1.0000	.9976	.9897	.9685	.8499	.6047	.3075	.0933	.0116	.0002
	8		.9996	.9978	.9917	.9417	.7880	.5141	.2195	.0439	.0015
	9		1.0000	.9997	.9983	.9825	.9102	.7207	.4158	.1298	.0092
	10			1.0000	.9998	.9961	.9713	.8757	.6448	.3018	.0441
	11				1.0000	.9994	.9935	.9602	.8392	.5519	.1584
	12					.9999	.9991	.9919	.9525	.8021	.4154
	13					1.0000	.9999	.9992	.9932	.9560	.7712
	14						1.0000	1.0000	1.0000	1.0000	1.0000



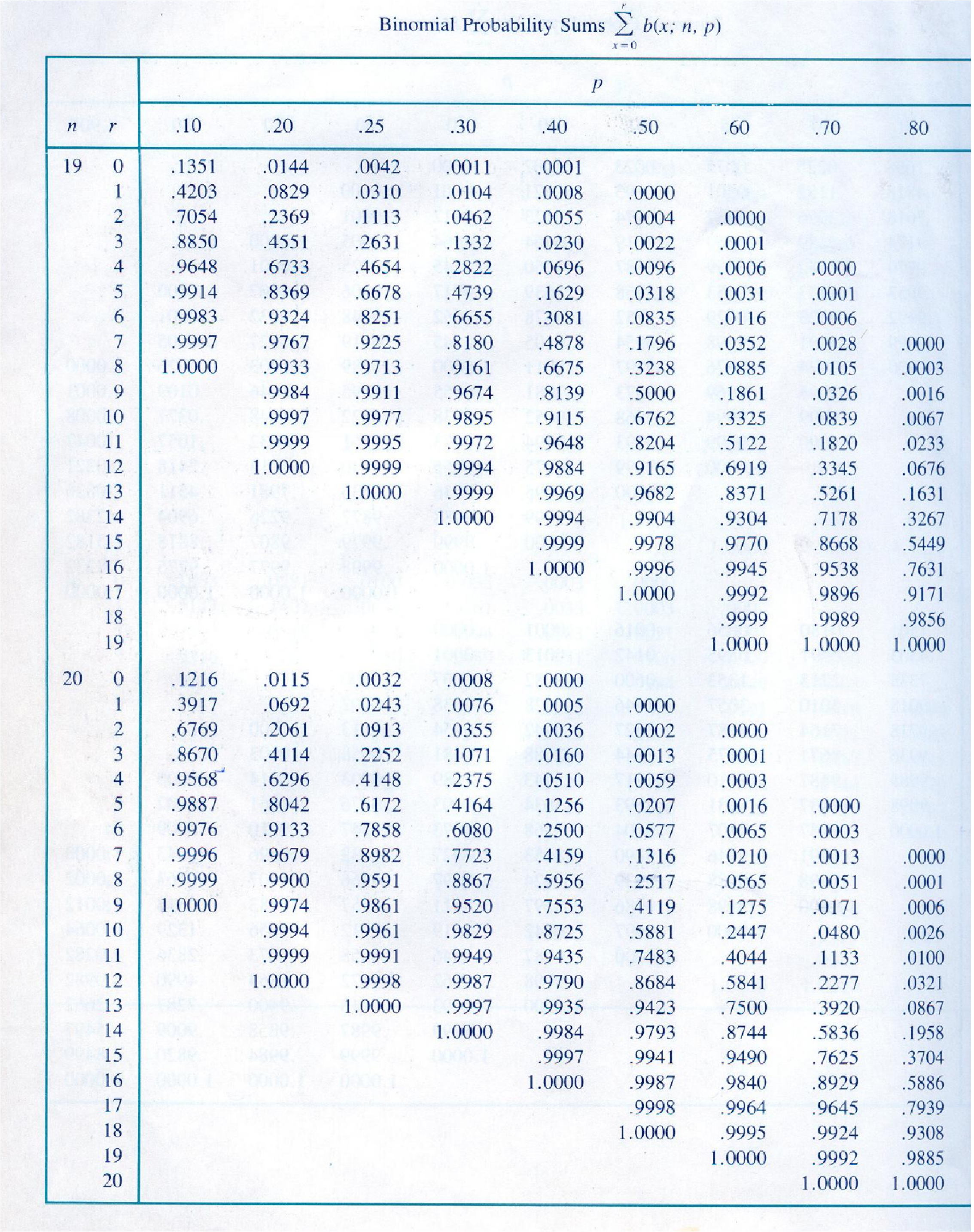
Binomial Probability Sums $\sum_{x=0}^{r} b(x; n, p)$

		p								
n	*r*	.10	.20	.25	.30	.40	.50	.60	.70	.80
15	0	.2059	.0352	.0134	.0047	.0005	.0000			
	1	.5490	.1671	.0802	.0353	.0052	.0005	.0000		
	2	.8159	.3980	.2361	.1268	.0271	.0037	.0003	.0000	
	3	.9444	.6482	.4613	.2969	.0905	.0176	.0019	.0001	
	4	.9873	.8358	.6865	.5155	.2173	.0592	.0094	.0007	.0000
	5	.9978	.9389	.8516	.7216	.4032	.1509	.0338	.0037	.0001
	6	.9997	.9819	.9434	.8689	.6098	.3036	.0951	.0152	.0008
	7	1.0000	.9958	.9827	.9500	.7869	.5000	.2131	.0500	.0042
	8		.9992	.9958	.9848	.9050	.6964	.3902	.1311	.0181
	9		.9999	.9992	.9963	.9662	.8491	.5968	.2784	.0611
	10		1.0000	.9999	.9993	.9907	.9408	.7827	.4845	.1642
	11			1.0000	.9999	.9981	.9824	.9095	.7031	.3518
	12				1.0000	.9997	.9963	.9729	.8732	.6020
	13					1.0000	.9995	.9948	.9647	.8329
	14						1.0000	.9995	.9953	.9648
	15							1.0000	1.0000	1.0000
16	0	.1853	.0281	.0100	.0033	.0003	.0000			
	1	.5147	.1407	.0635	.0261	.0033	.0003	.0000		
	2	.7892	.3518	.1971	.0994	.0183	.0021	.0001		
	3	.9316	.5981	.4050	.2459	.0651	.0106	.0009	.0000	
	4	.9830	.7982	.6302	.4499	.1666	.0384	.0049	.0003	
	5	.9967	.9183	.8103	.6598	.3288	.1051	.0191	.0016	.0000
	6	.9995	.9733	.9204	.8247	.5272	.2272	.0583	.0071	.0002
	7	.9999	.9930	.9729	.9256	.7161	.4018	.1423	.0257	.0015
	8	1.0000	.9985	.9925	.9743	.8577	.5982	.2839	.0744	.0070
	9		.9998	.9984	.9929	.9417	.7728	.4728	.1753	.0267
	10		1.0000	.9997	.9984	.9809	.8949	.6712	.3402	.0817
	11			1.0000	.9997	.9951	.9616	.8334	.5501	.2018
	12				1.0000	.9991	.9894	.9349	.7541	.4019
	13					.9999	.9979	.9817	.9006	.6482
	14					1.0000	.9997	.9967	.9739	.8593
	15						1.0000	.9997	.9967	.9719
	16							1.0000	1.0000	1.0000



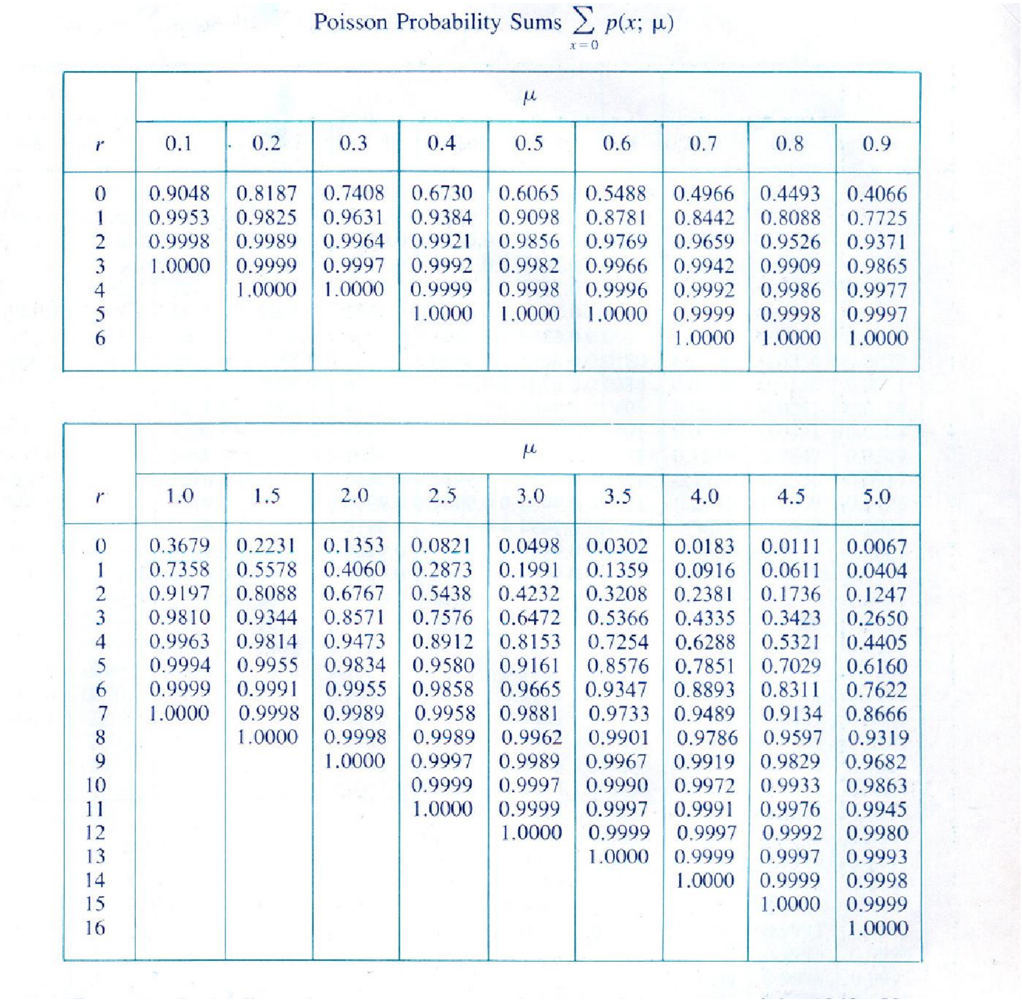
Binomial Probability Sums $\sum_{x=0}^{r} b(x; n, p)$

		p									
n	*r*	.10	.20	.25	.30	.40	.50	.60	.70	.80	.90
17	0	.1668	.0225	.0075	.0023	.0002	.0000				
	1	.4818	.1182	.0501	.0193	.0021	.0001	.0000			
	2	.7618	.3096	.1637	.0774	.0123	.0012	.0001			
	3	.9174	.5489	.3530	.2019	.0464	.0064	.0005	.0000		
	4	.9779	.7582	.5739	.3887	.1260	.0245	.0025	.0001		
	5	.9953	.8943	.7653	.5968	.2639	.0717	.0106	.0007	.0000	
	6	.9992	.9623	.8929	.7752	.4478	.1662	.0348	.0032	.0001	
	7	.9999	.9891	.9598	.8954	.6405	.3145	.0919	.0127	.0005	
	8	1.0000	.9974	.9876	.9597	.8011	.5000	.1989	.0403	.0026	.0000
	9		.9995	.9969	.9873	.9081	.6855	.3595	.1046	.0109	.0001
	10		.9999	.9994	.9968	.9652	.8338	.5522	.2248	.0377	.0008
	11		1.0000	.9999	.9993	.9894	.9283	.7361	.4032	.1057	.0047
	12			1.0000	.9999	.9975	.9755	.8740	.6113	.2418	.0221
	13				1.0000	.9995	.9936	.9536	.7981	.4511	.0826
	14					.9999	.9988	.9877	.9226	.6904	.2382
	15					1.0000	.9999	.9979	.9807	.8818	.5182
	16						1.0000	.9998	.9977	.9775	.8332
	17							1.0000	1.0000	1.0000	1.0000
18	0	.1501	.0180	.0056	.0016	.0001	.0000				
	1	.4503	.0991	.0395	.0142	.0013	.0001				
	2	.7338	.2713	.1353	.0600	.0082	.0007	.0000			
	3	.9018	.5010	.3057	.1646	.0328	.0038	.0002			
	4	.9718	.7164	.5787	.3327	.0942	.0154	.0013	.0000		
	5	.9936	.8671	.7175	.5344	.2088	.0481	.0058	.0003		
	6	.9988	.9487	.8610	.7217	.3743	.1189	.0203	.0014	.0000	
	7	.9998	.9837	.9431	.8593	.5634	.2403	.0576	.0061	.0002	
	8	1.0000	.9957	.9807	.9404	.7368	.4073	.1347	.0210	.0009	
	9		.9991	.9946	.9790	.8653	.5927	.2632	.0596	.0043	.0000
	10		.9998	.9988	.9939	.9424	.7597	.4366	.1407	.0163	.0002
	11		1.0000	.9998	.9986	.9797	.8811	.6257	.2783	.0513	.0012
	12			1.0000	.9997	.9942	.9519	.7912	.4656	.1329	.0064
	13				1.0000	.9987	.9846	.9058	.6673	.2836	.0282
	14					.9998	.9962	.9672	.8354	.4990	.0982
	15					1.0000	.9993	.9918	.9400	.7287	.2662
	16						.9999	.9987	.9858	.9009	.5497
	17						1.0000	.9999	.9984	.9820	.8499
	18							1.0000	1.0000	1.0000	1.0000



Binomial Probability Sums $\sum_{x=0}^{r} b(x; n, p)$

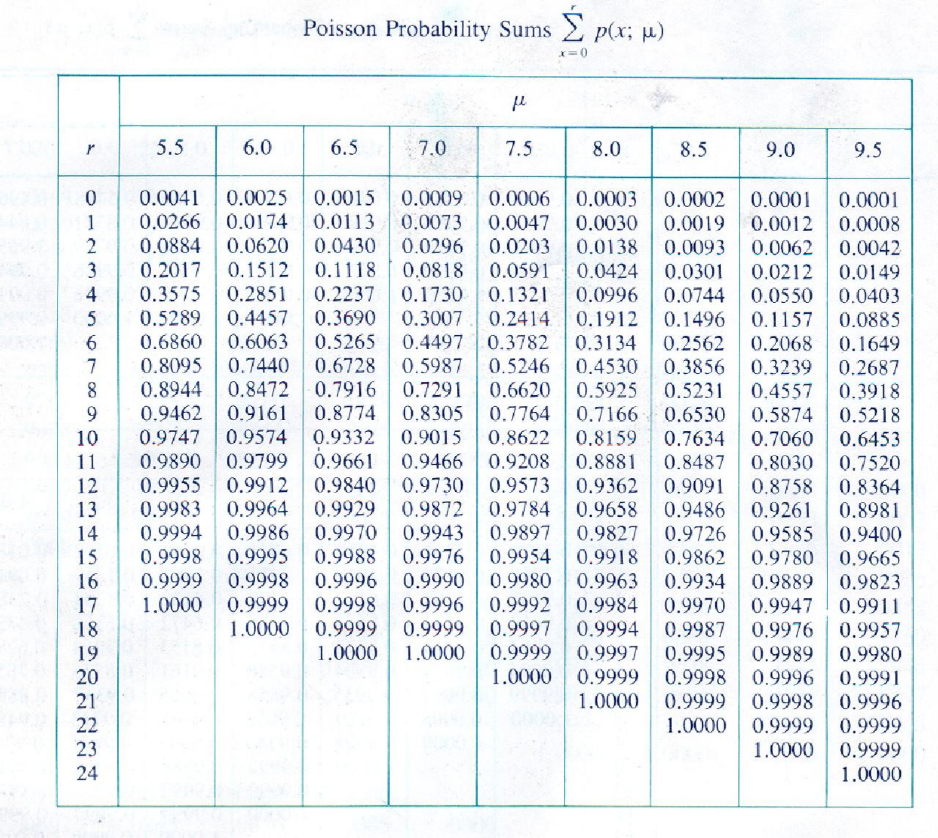
						p				
n	*r*	.10	.20	.25	.30	.40	.50	.60	.70	.80
19	0	.1351	.0144	.0042	.0011	.0001				
	1	.4203	.0829	.0310	.0104	.0008	.0000			
	2	.7054	.2369	.1113	.0462	.0055	.0004	.0000		
	3	.8850	.4551	.2631	.1332	.0230	.0022	.0001		
	4	.9648	.6733	.4654	.2822	.0696	.0096	.0006	.0000	
	5	.9914	.8369	.6678	.4739	.1629	.0318	.0031	.0001	
	6	.9983	.9324	.8251	.6655	.3081	.0835	.0116	.0006	
	7	.9997	.9767	.9225	.8180	.4878	.1796	.0352	.0028	.0000
	8	1.0000	.9933	.9713	.9161	.6675	.3238	.0885	.0105	.0003
	9		.9984	.9911	.9674	.8139	.5000	.1861	.0326	.0016
	10		.9997	.9977	.9895	.9115	.6762	.3325	.0839	.0067
	11		.9999	.9995	.9972	.9648	.8204	.5122	.1820	.0233
	12		1.0000	.9999	.9994	.9884	.9165	.6919	.3345	.0676
	13			1.0000	.9999	.9969	.9682	.8371	.5261	.1631
	14				1.0000	.9994	.9904	.9304	.7178	.3267
	15					.9999	.9978	.9770	.8668	.5449
	16					1.0000	.9996	.9945	.9538	.7631
	17						1.0000	.9992	.9896	.9171
	18							.9999	.9989	.9856
	19							1.0000	1.0000	1.0000
20	0	.1216	.0115	.0032	.0008	.0000				
	1	.3917	.0692	.0243	.0076	.0005	.0000			
	2	.6769	.2061	.0913	.0355	.0036	.0002	.0000		
	3	.8670	.4114	.2252	.1071	.0160	.0013	.0001		
	4	.9568	.6296	.4148	.2375	.0510	.0059	.0003		
	5	.9887	.8042	.6172	.4164	.1256	.0207	.0016	.0000	
	6	.9976	.9133	.7858	.6080	.2500	.0577	.0065	.0003	
	7	.9996	.9679	.8982	.7723	.4159	.1316	.0210	.0013	.0000
	8	.9999	.9900	.9591	.8867	.5956	.2517	.0565	.0051	.0001
	9	1.0000	.9974	.9861	.9520	.7553	.4119	.1275	.0171	.0006
	10		.9994	.9961	.9829	.8725	.5881	.2447	.0480	.0026
	11		.9999	.9991	.9949	.9435	.7483	.4044	.1133	.0100
	12		1.0000	.9998	.9987	.9790	.8684	.5841	.2277	.0321
	13			1.0000	.9997	.9935	.9423	.7500	.3920	.0867
	14				1.0000	.9984	.9793	.8744	.5836	.1958
	15					.9997	.9941	.9490	.7625	.3704
	16					1.0000	.9987	.9840	.8929	.5886
	17						.9998	.9964	.9645	.7939
	18						1.0000	.9995	.9924	.9308
	19							1.0000	.9992	.9885
	20								1.0000	1.0000



Poisson Probability Sums $\sum_{x=0} p(x; \mu)$

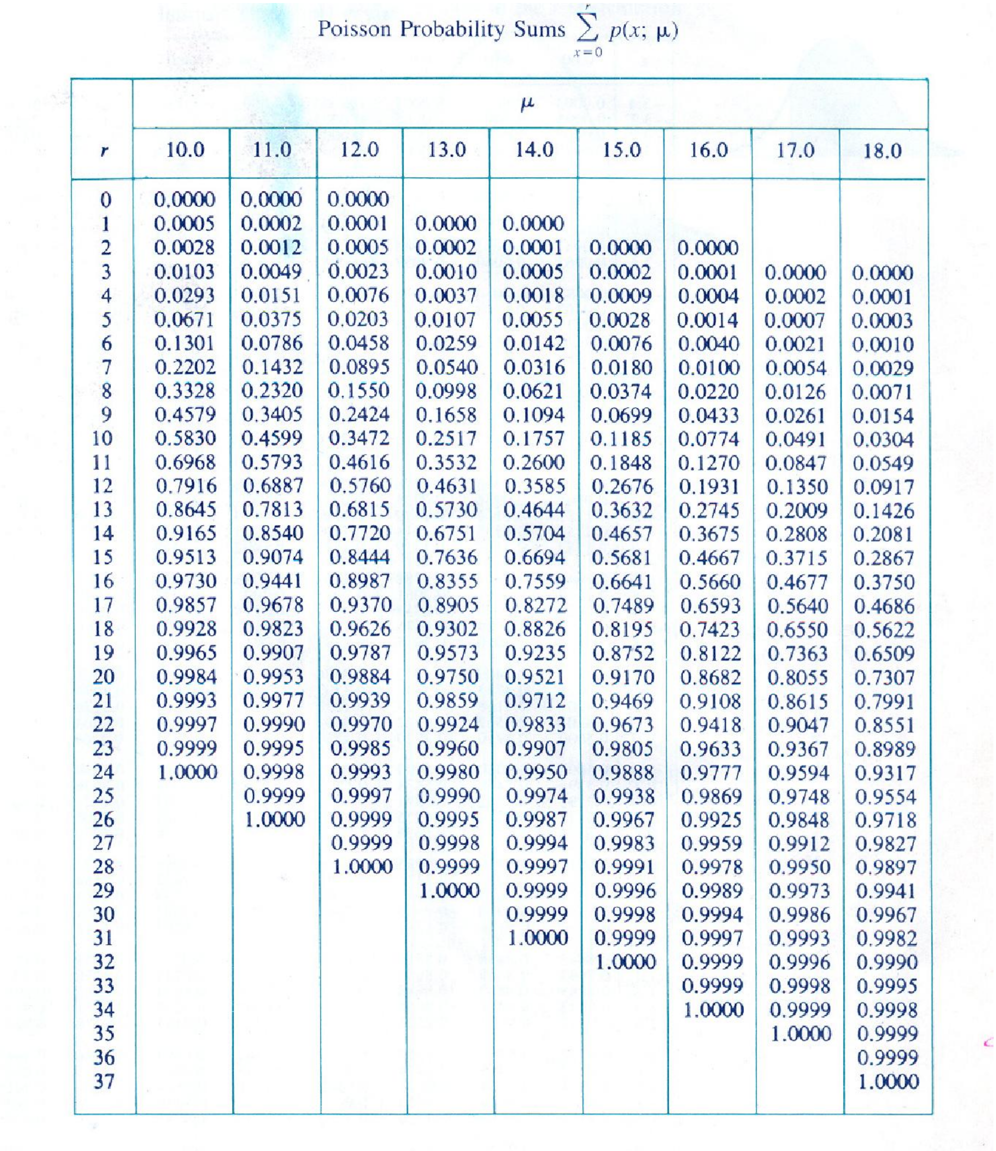
	μ								
r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.9048	0.8187	0.7408	0.6730	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865
4		1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977
5				1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
6							1.0000	1.0000	1.0000

	μ								
r	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
0	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247
3	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650
4	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160
6	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8		1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9			1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10				0.9999	0.9997	0.9990	0.9972	0.9933	0.9863
11				1.0000	0.9999	0.9997	0.9991	0.9976	0.9945
12					1.0000	0.9999	0.9997	0.9992	0.9980
13						1.0000	0.9999	0.9997	0.9993
14							1.0000	0.9999	0.9998
15								1.0000	0.9999
16									1.0000



Poisson Probability Sums $\sum_{x=0}^{r} p(x;\mu)$

	μ								
r	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5
0	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001
1	0.0266	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008
2	0.0884	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042
3	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149
4	0.3575	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403
5	0.5289	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885
6	0.6860	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649
7	0.8095	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687
8	0.8944	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918
9	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218
10	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453
11	0.9890	0.9799	0.9661	0.9466	0.9208	0.8881	0.8487	0.8030	0.7520
12	0.9955	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364
13	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981
14	0.9994	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400
15	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665
16	0.9999	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823
17	1.0000	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911
18		1.0000	0.9999	0.9999	0.9997	0.9994	0.9987	0.9976	0.9957
19			1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980
20					1.0000	0.9999	0.9998	0.9996	0.9991
21						1.0000	0.9999	0.9998	0.9996
22							1.0000	0.9999	0.9999
23								1.0000	0.9999
24									1.0000



Poisson Probability Sums $\sum_{x=0}^{r} p(x;\ \mu)$

	μ								
r	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0
0	0.0000	0.0000	0.0000						
1	0.0005	0.0002	0.0001	0.0000	0.0000				
2	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000		
3	0.0103	0.0049	0.0023	0.0010	0.0005	0.0002	0.0001	0.0000	0.0000
4	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001
5	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003
6	0.1301	0.0786	0.0458	0.0259	0.0142	0.0076	0.0040	0.0021	0.0010
7	0.2202	0.1432	0.0895	0.0540	0.0316	0.0180	0.0100	0.0054	0.0029
8	0.3328	0.2320	0.1550	0.0998	0.0621	0.0374	0.0220	0.0126	0.0071
9	0.4579	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0154
10	0.5830	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0304
11	0.6968	0.5793	0.4616	0.3532	0.2600	0.1848	0.1270	0.0847	0.0549
12	0.7916	0.6887	0.5760	0.4631	0.3585	0.2676	0.1931	0.1350	0.0917
13	0.8645	0.7813	0.6815	0.5730	0.4644	0.3632	0.2745	0.2009	0.1426
14	0.9165	0.8540	0.7720	0.6751	0.5704	0.4657	0.3675	0.2808	0.2081
15	0.9513	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2867
16	0.9730	0.9441	0.8987	0.8355	0.7559	0.6641	0.5660	0.4677	0.3750
17	0.9857	0.9678	0.9370	0.8905	0.8272	0.7489	0.6593	0.5640	0.4686
18	0.9928	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.6550	0.5622
19	0.9965	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.6509
20	0.9984	0.9953	0.9884	0.9750	0.9521	0.9170	0.8682	0.8055	0.7307
21	0.9993	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7991
22	0.9997	0.9990	0.9970	0.9924	0.9833	0.9673	0.9418	0.9047	0.8551
23	0.9999	0.9995	0.9985	0.9960	0.9907	0.9805	0.9633	0.9367	0.8989
24	1.0000	0.9998	0.9993	0.9980	0.9950	0.9888	0.9777	0.9594	0.9317
25		0.9999	0.9997	0.9990	0.9974	0.9938	0.9869	0.9748	0.9554
26		1.0000	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9718
27			0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9827
28			1.0000	0.9999	0.9997	0.9991	0.9978	0.9950	0.9897
29				1.0000	0.9999	0.9996	0.9989	0.9973	0.9941
30					0.9999	0.9998	0.9994	0.9986	0.9967
31					1.0000	0.9999	0.9997	0.9993	0.9982
32						1.0000	0.9999	0.9996	0.9990
33							0.9999	0.9998	0.9995
34							1.0000	0.9999	0.9998
35								1.0000	0.9999
36									0.9999
37									1.0000