Version 1

Math, FOIT

Part -I

Shafiq Ur Rehman

Lecturer Statistics
Department of Mathematics
Faculty of Science and Technology
University of Central Punjab, Lahore.

Table Of Contents

LECTURE 1: INTRODUCTION TO STATISTICS

1.1 DEFINITIONS OF STATISTICS

1. Statistics is a discipline concerned with collection of data, presentation of data, summarizing and analyzing the data, and making inference about the population.
2. Statistics is concerned with scientific method for collecting, organizing, summarizing, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

These definitions cover all the aspects and then tries to link them up with decision-making. After all, Statistics as a subject must help one to reach a reasonable and appropriate decision on the basis of the analysis of numerical data collected earlier.

1.1.1 TYPES OF STATISTICS

Descriptive Statistics:

1. Descriptive Statistics deals with the all methods of data collection, presentation, organization and summarizations techniques to find the important aspect from the collected data.
2. It includes any treatment designed to describe or summarize the given data. Such summaries of data, which may be tabular, graphical, or numerical, are referred to as *descriptive statistics*.
3. A collection of methods that enable us to organize, display and describe data using such devices as tables, graphs and summary measures.

Statistical Inference:

1. A collection of methods that enable us in making decisions about a population based on sample results.
2. Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data.
3. Inferential Statistics uses a number of quantitative techniques that enable us to make appropriate generalizations from limited observations.
4. Statistical Inference uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as statistical inference.

1.2 PROBABILITY, AND KEY TERMS

Some definitions of the probability

1. The quantitative measure of an uncertainty is called probability.

2. Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring.

E.g., If you toss a fair coin, The expected theoretical probability of heads in any one toss is 1/2 or 0.5.

Denoted by P(A) = $\frac{n(A)}{n(S)}$;

where, P(A) is the probability of an event;

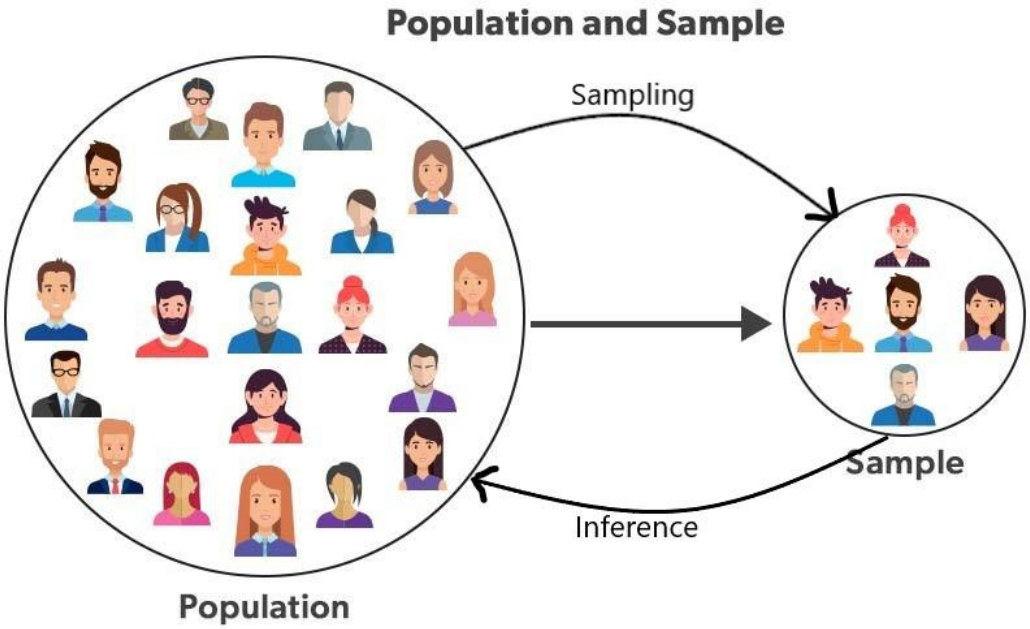
n(A) is the total number of outcomes in favor to event A

n(S) is the total number of outcomes of a random experiment.

1.3 KEY-TERMS

POPULATION: A population is the set of all elements of interest in a particular study. N is the total number of observations/elements/individuals in the population.

SAMPLE: A sample is a representative subset of the population. n is the total number of observations/elements/individuals in the sample.



PARAMETER: are the numerical results computed from population observations.

STATISTIC: are the numerical results computed from the sample observations.

1.4 SAMPLING:

Sampling is a procedure to select a sampling from a population is called Sampling.

There are two types of sampling,

1.4.1 1- PROBABILITY SAMPLING

When each and every sampling unit of a population has a known (non-zero) probability of its being included in the sample is called Probability Sampling.

i- Simple Random Sampling

ii- Stratified Sampling

iii- Systematic Sampling

iv- Cluster Sampling

1.4.2 2- NON-PROBABILITY SAMPLING

When some sampling units of a population has a zero probability of its being included in the sample is called Probability Sampling.

i- Purposive Sampling

ii- Judgmental Sampling

iii- Quota Sampling

1.5 OBSERVATION AND VARIABLE

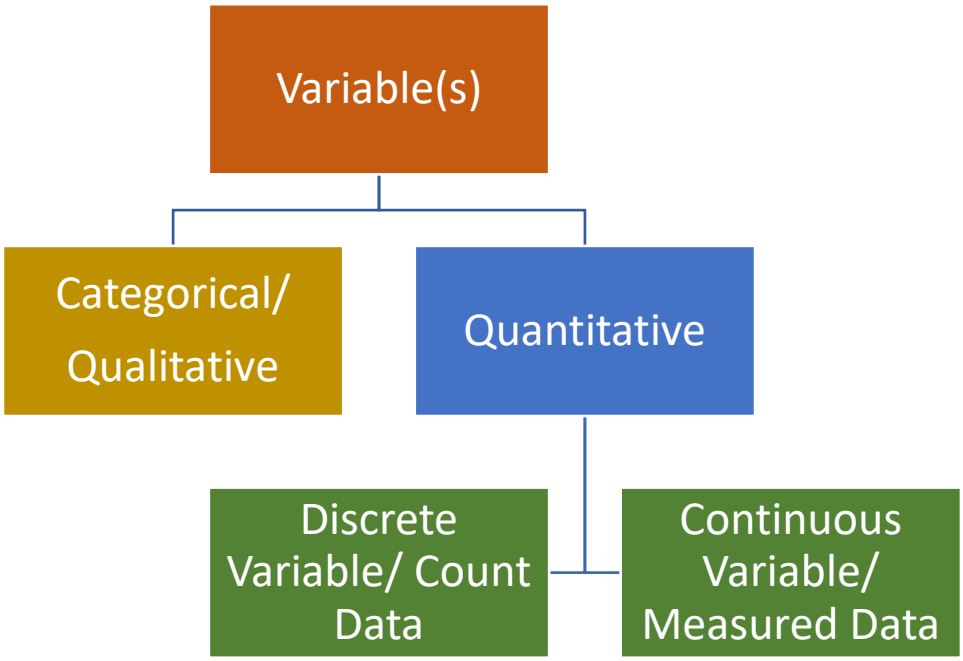
Observation:

- Any sort of numerical recording(s) of some or about some characteristic of interest is called observation(s).
- Observation is one of the methods of collecting data. It is used to get both past and current information.

Variable(s):

Variable(s) is the attribute or characteristic of interest whose values/observations changes individual by individual. Variable is denoted by the capital letters *i.e.*, X, Y, Z. Variables may be numerical or categorical.

1.5.1 TYPES OF VARIABLES:



1- Categorical/Qualitative Variable:

- Variables that can be expressed by a non-numerical property such as satisfaction of a customer, rich, poor and superior.
- Qualitative variable/data are the results of categorizing or describing attributes of a population. Qualitative data are also often called categorical data. Hair color, blood type, ethnic group, the car a person drives.

2- Quantitative Variable:

- Variables that can be expressed numerically expressed. For example, weight, height, income, expenditure and price are quantitative data.
- Quantitative variable/data are always numbers. Quantitative data are the result of counting or measuring attributes of a population.

1.5.2 TYPES OF QUANTITATIVE VARIABLE

1- Discrete Variable or Count Variable:

- A variable that can take only a complete count or whole number is called Discrete Variable. E.g., you can count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

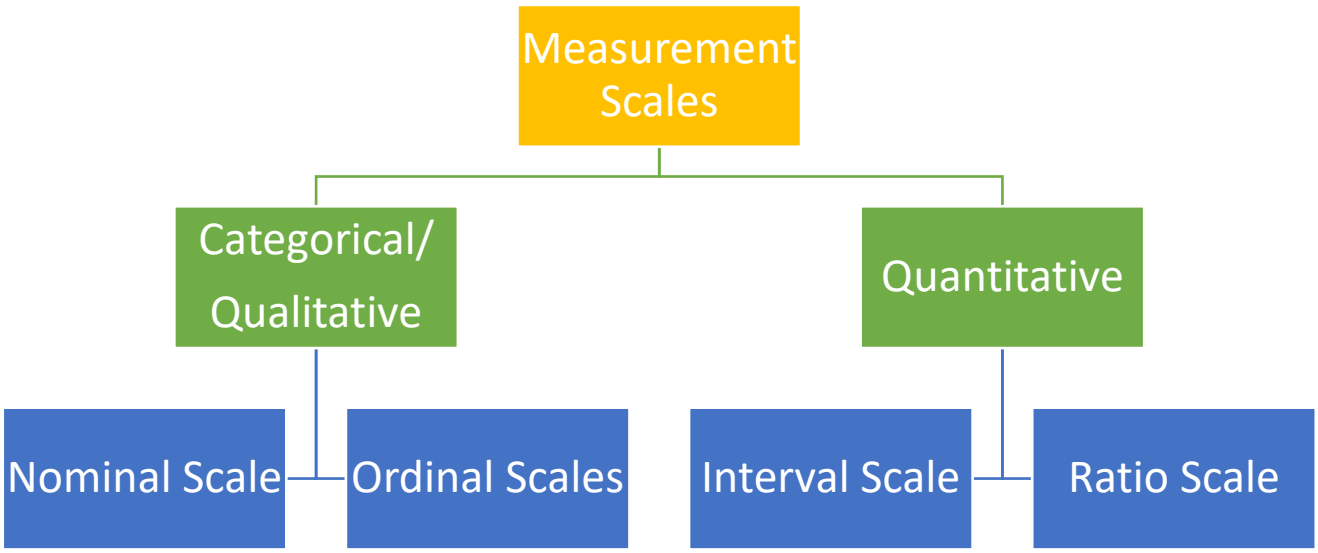
- All data that are the result of counting are called quantitative discrete variable. These data take on only certain numerical values.

2- Continuous Variable or Measured Variable:

- A variable that can take fractional or decimal values, or can take any possible values between an interval is called continuous variable or measured variable. Continuous data are often the results of measurements like lengths, weights, or times.
- Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called quantitative continuous variable. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

1.6 LEVELS OF MEASUREMENT

Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.



Nominal Scale: When the data for a variable consist of labels or names used to identify an attribute of the element. E.g., Gender (Male, Female, Transgender). Eye color (Blue, Green, Brown, Hazel). Type of house (Bungalow, Duplex, Ranch). Type of pet (Dog, Cat, Rodent, Fish, Bird).

Ordinal Scale: If the data exhibit the properties of nominal data and in addition, the order or rank of the data is meaningful. E.g., High school class rankings: 1st, 2nd, 3rd etc. Social economic class: working, middle, upper. The Likert Scale: agree, strongly agree, disagree etc.

Interval Scale: You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point and ratios are meaningless. E.g., The difference between any two adjacent temperatures is the same: one degree. But zero degrees is defined differently depending on the scale – it doesn't mean an absolute absence of temperature.

Ratio Scale: When the data can categorize, rank, and infer equal intervals between neighboring data points, and there is a true zero point. E.g., A true zero means there is an absence of the variable of interest. In ratio scales, zero does mean an absolute lack of the variable.

PRACTICE QUESTIONS

Page 48.
Questions
1-5, 11, 16, 39, 40, 53-64, 80-86

LECTURE 2: DATA COLLECTION EXPERIMENT

2.1 EXPERIMENTAL DESIGN AND ETHICS

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**.

In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable.

Treatments: The different values of the explanatory variable are called treatments.

Experimental unit is a single object or individual to be measured.

2.2 DATA COLLECTION EXPERIMENT

Collection of Data: It is a procedure of gathering the information regarding to the variable(s) of a particular study. That is a very important step in the field of statistics. There are two types of data that can be collected, which are as follows

- **Primary Data:** Primary data is the data that is collected for the first time through personal experiences or evidence, particularly for research. It is also described as raw data or first-hand information. The mode of assembling the information is costly, as the analysis is done by an agency or an external organization, and needs human resources and investment. The investigator supervises and controls the data collection process directly. The data is mostly collected through observations, physical testing, mailed questionnaires, surveys, personal interviews, telephonic interviews, case studies, and focus groups, etc.

- **Secondary Data:** Secondary data is a second-hand data that is already collected and recorded by some researchers for their purpose, and not for the current research problem. It is accessible in the form of data collected from different sources such as government publications, censuses, internal records of the organization, books, journal articles, websites and reports, etc.

  This method of gathering data is affordable, readily available, and saves cost and time. However, the one disadvantage is that the information assembled is for some other purpose and may not meet the present research purpose or may not be accurate.

Comparison Between Primary and Secondary Data

Primary	Secondary
Originality	
These are original because these are collected by the investigator for the first time.	These are not original because someone else has collected these for his own purpose.
Reliability and Suitability	
These are more reliable and suitable for the enquiry because these are collected for a particular purpose.	These are less reliable and less suitable as someone else has collected the data which may not perfectly match our purpose.
Time and Money	
Collecting primary data is quite expensive both in the terms of time and money.	Secondary data requires less time and money; hence it is economical.

2.3 MEASUREMENT ERROR

The difference between the actual value of the population and the sample estimated value is known as measurement error *i.e.*, $e = \mu - \bar{x}$, where μ is population result and $\bar{x}$ is sample result.
There are two types of errors.

1- Sampling Error: The natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

2- Non-Sampling Error: An issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

2.4 DATA ARRAY

The arrangement of data in ascending or descending order is called an array.

When we collect data, there are not one or two observations but a large number of them. Even the data array, particularly when the number of observations is very large, is not very helpful for any

analysis. Thus, it becomes necessary to organize the mass of data so that they are reduced to meaningful proportions. This brings us to tabular representation.

Example: Suppose there is a class of 30 management students and each student in the class is asked to toss a coin five times and record each time whether he gets a head. As a result of this experiment, the following figures emerge for the 30 students:

3, 2, 0, 4, 1, 2, 3, 2, 5, 3, 3, 1, 1, 3, 5, 4, 2, 2, 3, 1, 0, 4, 3, 2, 2, 4, 2, 3, 3, 1

It is seen that even with only 30 observations the data need some better display. One way of doing this is to show the data in a certain order. For instance, we may show the same data in an ascending order as follows:

0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5

2.5 CLASSIFICATION:

Classification means, data can be classified into groups/classes according to any different criterion. The classification schemes can be either qualitative or quantitative and either discrete or continuous. There are different ways to classify the data set into groups. Which are as follows

- **Simple or One-Way Classification**
  When the data set is divided into groups according to only one criterion.
- **Areal or Two-Way Classification**
  When the data set is divided into groups according to two different criterions.
- **Manifold or Cross Classification**
  When the data set is divided into groups according to three or more different criterion.

2.6 THE FREQUENCY TABLE: IMPORTANT POINTS

Before creating the frequency table, we should have a clear idea of certain terms, which we shall come across frequently.

Class Limits: These are the lowest and the highest values of a class. For example, take the class 30-50. Here, we find that the lowest limit is 30 and the highest limit is 50. When we categories individual observations within this class, it is clear that none of the included observations is below 30 or above 50.

Class Interval: The difference between the two successive lower or upper limits is known as a class interval. It is the width between the classes. Thus, the class interval of class 30-49 and 50-69 is 20 as the difference between two successive lower limits is 50-30=20.

Class Frequency: The number of observations belonging to a particular class is known as the frequency of that class or the class frequency.

Class mid-point: When we add up the lower and the upper-class limits of a class interval, we get a certain value. This value is divided by two, which gives us the class mid-point.

$$class\ mid-point = \frac{Lower\ class\ limit + upper\ class\ limit}{2}$$

2.7 DISCRETE SERIES: FREQUENCY TABLE

The arrangement and display of data in the form where the observed value is paired with its frequency is called a frequency distribution.

It can be noticed that the data relating to the toss of a coin five times by 30 students show that each figure 0 to 5 has occurred a certain number of times. The frequency of their occurrence varies in each case. The same data can be shown differently. We can condense these data by pairing each of that value with its frequency. This is shown in the following table,

Observation	Frequency
0	2
1	5
2	8
3	9
4	4
5	2

It can be seen from the Table that the largest number of frequencies are against observation 3 while the lowest number of frequencies are against observation 0 and 5.

2.7.1 THE FREQUENCY TABLE: RELATIVE FREQUENCY

A frequency distribution in which every observed value is paired with its relative frequency is known as a relative frequency distribution. **This is very simple; the concerned frequency is divided by the number of observations or the total frequencies.**

Observation	Frequency	(*f*/N)	(*f*/N)*100
0	2	2/30 = 0.07	7
1	5	5/30 = 0.17	17
2	8	0.27	27
3	9	0.30	30
4	4	0.12	12
5	2	0.07	7
Sum	30	1.00	100%

As a frequency distribution facilitates us in condensing a large mass of data, the process of data analysis and interpretation also become far more manageable than it would have been otherwise.

2.8 FORMATION OF A GROUPED FREQUENCY TABLE

The formation of a frequency distribution table comprises the following steps:

1. Deciding the appropriate **number of class/groups and the Range**

$k = 1+3.3\log(N)$

$R = X_m - X_0$

The number of class intervals depends mainly on the number of observations as well as their range. As a general rule, the number of classes should not be less than six nor should be more than 15. If the number of observations is small, obviously the classes will be few as we cannot classify small data into12 or 15 classes. If the classes are too few, then the original data will be so compressed that only limited information will be available.

2. Choosing a suitable size or width of a **class interval**

$h \text{ or } c = R/K$

Another major consideration while forming a frequency table is the size of the class width. It is desirable to have each class grouping of equal width. In order to ascertain the width of each class, the difference between the highest value and the lowest value, which is known as the range, should be divided by the number of class groupings desired.

FORMATION OF A GROUPED FREQUENCY TABLE STEPS:

1. Arrange the data into an array.
2. Find the Range of the data set.
3. Compute the required No. of Groups *i.e.,* denoted by K = 1+3.3log(N)
4. Find the interval between the classes *i.e.,* h = R/K (Range divided by No of Groups). For class interval, roundup the result by one value.
5. Establishing the Class Limits for each group
6. Classifying the data into the appropriate classes
7. Counting the number of items (i.e. frequency) in each class.
8. Compute Class Boundaries
9. Find Mid points, Cumulative Frequency

Inclusive Method (Class Limits): In the case of inclusive method, the upper limit of one class is included in that class itself. Suppose we have the following frequency distribution:

Profits (Rs in lakh)	*Number of Companies*
10–19	12
20–29	17
30–39	30
40–49	25
50–59	16
Total	**100**

In this case, the class intervals are formed on the basis of inclusive method, which shows that the upper limit of the class interval in reality is taken in that class.

Exclusive Method (Class Boundaries): In the case of inclusive method, the upper limit of one class is not included in that class itself. Suppose we have the following frequency distribution:

Profits Earned by Companies

Profits (Rs in lakh)	*Number of Companies*
10–20	12
20–30	17
30–40	30
40–50	25
50–60	16
Total	**100**

This is an exclusive method where the upper limit of one class is shown as the lower limit of the next class. This ensures continuity of data.

Example: The following data pertain to weights (in kg) of 33 students of a class:

42, 74, 40, 60, 82, 115, 41, 61, 75, 83, 63, 53, 110, 76, 84, 50, 67, 65, 78, 77, 56, 95, 68, 69, 104, 80, 79, 79, 54, 73, 59, 81 and 110.

Prepare a suitable frequency table.

Solution:

Arrange the values in ascending order.

40, 41, 42, 50, 53, 54, 56, 59, 60, 61, 63, 65, 67, 68, 69, 73, 74, 75, 76, 77, 78, 79, 79, 80, 81, 82, 83, 84, 95, 104, 110, 110, 115

Steps:

1. Range = Max – Min = 115 – 40 = 75
2. No of groups = K = 1+3.3log(N) = 1+3.3log (33) = 6.01 ≈ 6
3. Interval = h = R / K = 75/6 = 12.5 ≈ 13

Class Limits	Frequency	Class Boundaries	Mid-Points	Cumulative Frequency
40 – 52	4	39.5 - 52.5	46	4
53 – 65	8	52.5 – 65.5	59	12
66 – 78	9	65.5 – 78.5	72	21
79 – 91	7	78.5 – 91.5	85	28
92 – 104	2	91.5 – 104.5	98	30
105 - 117	3	104.5 – 117.5	111	33
Sum	33			

PRACTICE QUESTIONS

Page 48.
Questions
1-5, 11, 16, 39, 40, 53-64, 80-86

LECTURE 3: DESCRIPTIVE STATISTICS

3.1 STEM-AND-LEAF DIAGRAMS

This is a technique used for simultaneously sorting and displaying the data sets, where each number in the data set is divided into two parts, a *Stem* and a *Leaf*. A Stem is a leading digit(s) of each number and used in sorting, while Leaf is the rest of the number or trailing digit(s). A vertical line separates the leaf from stem. E.g., the numbers 25, 31, 45, 243 split into two parts. 1st digit 2 is the stem and 5 is the leaf digit. Similarly, 3 will be stem and 1 is leaf, the number 243, 24 is the stem and 3 is leaf digit. It will be displayed as

STEM	LEAF
2	5
3	1
4	5
24	3

EXAMPLE: For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

SOLUTION:

STEM	LEAF
3	3
4	2 9 9
5	3 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

3.2 HISTOGRAMS

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labelled with what the data represents (for instance, distance from your home to school). The vertical axis is labelled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then: (The formula for relative frequency)

$$RF = f / n$$

For example, if three students in Mr. Shafiq's Statistics class of 40 students received from 85% to 100%, then, f = 3, n = 40, and RF = f / n = 3/40 = 0.075. 7.5% of the students received 85–100%. 85–100% are quantitative measures.

3.2.1 STEPS TO CONSTRUCT THE HISTOGRAM:

1. First check how many bars or intervals, also called classes, representing the data. Generally, many histograms consist of five to 15 bars or classes.
2. A convenient starting point.
3. The height of the bars depends on the frequency of that class or the relative frequency.
4. There will be no gap between the classes' bars.

Example:

The following data are the heights (in inches to the nearest half inch) of 100 male students of the UCP. The heights are continuous data, since height is measured.

60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5
66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67;
67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74

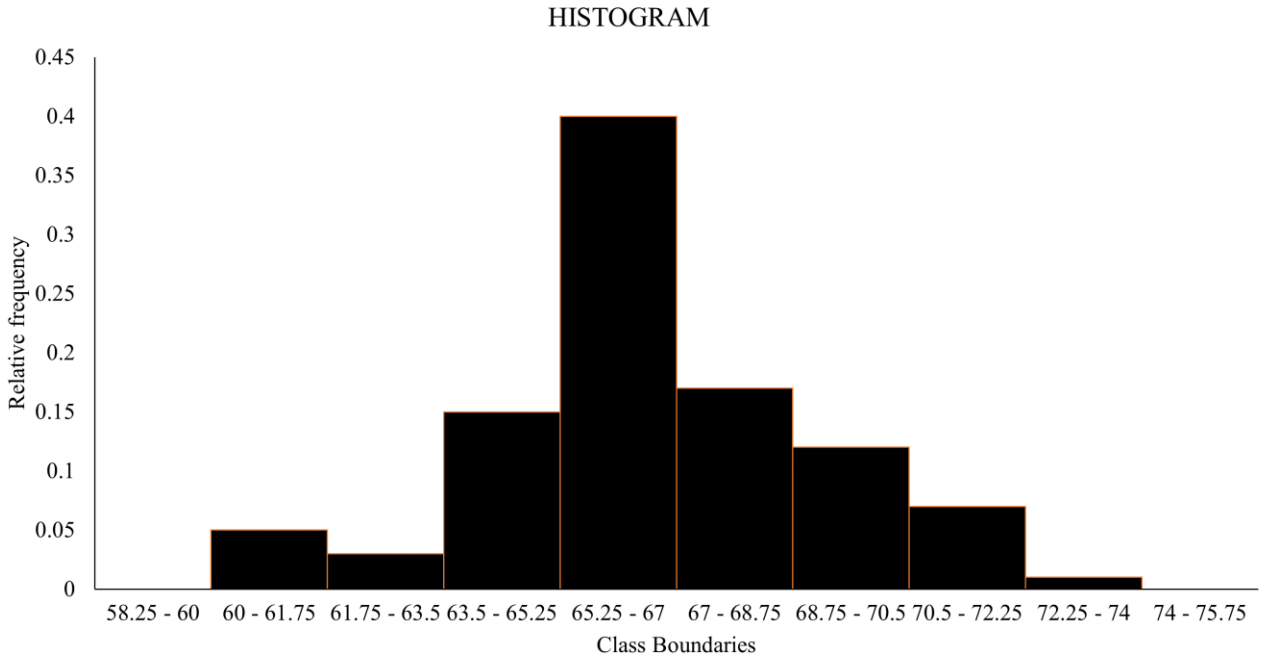
Solution:

Total number of classes = k = 8;

Range = 60-74 =14;

Interval = R/K = 1.75

Class-Boundaries	*f*	*R.F.*	Cumulative Freq.
60 - 61.75	5	0.05	5
61.75 - 63.5	3	0.03	8
63.5 - 65.25	15	0.15	23
65.25 - 67	40	0.4	63
67 - 68.75	17	0.17	80
68.75 - 70.5	12	0.12	92
70.5 - 72.25	7	0.07	99
72.25 - 74	1	0.01	100



PRACTICE QUESTIONS

Page 125
Questions
1-7, 12-20

LECTURE 4: MEASURES OF THE LOCATION

4.1 A FORMULA FOR FINDING THE KTH PERCENTILE

If you were to do a little research, you would find several formulas for calculating the kth percentile. Here is one of them.

k = the kth percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate i = $\frac{k}{100}(n+1)$
- If i is an integer, then the kth percentile is the data value in the ith position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

EXAMPLE:

Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.

b. Find the 83rd percentile.

SOLUTION:

a. k = 70

i = the index

n = 29

i = $\frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$

Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

b. k = 83rd percentile

i = the index

n = 29

i = $\frac{k}{100}(n+1) = \frac{83}{100}(29+1) = 24.9$; which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

4.2 A FORMULA FOR FINDING THE PERCENTILE OF A VALUE IN A DATA SET

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which
- you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total numbers of data.
- Calculate $\frac{x+0.5y}{n}*100$. Then round to the nearest integer.

EXAMPLE:

Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.

b. Find the percentile for 25.

SOLUTION:

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

X = 18 and y = 1. $\frac{x+0.5y}{n}*100 = \frac{18+0.5(1)}{29}*100$ = 63.80. 59 is the 64th percentile.

b. Counting from the bottom of the list, there are 3 data values less than 25. There is one value of 25.

X = 3 and y = 1. $\frac{x+0.5y}{n}*100 = \frac{3+0.5(1)}{29}*100$ = 12.07. 25 is the 12th percentile.

4.3 MEASURE OF CENTRAL TEDENCY

The measures of central tendency (also called measures of location) are the summarized value of the entire data, they are *The Arithmetic Mean*, *Weighted/Grouped Mean*, *The Median*, *The Mode*, *The Geometric Mean, and The Harmonic Mean*. Before discussing arithmetic mean or any other mean, the question arises: Why should we use such a mean? The answer is that there are two main objects of using mean.

- First, to get a single value that indicates the characteristic of the entire data. For instance, when we talk of per capita income of a country, it gives a broad idea of the standard of living of the people in that country.

- Second, to facilitate comparisons. Measures of central tendency enable us to compare two or more distributions pertaining to the same time period or within the same distribution over time. For example, the average consumption of tea in two different territories for the same period or in a territory for two years, say, 1999 and 2000, can be attempted by means of an average.

4.4 THE ARITHMETIC MEAN

The arithmetic mean is obtained by adding all the observations and dividing the sum by the number of observations. Suppose we have the following observations:

10, 15, 30, 7, 42, 79 and 83

4.4.1 UNGROUPED DATA SET

$\bar{x} = \Sigma x/n$, *where* $\bar{x}$ *is sample mean.*

This formula is the basic formula that forms the definition of arithmetic mean and is used in case of ungrouped data where weights are not involved.

$$= \frac{10+15+30+7+42+79+83}{7}$$

$$= \frac{266}{7} = 38$$

It may be noted that the Greek letter μ is used to denote the mean of the population and N to denote the total number of observations in a population. Thus, the population mean $\mu = \Sigma X/N$.

4.4.2 UNGROUPED WEIGHTED MEAN:

Example:

An investor is fond of investing in equity shares. During a period of falling prices in the stock exchange, a stock is sold at Rs 120 per share on one day, Rs 105 on the next and Rs 90 on the third day. The investor has purchased 50 shares on the first day, 80 shares on the second day and 100 shares on the third day. What average price per share did the investor pay?

Day	*Price Per Share (Rs) x*	*No. of Shares Purchased w*	*Amount Paid Rs wx*
1	120	50	6000
2	105	80	8400
3	90	100	9000
Total	—	**230**	**23,400**

$$\text{Weighted average} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} = \frac{\Sigma wx}{\Sigma w}$$

$$= \frac{6000 + 8400 + 9000}{50 + 80 + 100} = \text{Rs } 101.7$$

Thus, the investor paid an average price of Rs 101.7 per share.

It will be seen that if merely prices of the shares for the three days (regardless of the number of shares purchased) are taken into consideration, then the average price would be

$$\frac{120 + 105 + 90}{3} = \text{Rs } 105$$

4.5 GROUPED DATA (MEAN):

When the data set is given in the frequency table format or the data set is divided into groups corresponding to the frequencies, it is called grouped data.

Example The following table gives the marks of 58 students in Statistics. Calculate the average marks of this group.

Marks	*No. of Students*
0–10	4
10–20	8
20–30	11
30–40	15
40–50	12
50–60	6
60–70	2
Total	**58**

Solution:

Solution **Calculation of Arithmetic Mean by Direct Method**

Marks	*Mid-point m*	*No. of Students f*	*fm*
0–10	5	4	20
10–20	15	8	120
20–30	25	11	275
30–40	35	15	525
40–50	45	12	540
50–60	55	6	330
60–70	65	2	130
			$\Sigma fm = 1940$

$$\bar{x} = \frac{\Sigma fm}{n} = \frac{1940}{58} = 33.45 \text{ marks or } 33 \text{ marks approximately.}$$

It may be noted that the mid-point of each class is taken as a good approximation of the true mean of the class. This is based on the assumption that the values are distributed fairly evenly throughout the interval. When large numbers of frequency occur, this assumption is usually accepted.

Example The mean of the following frequency distribution was found to be 1.46.

No. of Accidents	*No. of Days (Frequency)*
0	46
1	?
2	?
3	25
4	10
5	5
Total	**200 days**

Calculate the missing frequencies.

Solution Here we are given the total number of frequencies and the arithmetic mean. We have to determine the two frequencies that are missing.

Let us assume that the frequency against 1 accident is X and against 2 accidents is Y. If we can establish two simultaneous equations, then we can easily find the values of X and Y.

$$\text{Mean} = \frac{(0\times 46) + (1\times x) + (2\times y) + (3\times 25) + (4\times 10) + (5\times 5)}{200}$$

$$1.46 = \frac{x + 2y + 140}{200}$$

$$x + 2y + 140 = (200)\,(1.46)$$

$$x + 2y = 152 \quad \text{(i)}$$

$$x + y = 200 - \{46 + 25 + 10 + 5\}$$

$$x + y = 200 - 86$$

$$x + y = 114 \quad \text{(ii)}$$

Now subtracting equation (ii) from equation (i), we get

$$\begin{array}{r} x + 2y = 152 \\ x + y = 114 \\ \hline y = 38 \end{array}$$

Substituting the value of $y = 38$ in equation (ii) above, $x + 38 = 114$.
Therefore, $x = 114 - 38 = 76$.
Hence, the missing frequencies are:
Against accident 1 : 76
Against accident 2 : 38

Formula (Arithmetic Mean)	Ungroup	Group
Sample	$\bar{x} = \frac{\sum X_i}{n}$; i =1, 2, …, n	$\bar{x} = \frac{\sum f_i X_i}{n}$; i =1, 2, …, n
Population	$\mu = \frac{\sum X_i}{N}$; i =1, 2, …, N	$\mu = \frac{\sum f_i X_i}{N}$; i =1, 2, …, N

4.6 MERIT AND DEMERIT OF ARITHMETIC MEAN:

Merits:

i. It can be easily calculated; and can be easily understood. It is the reason that it is the most used measure of central tendency.
ii. As every item is taken in calculation, it is affected by every item.
iii. As the mathematical formula is rigid one, therefore the result remains the same.
iv. It can further be subjected to algebraic treatment unlike other measures i.e., mode and median.
v. As it is rigidly defined, it is mostly used for comparing the various issues.

Demerits or Limitations:

i. It is highly affected by extreme values.
ii. It provides a high value on having one very large value in the data set.
iii. Qualitative forms such as Cleverness, Riches etc. cannot give X as data can't be expressed numerically.

PRACTICE QUESTIONS

Page 132
Questions
22-36
Page 137
Questions
110-116

LECTURE 5: CUMULATIVE HISTOGRAM & BOX-PLOT

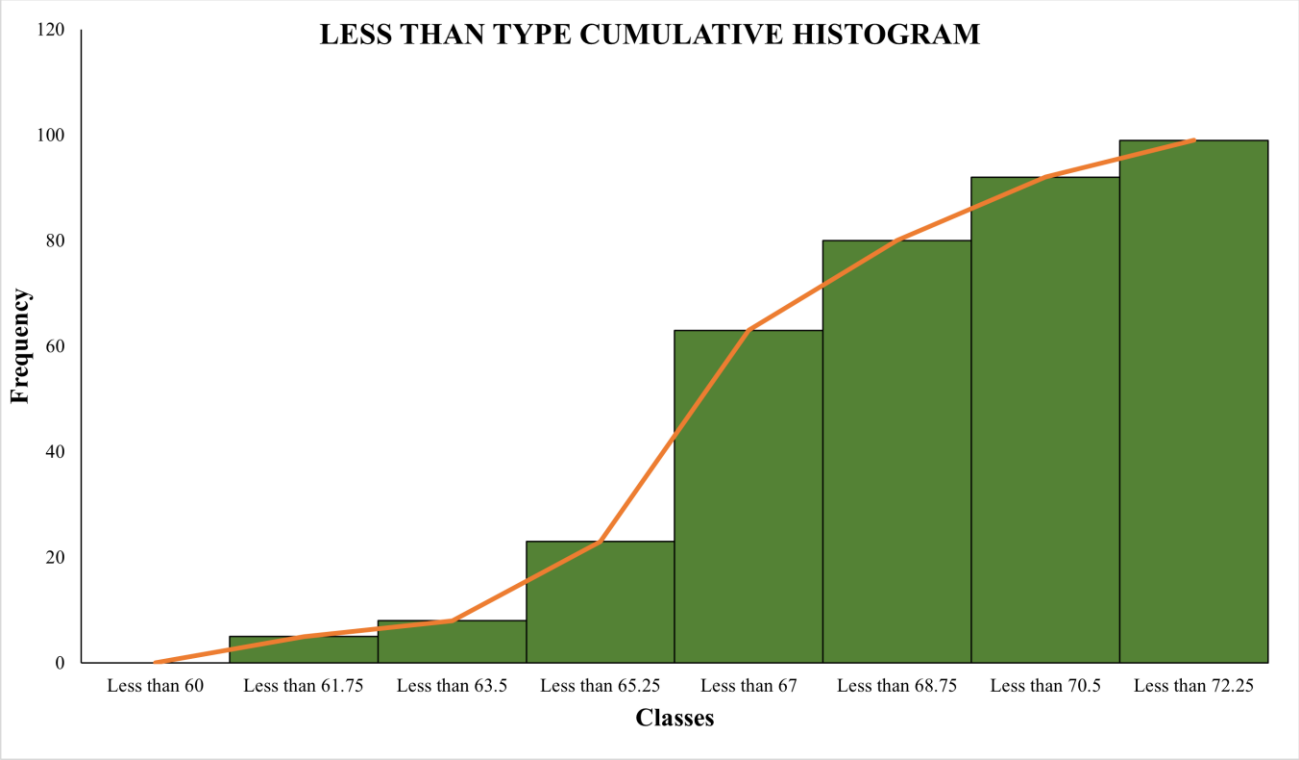
5.1 CUMULATIVE FREQUENCY GRAPHS

A cumulative frequency, mainly known as Ogive is a graph obtain by plotting the cumulated frequencies of a distribution against the upper- or lower-class boundaries. Depending upon whether the cumulation is of "less than" or "more than" type. The curve (also called polygon) should start from zero at the lower boundary of the first interval, the last point is also joined with the upper-class boundary.

For instance, let`s take the example of Histogram. We have the data

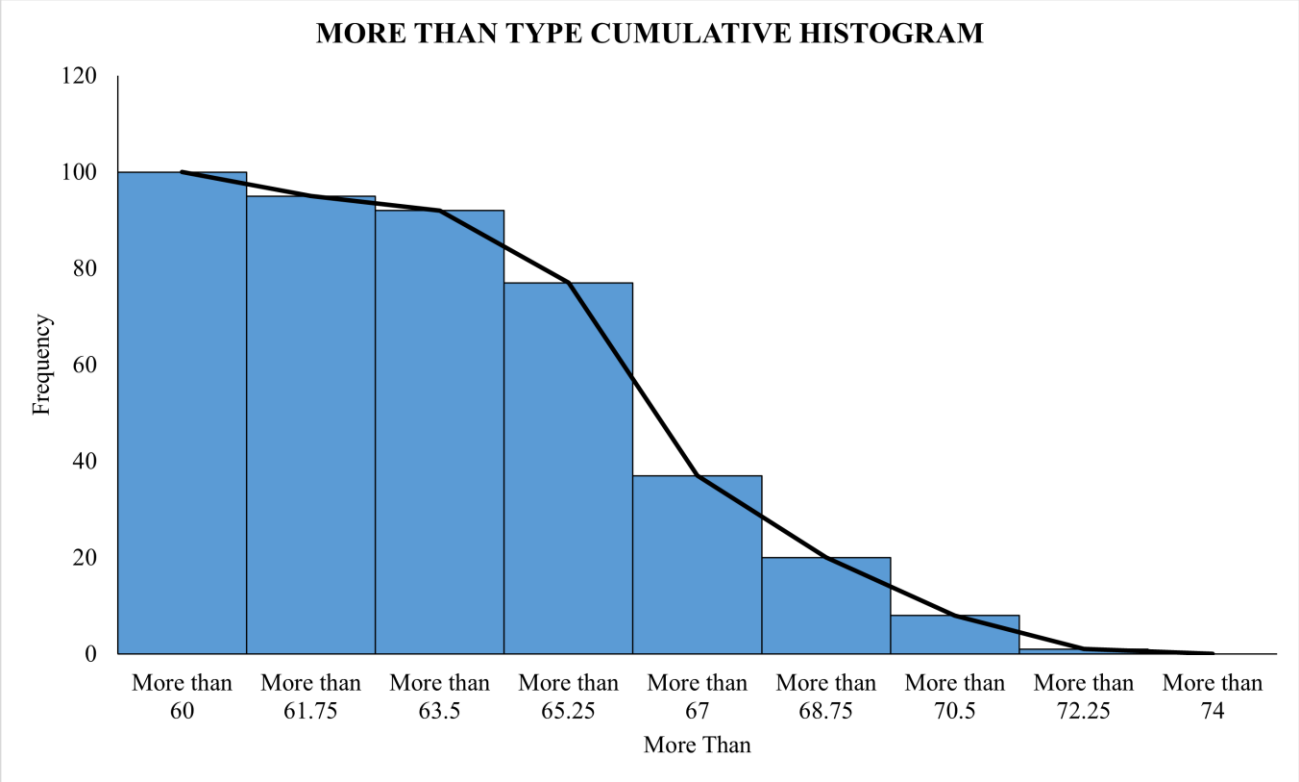
Less than type

Class-Boundaries	*f*	Less than	Cumulative Freq.
60 - 61.75	5	Less than 60	0
61.75 - 63.5	3	Less than 61.75	5
63.5 - 65.25	15	Less than 63.5	8
65.25 - 67	40	Less than 65.25	23
67 - 68.75	17	Less than 67	63
68.75 - 70.5	12	Less than 68.75	80
70.5 - 72.25	7	Less than 70.5	92
72.25 - 74	1	Less than 72.25	99
-	-	Less than 74	100

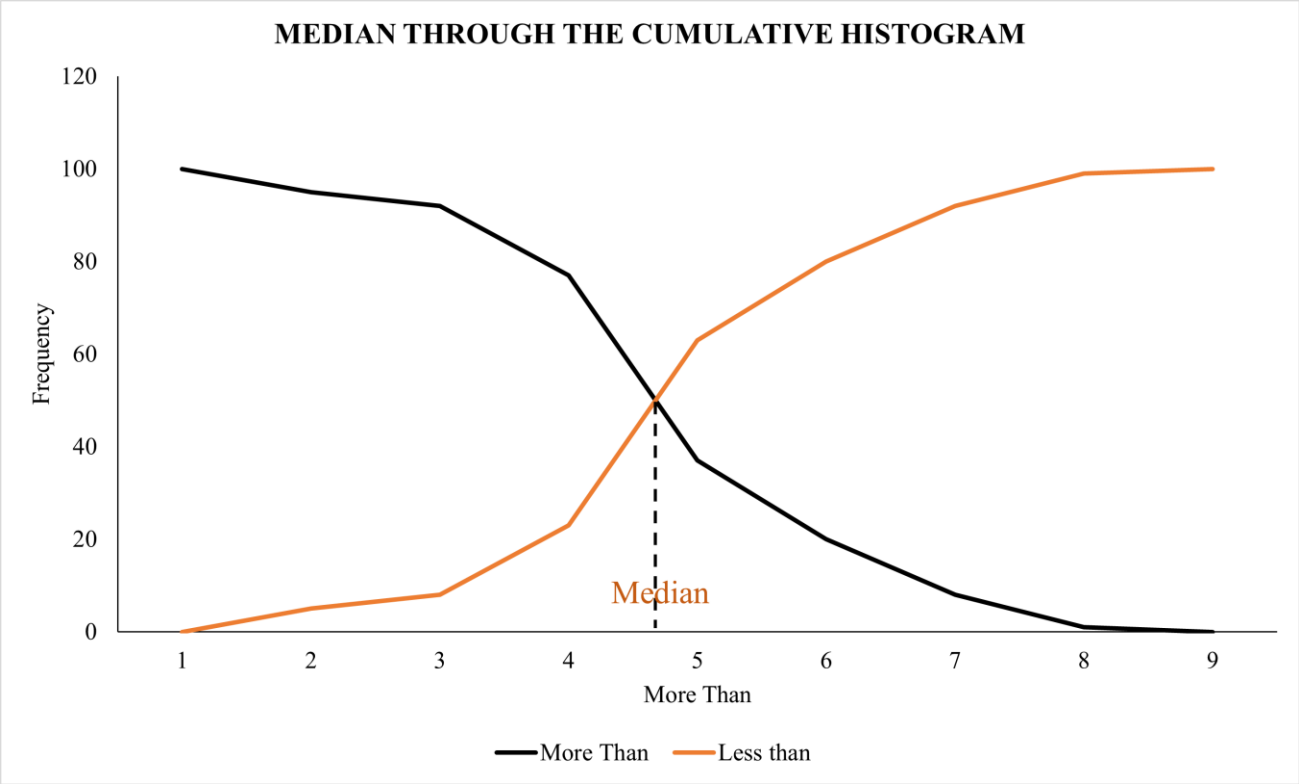


More than type

Class-Boundaries	*f*	More Than	Cumulative Freq.
60 - 61.75	5	More than 60	100
61.75 - 63.5	3	More than 61.75	95
63.5 - 65.25	15	More than 63.5	92
65.25 - 67	40	More than 65.25	77
67 - 68.75	17	More than 67	37
68.75 - 70.5	12	More than 68.75	20
70.5 - 72.25	7	More than 70.5	8
72.25 - 74	1	More than 72.25	1
-	-	More than 74	0



Median Through the Combine Less Than and More Than Type Cumulative Histogram



5.2 BOX-PLOT (WHISKERS PLOT)

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

IMPORTANT POINTS:

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately the middle 50 percent of the data fall inside the box. The "whiskers" extend from the ends of the box to the smallest and largest data values.

STEPS:

1. Arrange the data set into ascending order
2. Compute Median
3. Find 1st and 3rd Quartiles *i.e.*, Q1 and Q3
4. Check for the Minimum and Maximum value
5. Draw the Box-Plot (Whiskers Plot)

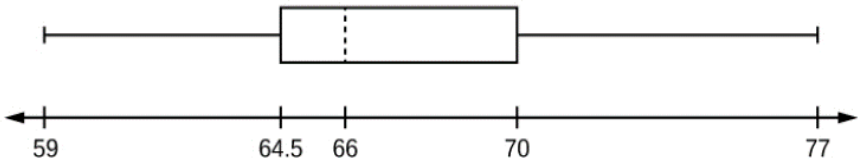
Example

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator intructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- *Q*1: First quartile = 64.5
- *Q*2: Second quartile or median= 66
- *Q*3: Third quartile = 70



5.3 DETECTING OUTLIERS

Sometimes a data set will have one or more observations with unusually large or unusually small values. These extreme values are called **outliers**.

An outlier may be a data value that has been incorrectly recorded. If so, it can be corrected before further analysis. An outlier may also be from an observation that was incorrectly included in the data set; if so, it can be removed. Finally, an outlier may be an unusual data value that has been recorded correctly and belongs in the data set. In such cases it should remain.

Two methods ate discussed below to detect the outlier from the data.

1- Standardized values (z-scores)

Standardized values (z-scores) can be used to identify outliers. Hence, in using z-scores to identify outliers, we recommend treating any data value with a z-score less than −3 or greater than +3 as an outlier. Such data values can then be reviewed for accuracy and to determine whether they belong in the data set.

$$Z = \frac{x - \mu}{\sigma}$$

Where; μ is the population mean; σ is the standard deviation of population

2- By Using the Quartiles:

Another approach to identifying outliers is based upon the values of the first and third quartiles (Q1 and Q3) and the interquartile range (IQR). Using this method, we first compute the following lower and upper limits:

$$\text{Lower Limit} = Q1 - 1.5(IQR)$$

$$\text{Upper Limit} = Q3 + 1.5(IQR)$$

An observation is classified as an outlier if its value is less than the lower limit or greater than the upper limit. **For Example**, we have the monthly starting salary data that shown, Q1 = 5857.5, Q3 = 6025, IQR = 167.5, and the lower and upper limits are

$$\text{Lower Limit} = Q1 - 1.5(IQR) = 5857.5 - 1.5(167.5) = 5606.25$$

$$\text{Upper Limit} = Q3 + 1.5(IQR) = 6025 + 1.5(167.5) = 6276.25$$

Looking at the data set, any value that is less than 5606.25 or greater than the 6276.25 value will be considered as the outlier of the data set.

PRACTICE QUESTIONS

Page 147

Questions

83, 84, 114, 115, 119

LECTURE 6: MEASURES OF THE CENTER OF THE DATA

6.1 THE MEDIAN

The median is exact central value of the order data set. Median value will divide the data set into two equal parts.

6.1.1 UNGROUP DATA (MEDIAN):

Suppose we have the following observations:

10, 15, 30, 7, 42, 79 and 83

CASE-I

Arrange the data set in a array *i.e.*, 07, 10, 15, 30, 42, 79, 83

i. Divide the total number of observations by 2 *i.e.*, $\frac{n}{2}$

ii. Not an integer: $\frac{n}{2} = \frac{7}{2} = 3.5$ then the median value will be

$$\frac{(n+1)}{2} th\ value = \frac{8}{2} = 4th\ value\ in\ the\ data$$

CASE-II

Arrange the data set in a array *i.e.*, 07, 08, 10, 15, 30, 42, 79, 83

i. Divide the total number of observations by 2 *i.e.*, $\frac{n}{2}$

ii. An integer: $\frac{n}{2} = \frac{8}{2} = 4$ then the median value will be

$$\frac{\left(\frac{n}{2}\right) th\ value + \left(\frac{n}{2}+1\right) th\ value}{2} = \frac{15+30}{2} = \frac{45}{2} = 22.5$$

6.1.2 GROUP DATA (MEDIAN):

The median is exact central value of the order data set. Median value will divide the data set into two equal parts. Suppose we have the following observations:

$$\text{Median} = \text{M} = \tilde{x} = l + \frac{h}{f}\left(\frac{n}{2} - C\right)$$

Where;

l = lower limit of the selected class; h = class interval; f = selected class's frequency

$n/2$ = computed value; C = previous class's cumulative frequency

Example

Monthly Wages (Rs)	*No. of Workers*
800–1,000	18
1,000–1,200	25
1,200–1,400	30
1,400–1,600	34
1,600–1,800	26
1,800–2,000	10
Total	**143**

Solution:

Step-wise Calculation

1- Find the cumulative frequency of the given data set.

2- Find the value of n/2 and find in which class this (n/2)th value will fall.

As n = 143 so n/2 = 71.5

Monthly Wages	*Frequency*	*Cumulative Frequency*
800–1,000	18	18
1,000–1,200	25	43
1,200–1,400	30	73
1,400–1,600	34	107
1,600–1,800	26	133
1,800–2,000	10	143

We can see that our value falls in the 3rd group. Now select the required values of the formula.

l = 1200; *h = 200;* *f = 30;* *C = 43;* *n/2 = 71.5 ≈ 72*

By replacing and simplifying the values by using the formula. The median value will be

$$\text{Median} = \tilde{x} = l + \frac{h}{f}\left(\frac{n}{2} - C\right)$$

$$\tilde{x} = 1200 + \frac{200}{30}(72 - 43)$$

$$\tilde{x} = 1200 + \frac{200}{30}(29) = 1393.3$$

Advantages of Median

1. Unlike arithmetic mean, median is not affected at all by extreme values as it is a positional average. As such, median is particularly very useful when a distribution happens to be skewed.

2. Another point that goes in favour of median is that it can be computed when a distribution has open-end classes.

3. Another merit of median is that when a distribution contains qualitative data, it is the only average that can be used. No other average is suitable in case of such a distribution.

6.2 THE MODE

The mode is the most frequent or repeated value in the data/order data set.

6.2.1 UNGROUP DATA (MODE):

1- Suppose we have the following observations:

10, 15, 30, 7, 42, 79 and 83

In this data set there is no such value(s) that repeats a greater number of times than others. So, in that case will say that mode does not exists.

2-Suppose we have the following observations:

10, 15, 30, 7, 42, 79, 30 and 83

In the following data set we can see that 30 values repeats twice in the data. So, the mode is 30.

6.2.2 GROUP DATA (MODE):

The mode is the most frequent or repeated value in the data/order data set.

Suppose we have the following observations:

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} h$$

Where;

f_m = Class with maximum frequency; f_1 = previous class frequency;

f_2 = next class frequency; h = class interval; l = lower class-limit of the selected class

Example

Monthly Wages (Rs)	*No. of Workers*
800–1,000	18
1,000–1,200	25
1,200–1,400	30
1,400–1,600	34
1,600–1,800	26
1,800–2,000	10
Total	**143**

Solution:

Select the class with maximum frequency.

Monthly Wages (Rs)	*No. of Workers*
800–1,000	18
1,000–1,200	25
1,200–1,400	30
1,400–1,600	34
1,600–1,800	26
1,800–2,000	10
Total	**143**

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} h$$

f_m= Class with maximum frequency = 34;

f_1= previous class frequency = 30;

f_2= next class frequency = 26;

h = class interval = 1400 – 1200 = 200;

l = lower class limit of the maximum frequency class = 1400

Now putting the values into the formula and solving the equation

$$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} h = 1400 + \frac{34 - 30}{(34 - 30) + (34 - 26)} * 200 = 1466.67$$

Mode = 1466.67 for the given group data.

PRACTICE QUESTIONS

Page 133 Questions 42-45,
Page 139 Questions 73, 91, 92

LECTURE 7: MEASURES OF THE DISPERSION

7.1 MEASURES OF THE DISPERSION

In the preceding topics, we have seen different types of means and learnt how they can be calculated in varying types of distributions. The means are just the measures of central tendency and do not indicate the extent of variability in a distribution.

The main theme of dispersion is variability, which provides us one more step in increasing our understanding of the pattern of the data.

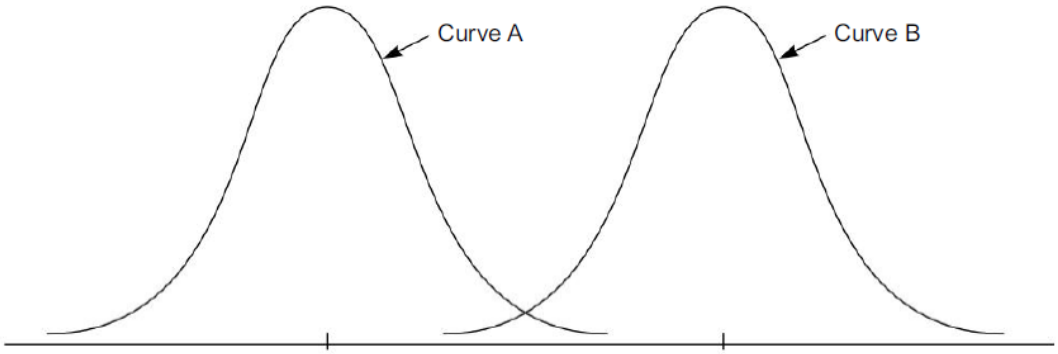


Fig. 7.1 **Curves A and B with Different Means and Same Variability**

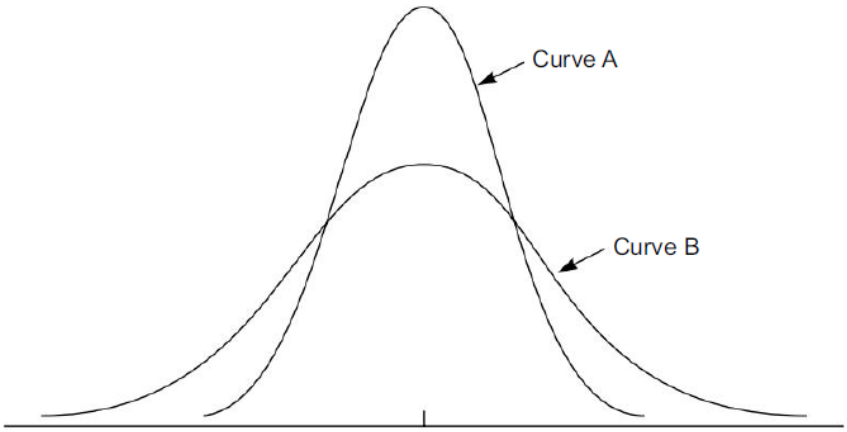


Fig. 7.2 **Curves A and B with Equal Means and Different Variability**

Some Example(s):

1. Suppose, of the two applicants, a firm is planning to appoint one worker. It may assign certain work to each of them. Though the average output of both the workers is the same, the output of the first worker is more stable as compared to the second worker, whose output shows considerable fluctuation. Therefore, the firm should appoint the first worker, as his output has less variability.

2. In the sphere of quality control, dispersion, i.e., variability, is more relevant than average measurement. A manufacturing company may first fix a standard for a prospective product and then ascertain as to the extent to which it has deviated from the previously-set standard. This would enable the company to take necessary steps to ensure that the output is in conformity with the laid down standard.

7.2 TYPES OF DISPERSION:

There are four major types of dispersion

1. Range
2. Inter Quartile Range or Quartile Deviation
3. Mean Deviation
4. The Standard Deviation or Variance

7.2.1 1- RANGE

The simplest measure of dispersion is the range, which is the difference between the maximum value and the minimum value of data.

Range = $X_m - X_o$

- *In a frequency distribution, range is calculated by taking the difference between the upper limit of the highest class and the lower limit of the lowest class.*

Example Find the range for the following frequency distribution:

Size of Item	*Frequency*
20–40	7
40–60	11
60–80	30
80–100	17
100–120	5
Total	**70**

Solution Here, the upper limit of the highest class is 120 and the lower limit of the lowest class is 20. Hence, the range is 120 – 20 = 100. Note that the range is not influenced by the frequencies. Symbolically, the range is calculated by the formula $L–S$, where L is the largest value and S is the smallest value in a distribution. The coefficient of range is calculated by the formula:

$$\frac{L-S}{L+S}$$

7.2.2 INTER QUARTILE RANGE OR QUARTILE DEVIATION

The interquartile range or the quartile deviation is a better measure of variation in a distribution than the range. Here, the middle 50 per cent of the distribution is used by avoiding the 25 per cent of the distribution at both the ends. In other words, the interquartile range denotes the difference between the third quartile and the first quartile.

$$\text{Interquartile Range (IQR)} = Q_3 - Q_1$$

Many times, the interquartile range is reduced in the form of semi-interquartile range or quartile deviation to know the dispersion of the central 50 percent items.

$$\text{Semi} - \text{Interquartile Range (SIQR) or Quartile Deviation (QD)} = \frac{Q_3 - Q_1}{2}$$

The term/rule/formula of interquartile range or the quartile deviation is an absolute measure of dispersion. It can be changed into a relative measure of dispersion as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

7.2.3 GROUP DATA (I.Q.R):

Example: Find the IQR's Absolute and relative measure for the following data set.

Class-interval	*Frequency*	*Cumulative frequency*
30–40	4	4
40–50	6	10
50–60	8	18
60–70	12	30
70–80	9	39
80–90	7	46
90–100	4	50

Q_1 = n/4 = 50/4 = 12.5th value

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) = 50 + \frac{10}{8}(12.5 - 10) = 53.125$$

$Q_2 = \frac{2*n}{4} = {}^{n}/_{2}$ = 50/2 = 25th value

$$Q_2 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) = 60 + \frac{10}{12}(25 - 18) = 65.83$$

Q_3 = 3n/4 = 3*50/4 = 37.5th value

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) = 70 + \frac{10}{9}(37.5 - 30) = 78.33$$

Inter-Quartile Range:

$$\text{IQR} = Q_3 - Q_1 = 78.33 - 53.125 = 25.205$$

Quartile Deviation (QD) or S.I.Q.R:

$$\text{QD} = \frac{Q_3 - Q_1}{2} = \frac{25.205}{2} = 12.6025$$

Co-efficient of Quartile Deviation:

$$\text{Coefficient of Q. D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{23.205}{129.455} = 0.179$$

7.2.4 MEAN DEVIATION:

The mean deviation is also known as the average deviation. As the name implies, it is the average of absolute amounts by which the individual items deviate from the mean. While computing the mean deviation, we ignore positive and negative signs. Symbolically,

$$\text{Mean Deviation} = \text{MD} = \frac{\sum|x - \bar{x}|}{n}$$

Where;

$|x - \bar{x}|$ = modulus of deviations from the mean

n = the total number of observations

UNGROUP DATA (M.D.)

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
20	2	2
15	-3	3
19	1	1
24	6	6
16	-2	2
14	-4	4
$\sum x$ = 108	**Total**	18

Calculations:

$$\bar{x} = \frac{\sum x}{n} = 18$$

$$\text{MD} = \frac{\sum|x - \bar{x}|}{n} = \frac{18}{6} = 3$$

GROUP DATA (MEAN DEVIATION):

Size of Item	*Frequency*
2–4	20
4–6	40
6–8	30
8–10	10

Solution We set up the worksheet for calculating the mean deviation.

Size of Item	*Mid-points (m)*	*Frequency (f)*	*mf*	*d from* $\bar{x}$	$f\|d\|$
2–4	3	20	60	– 2.6	52
4–6	5	40	200	– 0.6	24
6–8	7	30	210	1.4	42
8–10	9	10	90	3.4	34
	Total	**100**	**560**		**152**

$$\bar{x} = \frac{\Sigma fm}{n} = \frac{560}{100} = 5.6$$

$$\text{MD } (\bar{x}) = \frac{\Sigma f\,|\,d\,|}{n} = \frac{152}{100} = 1.52$$

7.2.5 STANDARD DEVIATION OR VARIANCE: (UNGROUP DATA)

The mean deviation is also known as the average deviation. As the name implies, it is the average of absolute amounts by which the individual items deviate from the mean. While computing the mean deviation, we ignore positive and negative signs. Symbolically,

For Population Data set

Method	Notation	Formula I	Formula II
Variance	σ^2	$= \frac{\sum(x_i - \mu)^2}{N}$	$= \frac{\sum X_i^2}{N} - \left(\frac{\sum X_i}{N}\right)^2$
Standard Deviation	σ	$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$	$= \sqrt{\frac{\sum X_i^2}{N} - \left(\frac{\sum X_i}{N}\right)^2}$

i = 1, 2, 3, …, n/N

For Sample Data set

Method	Notation	Formula I	Formula II
Variance	s^2	$= \frac{\sum(x_i - \bar{x})^2}{n}$	$= \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2$
Standard Deviation	s	$= \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$	$= \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2}$

Where; i = 1, 2, 3, …, n/N

$(x - \bar{x})^2$ = Squared deviation of an item from the mean, n = Total number of sample observations

UNGROUP DATA (S.D & VARIANCE)

x	$x - \bar{x}$	$(x - \bar{x})^2$	x^2
20	2	4	400
15	-3	9	225
19	1	1	361
24	6	12	576
16	-2	4	256
14	-4	16	196
$\sum x = 108$	**Total**	70	2014

Formula – I:

$$\text{Variance} = \frac{\sum(x - \bar{x})^2}{n} = \frac{70}{6} = 11.67$$

$$\text{SD} = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = \sqrt{11.67} = 3.42\ points$$

Formula – II:

$$\text{SD} = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{2014}{6} - \left(\frac{108}{6}\right)^2} = \sqrt{11.67} = 3.42\ points$$

GROUP DATA (S.D & VARIANCE)

For Sample Data Set

Method	Notation	Formula I	Formula II
Variance	s^2	$= \frac{\sum f(x_i - \bar{x})^2}{n}$	$= \frac{\sum f_i X_i^2}{n} - \left(\frac{\sum f_i X_i}{n}\right)^2$
Standard Deviation	s	$= \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{n}}$	$= \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum f_i X_i}{n}\right)^2}$

i = 1, 2, 3, …, n/N

FORMULA 1:

C-L	Frequency	Mid-Point (x)	fx	$(x\text{-mean})^2$	$f(x\text{-mean})^2$
2-4	20	3	60	6.76	135.2
4-6	40	5	200	0.36	14.4
6-8	30	7	210	1.96	58.8
8-10	10	9	90	11.56	115.6
Total	100	-	560	-	324

$$\text{Mean } = \frac{\sum fx}{n} = \frac{560}{100} = 5.6$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum f(x-\bar{x})^2}{n}} = \sqrt{\frac{324}{100}} = \sqrt{3.24} = 1.8$$

FORMULA 2:

C-L	Frequency	Mid-Point (x)	x^2	fx	fx^2
2-4	20	3	9	60	180
4-6	40	5	25	200	1000
6-8	30	7	49	210	1470
8-10	10	9	81	90	810
Total	100			560	3460

$$\text{Standard Deviation} = \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2} = \sqrt{\frac{3460}{100} - \left(\frac{560}{100}\right)^2}$$

$$= \sqrt{34.6 - (5.6)2} = \sqrt{3.24} = 1.8$$

7.2.6 COEFFICIENT OF VARIATION:

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

$$\text{CV } = \frac{SD}{mean} x100$$

$$\text{CV } = \frac{\text{SD}}{\text{mean}} x100 = \frac{1.8}{5.6} x100 = 32.14\%$$

PRACTICE QUESTIONS

Page 133
Questions
42-45,
Page 139
Q. 73, 91, 92

LECTURE 8: SKEWNESS

8.1 COMBINE MEAN:

$$Combine\ mean = \bar{x}_c = \frac{n1\bar{x}_1 + n2\bar{x}_2 + n_3\bar{x}_3 + \ \ldots + n_k\bar{x}_k}{n_1 + n_2 + n_3 + \ \ldots + n_k}$$

Example:

The mean height of 25 male workers in a factory is 61 inches and the mean height of 35 female workers in the same factory is 58 inches. Find the combined mean height of 60 workers in the factory.

Solution:

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{(25 * 61) + (35 * 58)}{25 + 35} = \frac{1525 + 2030}{60} = 59.25\ inches$$

Example:

The mean annual salary of employees of a company is Rs 30000. The mean annual salaries of male and female employees are Rs 35000 and Rs 23000, respectively. Find out the percentage of male and female employees working in the company.

Solution:

Given: $\bar{x}_1 = \text{Rs } 35000$, $\bar{x}_2 = \text{Rs } 23000$ and $\bar{x}_{12} = \text{Rs } 30000$

Now, $$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

or $$(\bar{x}_{12})(n_1 + n_2) = n_1\bar{x}_1 + n_2\bar{x}_2$$

or $$30000(n_1 + n_2) = n_1\,(35000) + n_2\,(23000)$$

or $$30000\,n_1 + 30000\,n_2 = 35000n_1 + 23000\,n_2$$

or $$30000\,n_1 - 35000\,n_1 = 23000\,n_2 - 30000\,n_2$$

or $$-5000\,n_1 = -7000\,n_2$$

or $$n_1/n_2 = -5000/-7000$$

$$= 5/7$$

Hence, the percentage of male employees is

$$5/7 \times 100 = 500/7 = 71.43$$

and the percentage of female employees is

$$1 - 5/7 = 2/7$$

$$2/7 \times 100 = 200/7 = 28.57$$

8.2 COMBINE VARIANCE:

If two groups contain n_1 and n_2 observations with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2 respectively, then the standard deviation (σ) of the combined group is given by

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

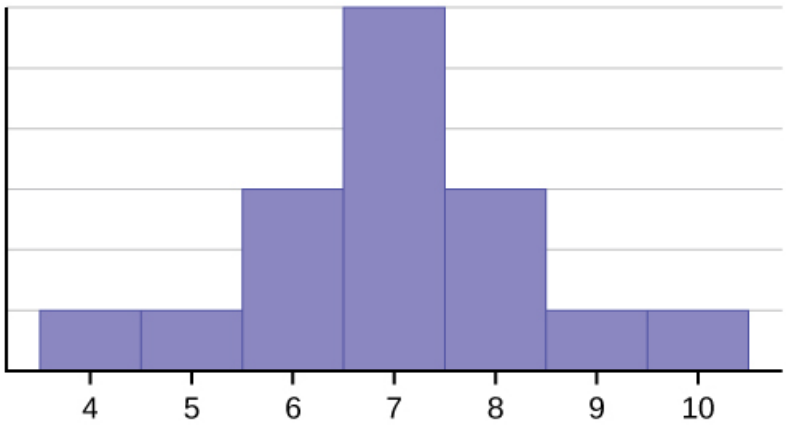
where σ_{12} *is the combined standard deviation of the two groups,* $d_1 = \bar{x}_{12} - \bar{x}_1$, $d_2 = \bar{x}_{12} - \bar{x}_2$ *and*

$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

8.3 SKEWNESS

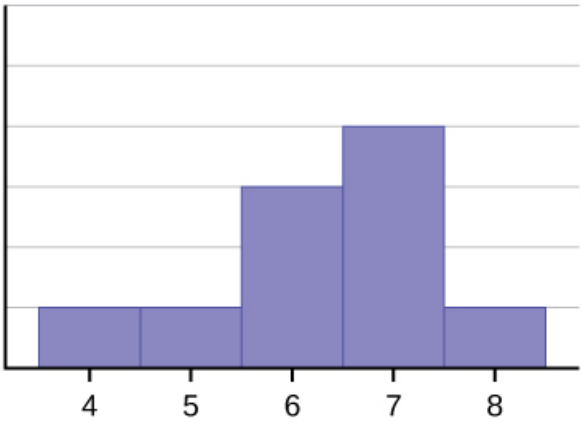
Skewness shows the tendency of the data to the measure of central tendency. It could either be symmetrical distribution (equally distributed on right and left side) or Asymmetrical distribution *i.e.*, positive skewed (right-tailed) or negative skewed (left-tailed).

The histogram displays a symmetrical distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. In a perfectly symmetrical distribution, the mean and the median are the same. This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.



A perfect skewed distribution or Symmetrical Distribution.

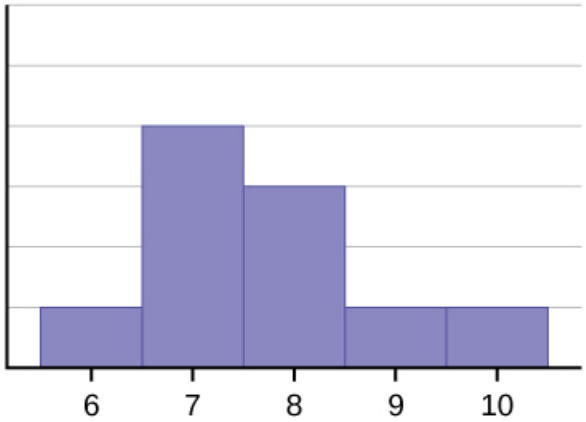
Data is divided equally to the right and left tail.



A negative skewed distribution or Asymmetrical Distribution.

Data is dispersed mostly left side. It is also called left-tailed distribution.

The mean is 6.3, the median is 6.5, and the mode is seven. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so. If the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode.



A positive skewed distribution or Asymmetrical Distribution.

Data is dispersed mostly right side. It is also called right-tailed distribution.

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

$$Sk = \frac{Mean - Mode}{S.d} = \frac{3(Mean - Median)}{S.d}$$

Interpretation of the Pearson Skewness Method:

- The direction of skewness is given by the sign.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.
- A value of zero means no skewness at all.
- A large negative value means the distribution is negatively skewed.
- A large positive value means the distribution is positively skewed.

PRACTICE QUESTIONS

Page 137
Questions
110-116

LECTURE 9: PROBABILITY

9.1 PROBABILITY

1. A quantitative measure of an uncertainty is called probability.
2. A measure of degree of belief in a particular statement or problem.
3. Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity.

$P(A) = n(A) / n(S)$

Where, n(A) = favorable outcomes of an event; n(S) = Total No. of outcomes of a random experiment

9.2 TERMINOLOGY

9.2.1 EXPERIMENT:

An experiment is a planned operation carried out under controlled conditions with predetermined results.

9.2.2 RANDOM EXPERIMENT:

An experiment is a planned operation carried out under controlled conditions. If the results are not predetermined then it is called random experiment.

9.2.3 OUTCOME:

A result of an experiment is called an outcome. These outcomes are also called sample points.

9.2.4 TRAIL:

Single performance of an experiment is called a trail.

9.2.5 SAMPLE SPACE:

A set consisting of all possible outcomes that can result from a random experiment (real or conceptual), is defined to be sample space. Denoted by *S*. *n(s)* means total number of outcomes in the sample space.

Example 1:

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is {HH, TH, HT, TT} where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition {HT, TH}, so P(A) = 2/4 = 0.5.

Example 2:

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event E = rolling a number that is at least five. There are two outcomes {5, 6}. P(E) = 2/6 = 1/3.

9.3 AXIOMS OF PROBABILITY:

1. P(S) = 1
2. P(A) = n(A)/n(S)
3. $P(\bar{A}) = 1 - P(A)$
4. Probability always lies between zero and 1.

9.4 INDEPENDENT EVENTS:

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. Otherwise, events are called *Dependent.*

9.5 SAMPLING:

Selection of a unit from the population is called sampling. Sampling may be done with replacement or without replacement.

9.5.1 WITH REPLACEMENT:

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.

9.5.2 WITHOUT REPLACEMENT:

When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

For instance:

Suppose you have a fair, well-shuffled deck of 52 cards. You have to pick 3 cards from the deck, let's say

a. Sampling with replacement:

The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:

The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

9.6 MUTUALLY EXCLUSIVE EVENTS:

A and B are mutually exclusive events if they cannot occur at the same time. This means that A and B do not share any outcomes and P(A AND B) = P(A∩B) = 0.

For example, suppose the sample space S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let A = {1, 2, 3, 4, 5}, B = {4, 5, 6, 7, 8}, and C = {7, 9}. A AND B = {4, 5}. P(A AND B) = 2/10 and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so

P(A AND C) = 0. Therefore, A and C are mutually exclusive.

EXAMPLE:

Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose P(G) = 0.6, P(H) = 0.5, and P(G AND H) = 0.3. Are G and H independent?

SOLUTION:

If G and H are independent, then you must show ONE of the following:

• P(G|H) = P(G)

• P(H|G) = P(H)

• P(G AND H) = P(G)P(H)

a. Show that P(G|H) = P(G).

P(G|H) = P(G AND H)

P(H) = 0.3/0.5 = 0.6 = P(G)

b. Show P(G AND H) = P(G)P(H).

P(G)P(H) = (0.6)(0.5) = 0.3 = P(G AND H)

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that P(H|G) = P(H) to show that G and H are independent events.

9.7 CONDITIONAL PROBABILITY:

C. Probability is the likelihood that an event will occur given that another event has already occurred.

The conditional probability of A given B is written P(A|B). P(A|B) is the probability that event A will occur given that the event B has already occurred. A conditional reduces the sample space. We calculate the probability of A from the reduced sample space B. The formula to calculate P(A|B) is P(A|B) = $\frac{P(A \text{ and } B)}{P(B)}$ where P(B) is greater than zero.

EXAMPLE: Suppose we toss one fair, six-sided die. The sample space S = {1, 2, 3, 4, 5, 6}. Let A = face is 2 or 3 and B = face is even (2, 4, 6). To calculate P(A|B), we count the number of outcomes 2 or 3 in the sample space B = {2, 4, 6}. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{n(A \cap B)/n(S)}{n(B)/n(S)} = \frac{1/6}{3/6} = \frac{1}{3}$$

EXAMPLE: Two coins are tossed. What is the conditional probability that two head occurs given that there is at least one head?

SOLUTION:

The Sample space for the experiment is

S = {HH, HT, TH, TT}

Let A represents the event that two head appears, and B is the event that there is at least one head. Then P(A|B) =?

Since A = {HH}; B = {HH, HT, TH}; A∩B = {HH}

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{n(A \cap B)/n(S)}{n(B)/n(S)} = \frac{1/4}{3/4} = \frac{1}{3}$$

PRACTICE QUESTIONS

Page 365 – 378
Questions
1 – 30
74 – 80

LECTURE 10: BASIC RULES

10.1 TWO BASIC RULES OF PROBABILITIES:

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

10.1.1 THE ADDITION RULE:

If A and B are defined on a sample space, then: P(A OR B) = P(A U B) = P(A) + P(B) - P(A ∩ B).

If A and B are mutually exclusive, then P(A ∩ B) = 0. Then P(A U B) = P(A) + P(B) - P(A ∩ B) becomes P(A U B) = P(A) + P(B).

Example:

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska. Klaus can only afford one vacation. While the probability that he chooses A is P(A) = 0.6 and the probability that he chooses B is P(B) = 0.35.

Solution:

- P(A AND B) = 0 because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is P(A OR B) = P(A) + P(B) = 0.6 + 0.35 = 0.95. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

10.1.2 THE MULTIPLICATION RULE:

If A and B are two events defined on a sample space, then: P(A∩B) = P(B)P(A|B).

This rule may also be written as: P(A|B) = P(A ∩ B)/P(B)

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are independent, then P(A|B) = P(A). Then P(A∩B) = P(A|B)P(B) becomes P(A ∩ B) = P(A)P(B).

EXAMPLE:

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The **remainder** are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?

P(A) = n(A)/n(S) = 28/150

b. What is the probability that the member practices four times a week?

P(B) = n(B)/n(S) = 80/150

c. What is the probability that the member is an advanced swimmer and practices four times a week?

P(Advanced and Four hour) = n(A∩B)/n(S) = 40/150

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

P(advanced AND intermediate) = 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

EXAMPLE:

Tahir attends Modesto JC in Modesto, CA. The probability that Tahir enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, M|S = math given speech

a. What is the probability that Tahir enrolls in math and speech?

Find P(M AND S) = P(M|S)P(S).

b. What is the probability that Tahir enrolls in math or speech classes?

Find P(M OR S) = P(M) + P(S) - P(M AND S).

c. Are M and S independent? Is P(M|S) = P(M)?

d. Are M and S mutually exclusive? Is P(M AND S) = 0?

SOLUTION:

a. P(M AND S) = P(M|S)P(S) = 0.25*0.65 = 0.1625

b. P(M OR S) = P(M) + P(S) - P(M AND S) = 0.2 + 0.65 – 0.1625 = 0.6875

c. No

d. No

10.2 CONTINGENCY TABLES:

A contingency table is a tabular representation of categorical data. A contingency table usually shows frequencies for particular combinations of values of two discrete random variable s X and Y. Each cell in the table represents a mutually exclusive combination of X-Y values.

A contingency table provides a way of portraying data that can facilitate calculating probabilities.

EXAMPLE:

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

a. Find *P*(Driver is a cell phone user).
b. Find *P*(driver had no violation in the last year).
c. Find *P*(Driver had no violation in the last year AND was a cell phone user).
d. Find *P*(Driver is a cell phone user OR driver had no violation in the last year).
e. Find *P*(Driver is a cell phone user GIVEN driver had a violation in the last year).
f. Find *P*(Driver had no violation last year GIVEN driver was not a cell phone user)

SOLUTION:

a. No. of cell phone users / total No. in study = 305/755

b. No. that had no violation/total No. of study = 685/755

c.280/755

d. $\left(\frac{308}{755}+\frac{685}{755}\right)-\frac{280}{755}=\frac{710}{755}$

e.25/70 (The sample space is reduced to the number of drivers who had a violation).

f. 405/450 (The sample space is reduced to the number of drivers who were not cell phone users).

PRACTICE QUESTIONS

Page 365 – 378
Questions
1 – 30
74 – 80

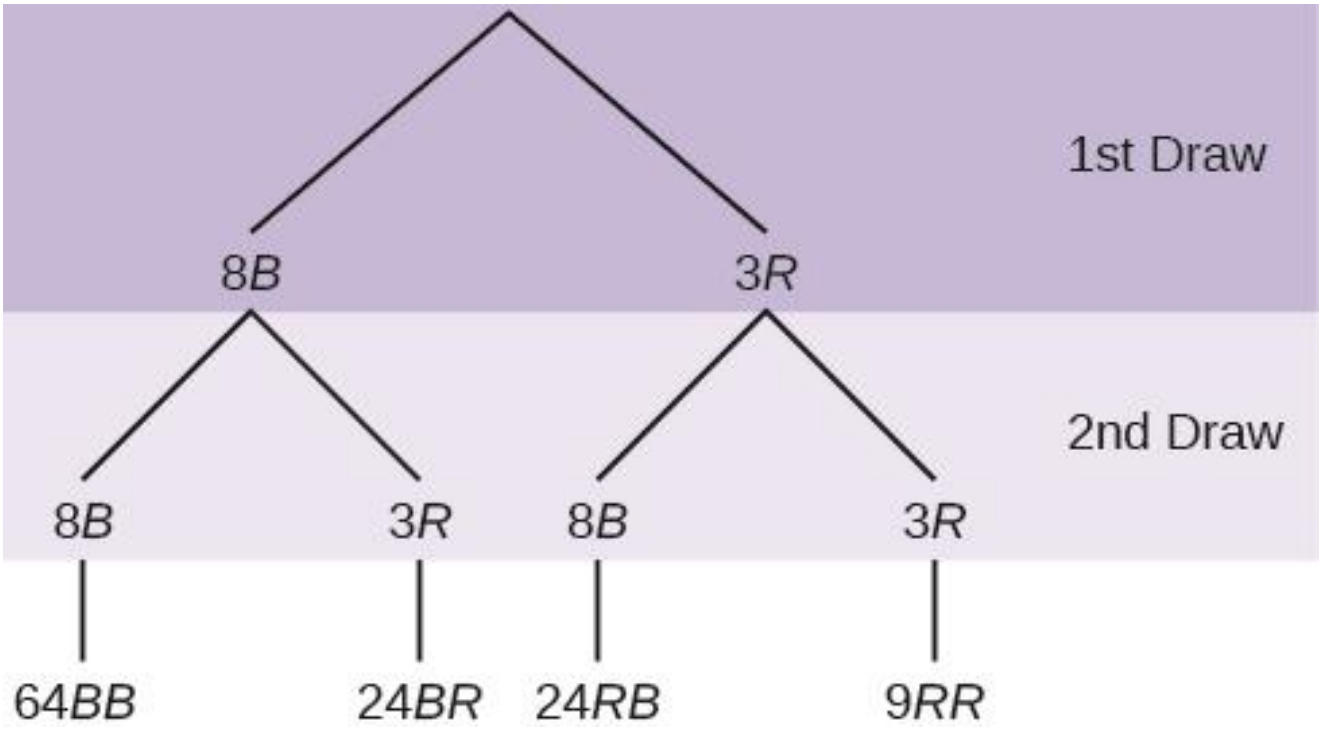
LECTURE 11: PROBABILITY TOPICS

11.1 TREE AND VENN DIAGRAMS

A tree diagram is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labelled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve.

EXAMPLE:

In an urn, there are 11 balls. Three balls are red (*R*) and eight balls are blue (*B*). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.



Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, andB8. Then the nine RR outcomes can be written as:

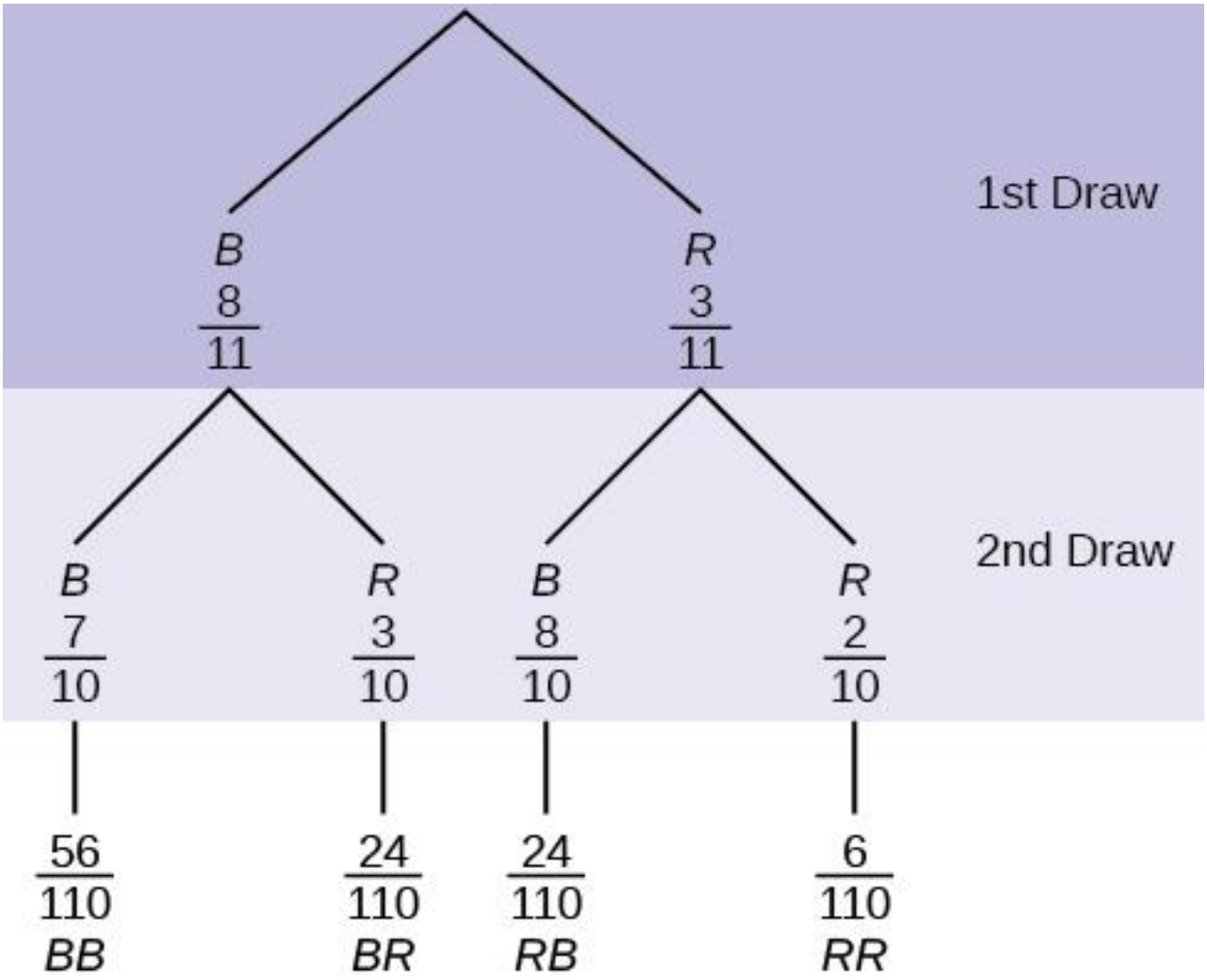
R1R1 R1R2 R1R3 R2R1 R2R2 R2R3 R3R1 R3R2 R3R3

The other outcomes are similar.

Solution: There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the sample space.

EXAMPLE 2:

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. "Without replacement" means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, (311)*(210) = 6110.



Total = (56+24+24+6)/110 = 110/110 = 1

PRACTICE QUESTIONS

Page 365 – 378
Questions
1 – 30
74 – 80

Version 1

Math, FOIT

Part -II

Mehwish Omer

Lecturer Statistics
Department of Mathematics
Faculty of Science and Technology
University of Central Punjab, Lahore

DISCRETE PROBABILITY DISTRIBUTION.

Objectives:

- **Define the terms probability distribution and random variable.**
- **Distinguish between discrete and continuous probability distributions.**
- **Calculate the mean, variance, and standard deviation of a discrete probability distribution.**

Probability Distribution:

A listing of all the outcomes of an experiment and the probability associated with each outcome.

Probability Distribution Function (PDF):

A probability distribution has two characteristics.

1. Each probability is between zero and one, inclusive
2. The sum of the probabilities is one**.**

Example:

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother `after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight.

SOLUTION:

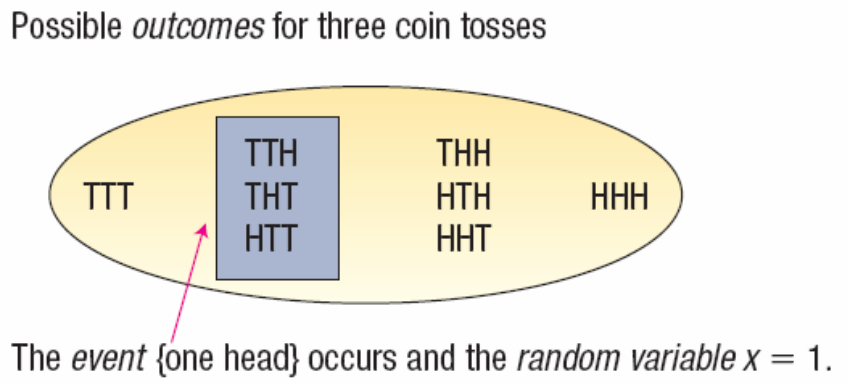
For this example, x = 0, 1, 2, 3, 4, P(x) = probability that X takes on a value X takes on the values 0, 1, 2, 3, 4, 5

x	*P(x)*
0	$P(x=0)=\frac{2}{50}$
1	$P(x=1)=\frac{11}{50}$
2	$P(x=2)=\frac{23}{50}$
3	$P(x=3)=\frac{9}{50}$
4	$P(x=4)=\frac{4}{50}$
5	$P(x=5)=\frac{1}{50}$

a. Each P(x) is between zero and one, inclusive
b. The sum of the probability is one.

RANDOM VARIABLE:

A quantity resulting from an experiment that, by chance, can assume different values.



TYPES OF RANDOM VARIABLES.

DISCRETE RANDOM VARIABLE:

A random variable that can assume only certain clearly separated values. It is usually the result of counting something.
e.g.: no. of students, no of cars in car parking.

CONTINUOUS RANDOM VARIABLE:

A variable that can assume an infinite number of values within a given range. It is usually the result of some type of measurement.
e.g.: Heights of students, C.G.P. A of students.

DISCRETE RANDOM VARIABLE

A random variable can assume only certain clearly separated values. It is usually the result of counting something.

EXAMPLES

1. The number of students in a class.
2. The number of children in a family.
3. The number of cars entering a carwash in a hour.
4. The number of passes completed by a football team in a match.
5. The number of heads that appear in a coin toss, when it is tossed 3 times.

MEAN

- The mean is a typical value used to represent the central location of a probability distribution.
- The mean of a probability distribution is also referred to as its **expected value.**

MEAN OF A PROBABILITY DISTRIBUTION $\mu = \Sigma[xP(x)]$

VARIANCE AND STANDARD DEVIATION

- Measures the amount of spread in a distribution
- The computational steps are:
  1. Subtract the mean from each value i.e. find(X-μ).
  2. Square this difference i.e. find $(X-\mu)^2$.
  3. Multiply each squared difference by its probability.
  4. Sum the resulting products to arrive at the variance.
  5. The standard deviation is found by taking the positive square root of the variance

$$\sigma^2 = \Sigma[(x - \mu)^2 P(x)]$$

Example:

Suppose you play a game with a biased coin. You play each game by tossing the coin once. P(heads) = 2/ 3 and P(tails) = 1 /3. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

a. Define a random variable X.

b. Complete the following expected value table

Solution:

X = amount of profit.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

	x	*P(x)*	*xP(x)*
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

The expected value **μ = –2/3**. You lose, on average, about 67 cents each time you play the gam

Mean, Variance, and Standard Deviation of a Probability Distribution

John Ragsdale sells new cars for Pelican Ford. John usually sells the largest number of cars on Saturday. He has developed the following probability distribution for the number of cars he expects to sell on a particular Saturday?

$$\mu = \Sigma[xP(x)]$$
$$= 0(.10) + 1(.20) + 2(.30) + 3(.30) + 4(.10)$$
$$= 2.1$$

Number of Cars Sold, x	Probability, $P(x)$	$x \cdot P(x)$
0	.10	0.00
1	.20	0.20
2	.30	0.60
3	.30	0.90
4	.10	0.40
Total	1.00	$\mu = 2.10$

VARIANCE & STANDARD_DEVIATION:

$$\sigma^2 = \Sigma[(x - \mu)^2 P(x)]$$

Number of Cars Sold, x	Probability, $P(x)$	$(x - \mu)$	$(x - \mu)^2$	$(x - \mu)^2 P(x)$
0	.10	0 − 2.1	4.41	0.441
1	.20	1 − 2.1	1.21	0.242
2	.30	2 − 2.1	0.01	0.003
3	.30	3 − 2.1	0.81	0.243
4	.10	4 − 2.1	3.61	0.361
				$\sigma^2 = 1.290$

We also calculate it with the help of other formulae:

Variance Formula:

$$\sigma^2{}_x = \Sigma[x^2 * P(x)] - \mu_x^2$$

Example:

Calculate variance of the discrete probability distribution.

X	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
0	0.2	$0(0.2) = 0$	$0^2 = 0$	$0(0.2) = 0$
1	0.3	$1(0.3) = 0.3$	$1^2 = 1$	$1(0.3) = 0.3$
2	0.2	$2(0.2) = 0.4$	$2^2 = 4$	$4(0.2) = 0.8$
3	0.2	$3(0.2) = 0.6$	$3^2 = 9$	$9(0.2) = 1.8$
4	0.1	$4(0.1) = 0.4$	$4^2 = 16$	$16(0.1) = 1.6$
		$\Sigma[x \cdot P(x)]=1.7$		$\Sigma[x^2 \cdot P(x)]=4.5$

$$\mu = \sum [x \cdot P(x)] = 1.7$$

$$\sigma^2 = \sum [x^2 \cdot P(x)] - \mu^2$$

$$= 4.5 - 1.7^2$$

$$= \boxed{1.61} \longleftarrow \text{Variance}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.61} \approx 1.27$$

Standard Deviation

PRACTICE QUESTIONS:

FOR FOIT: Introductory Statistics from OpenStax Book

Page No: (243-253)

Questions No: (1-9,18-27)

BINOMIAL DISTRIBUTION

OBJECTIVES:

Describe the characteristics and compute probabilities using the binomial probability distribution.

Compute mean, variance and standard deviation.

Binomial distribution:

The distribution is derived from a process known as the Bernoulli trial, named in honour of the Swiss mathematician James Bernoulli (1654–1705), who made significant contributions in the field of probability, including, in particular, the binomial distribution.

$$\boldsymbol{P(x) = c_x^n p^x q^{n-x}}$$

1. There are only two possible outcomes on a particular trial of an experiment.
2. The outcomes are mutually exclusive.
3. The random variable is the result of counts.
4. Each trial is *independent* of any other trial.

Binomial Probability Experiment:

1. An outcome on each trial of an experiment is classified into one of two mutually exclusive categories—a success p, or a failure(1-p), q.

2. The random variable counts the number of successes in a fixed number of trials.
3. The probability of success and failure stays the same for each trial.
4. The trials are independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

The Binomial Parameters:

The binomial distribution has two parameters, n and p. They are parameters in the sense that they are sufficient to specify a binomial distribution. The binomial distribution is really a family of distributions with each possible value of n and p designating a different member of the family.

Binomial Distribution examples in Real Life:

1. A real-life example of the binomial distribution is the performance of students in a given test. Binomial distribution discerns the number of students who passed or failed in the test. Notably, the pass implies success and fail implies failure.
2. Another real-life example of Binomial Distribution is the introduction of a new vaccine that can be used to cure a disease. The vaccine curing the disease is the trial success and if it does not cure the disease it is considered as a failure.
3. Another example of Binomial Distribution is winning a bet. In the case of betting, the possibility of winning is the success whereas not winning implies failure.

.

Binomial Distribution: Mean and Variance:

The mean, μ, and variance, σ^2, for the binomial probability distribution are

$\mu = np$,

$\sigma^2 = npq$ and standard deviation is then $\sigma = \sqrt{npq}$.

EXAMPLE:

A survey found that one out of five Americans say he or she has visited a production company in any given month. If 10 people are selected at random, find the probability that exactly 3 will have visited a production company last month?

SOLUTION:

In this case, $n = 10, X = 3, p = \frac{1}{5}$, and $q = \frac{4}{5}$. Hence,

$$P(3) = \frac{10!}{(10-3)!3!}\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^7 = 0.201$$

EXAMPLE: According to a Gallup poll, 30% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 5 who prefer saving to spending. If 5 adults are selected at random, find the probability that at least 3 of them Will prefer saving to spending?

Solution:

$$P(3) = \frac{5!}{(5-3)!3!}(0.3)^3(0.7)^2 = 0.132$$

$$P(4) = \frac{5!}{(5-4)!4!}(0.3)^4(0.7)^1 = 0.028$$

$$P(5) = \frac{5!}{(5-5)!5!}(0.3)^5(0.7)^0 = 0.002$$

Probability that at least 3 prefer saving to spending =0.132+0.028+0.002=0.162

Example: Let's say that 80% of all business start-ups in the IT industry report that they generate a profit in their first year. If a sample of 10 new IT business start-ups is selected, find the probability that

(a) exactly seven will generate a profit in their first year?

Solution:
(a)The probability of seven IT start-ups to generate a profit in their first year is:

n = 10, p=0.80, q=0.20, x=7

$$P(x = 7) = \frac{10!}{7!(10-7)!}0.80^7(1-0.80)^{10-7}$$

=0.2013

Example:

A die is rolled 480 times. Find the mean, variance, and standard deviation of the number of 3s that will be rolled?

Solution:

This is a binomial experiment since getting a 3 is a success and not getting a 3 is considered a failure.

Hence $n = 480$, $p = \frac{1}{6}$, and $q = \frac{5}{6}$.

$$\mu = n \cdot p = 480 \cdot \tfrac{1}{6} = 80$$
$$\sigma^2 = n \cdot p \cdot q = 480 \cdot \tfrac{1}{6} \cdot \tfrac{5}{6} = 66.67$$
$$\sigma = \sqrt{66.67} = 8.16$$

PRACTICE QUESTIONS

For FOIT:

Introductory Statistics from OpenStax Book

Page no :(250-263)

Question no (88-100,104-110).

GEOMETRIC DISTRIBUTION

A discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success.

Notation: X ~ G(p).

The probability density function of geometric distribution is:

$$p(x) = (1-p)^{x-1}p$$

X = the number of independent trials until the first success X takes on the values.

x = 1,2.3….

p = the probability of success for any trial q = the probability of a failure for any trial

p + q = 1.

In such a sequence of trials, the geometric distribution is useful to model the number of failures before the first success since the experiment can have an indefinite number of trials until success, unlike the binomial distribution which has a set number of trials.

Geometric Experiment:

A statistical experiment that has the following properties:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p, of success and the probability, q, of a failure do not change from trial to trial.

The geometric distribution is widely used in several real-life scenarios. The geometric distribution is used to do a cost-benefit analysis to estimate the financial benefits of making a certain decision.

Difference between binomial and geometric distribution:

- There is NOT a fixed number of trials in geometric distribution but in binomial, there is a fixed number of trials
- Binomial random variables start with 0 while geometric random variables start with 1.

$$\text{Mean } \mu = \frac{1}{p}$$

$$\text{Variance } \sigma^2 = \frac{1-p}{p^2}$$

Example: What is the probability that the first son is born is at most the fourth child?

Solution:

P(x=4) $= p(1-p)^{x-1}$

$= 0.5(1-0.5)^{4-1}.$

=0.625I

Example: The probability of a successful optical alignment in the assembly of an optical data storage product is 0.8. Assume the trials are independent. Compute the probability that the first successful alignment

a. requires exactly four trials,

b. requires at least three trials,

c.requires at most three trials

Solution: Let X denote the number of trials required for first successful optical alignment. Thus the random variable X take values X=1,2,3,…

Given that the probability of successful optical alignment in the assembly of an optical data storage product is p=0.8. The trials are independent.

Thus, random variable X follows a geometric distribution with probability mass function:

a. The probability that the first successful alignment requires exactly 4 trials is:

$$\begin{aligned} P(X=4) &= 0.8(0.2)^{4-1} \\ &= 0.8(0.008) \\ &= 0.0064. \end{aligned}$$

.

b. The probability that the first successful alignment requires at least 3 trials is:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - \sum_{x=1}^{2} P(X=x) \\ &= 1 - \big(P(X=1) + P(X=2)\big) \\ &= 1 - (0.8 + 0.16) \\ &= 1 - 0.96 \\ &= 0.04. \end{aligned}$$

c. The probability that the first successful alignment requires at most 3 trials is:

$$\begin{aligned} P(X \leq 3) &= \sum_{x=1}^{3} P(X=x) \\ &= P(X=1) + P(X=2) + P(X=3) \\ &= 0.8 + 0.16 + 0.032 \\ &= 0.992. \end{aligned}$$

Example: The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Let X= the number of people you ask until one says he or she has pancreatic cancer. Then X is a discrete random variable with a geometric distribution: X ~ G (1/78)

(a)What is the probability of that you ask ten people before one says he or she has pancreatic cancer?

Solution:

$$P(x=10) = p(1-p)^{x-1}$$

$$=0.0128(1-0.0128)^{10-1}.$$

$$=0.0114$$

(b). Find the (i) mean and (ii) standard deviation of X

$$\text{Mean} = \mu = \frac{1}{p} = \frac{1}{0.0128} = 78$$

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.0128}{0.0128^2}} \approx 77.6234$$

PRACTICE QUESTIONS:

FOR FOIT: Introductory Statistics from OpenStax Book

Page No: (259-263)

Questions No: (104_109)

THE NORMAL DISTRIBUTION.

Objectives:

Recognize the normal probability distribution and apply it appropriately.

Recognize the standard normal probability distribution and apply it appropriately.

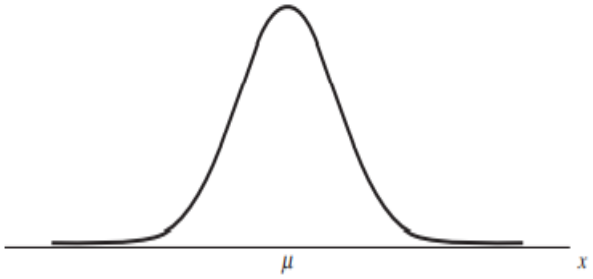
THE NORMAL DISTRIBUTION APPLICATIONS:

Although its importance in the field of statistics is indisputable, one should realize that normal distribution is not a law that is adhered to by all measurable characteristics in nature. A continuous distribution is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. We use the normal distribution to help determine the grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

Characteristics of the Normal Distribution:

The following are some important characteristics of the normal distribution.

1. It is symmetrical about its mean, As is shown in Figure, the curve on either side of is a mirror image of the other side



2. The mean, the median, and the mode are all equal.

3. The total area under the curve above the x-axis is one.

PROBABILITY DENSITY FUNCTION:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

Z-Scores:

If X is a normally distributed random variable and X ~ N (μ, σ), then the z-score is

$$\mathbf{z = x - \mu / \sigma}$$

The z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ

. A z-score for an individual value can be interpreted as follows:

- **Positive z-score:** The individual value is greater than the mean.
- **Negative z-score:** The individual value is less than the mean.
- **z-score of 0:** The individual value is equal to the mean.

The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation z = x − μ/ σ produces the distribution Z ~ N (0, 1). The value x in the given equation comes from a normal distribution with mean μ and standard deviation σ.

Example:

Suppose the scores for a certain exam are normally distributed with a mean of 80 and a standard deviation is 4. Find the z-score for an exam score of 87.

We can use the following steps to calculate the z-score:

- The mean is μ = 80
- The standard deviation is σ = 4
- The individual value we're interested in is X = 87
- Thus, z = (X – μ) / σ = (87 – 80) /4 = **1.75**.

This tells us that an exam score of 87 lies **1.75 standard deviations** *above* **the mean**.

Example:

Suppose X ~ N(5, 6). This says that X is a normally distributed random variable with mean μ = 5 and standard deviation σ = 6. Suppose x = 17.

Solution:

z = x – μ/ σ = 17 – 5/ 6 = 2

This means that x = 17 is two standard deviations (2σ) above or to the right of the mean μ = 5.

Notice that: 5 + (2)(6) = 17 (The pattern is μ + zσ = x)

Now suppose x = 1. Then:

z = x – μ/σ = 1 – 5/ 6 = –0.67 (rounded to two decimal places)

This means that x = 1 is 0.67 standard deviations (–0.67σ) below or to the left of the mean μ = 5.

Notice that: 5 + (–0.67)(6) is approximately equal to one .

(This has the pattern μ + (–0.67)σ = 1)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than

Why Are Z-Scores Useful?

Z-scores are useful because they give us an idea of how an individual value compares to the rest of a distribution.

The Empirical Rule:

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the Empirical Rule states the following:

• About 68% of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).

• About 95% of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).

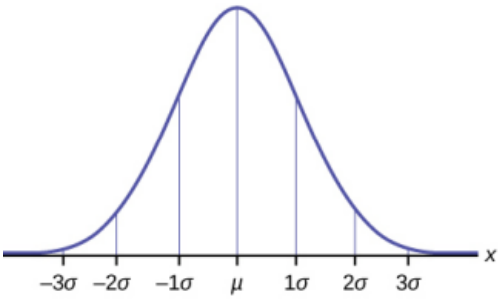
• About 99.7% of the x values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.

• The z-scores for +1σ and –1σ are +1 and –1, respectively.

• The z-scores for +2σ and –2σ are +2 and –2, respectively

. • The z-scores for +3σ and –3σ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.



EXAMPLE:

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males in 1984 to 1985. Then Y ~ N(172.36, 6.34).

a. About 68% of the y values lie between what two values? These values are _____. The z-scores
b. . About 95% of the y values lie between what two values? These values are _____. The z-scores are _____ respectively.
c. . About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Solution:

a. About 68% of the values lie between 166.02 cm and 178.7 cm. The z-scores are –1 and 1.
b. About 95% of the values lie between 159.68 cm and 185.04 cm. The z-scores are –2 and 2
c. About 99.7% of the values lie between 153.34 cm and 191.38 cm. The z-scores are –3 and 3.

PRACTICE QUESTIONS

For FOIT:

Introductory Statistics from OpenStax Book

Page no :(365_378)

Question no (1-30).

USING THE NORMAL DISTRIBUTION

Example:

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five. Find the probability that

a) randomly selected student scored more than 65 on exam.

Solution:

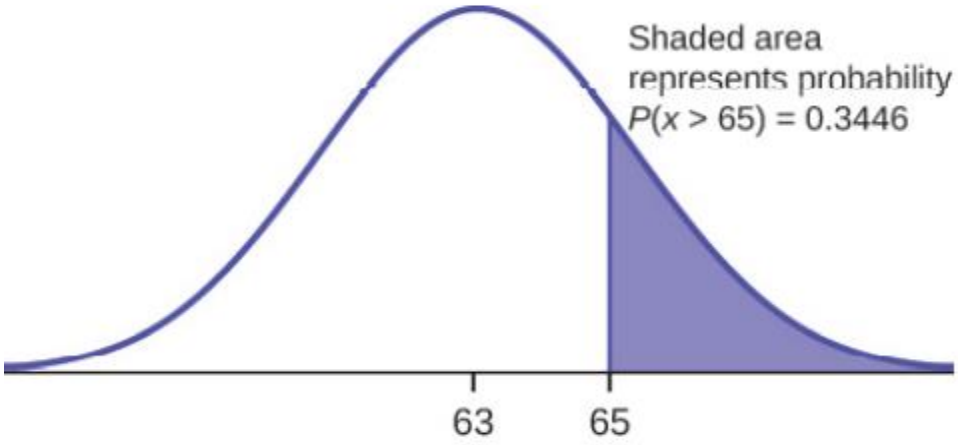
(a). Let X = a score on the final exam. X ~ N (63, 5), where μ = 63 and σ = 5.

P (x > 65) = ?

$$z = \frac{65 - 63}{5} = 0.4$$

Area to the left is 0.6554.

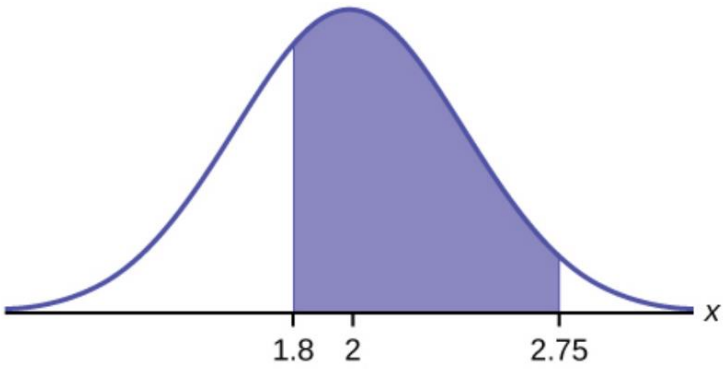
$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$



.

Example: For a certain type of computer, the length of time between charges of the battery is normally distributed with a mean of two hours and a standard deviation of half an hour. John owns one of these computers and wants to know the probability that the length of time will be between 1.8 and 2.75 hours per day.

Solution: Let x be the random variable that represents the length of time. It has a mean of 2 and a standard deviation of 0.5. We have to find the probability that x is between 1.8 and 2.75 or P(1.8< x < 2.75)=?
For x = 1.8 , z = (1.8 - 2) / 0.5 = -0.4
For x = 2.75 , z = (2.75 - 2) / 0.5 = 1.5
P(1.8< x < 2.75) = P(-0.4< z < 1.5) = [area to the left of z = 1.5] - [area to the left of z =- 0.4]
= 0.93319 - 0.34458 = 0.58861
The probability that John's computer has a length of time between 1.8 and 2.75 hours is 0.58861.



Example: In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information, answer the following questions (round answers to one decimal place).

a. Calculate the interquartile range (IQR).

b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution:

a

$IQR = Q_3 - Q_1$
Calculate $Q_3 = 75^{th}$ percentile and $Q_1 = 25^{th}$ percentile.
invNorm(0.75,36.9,13.9) = Q_3 = 46.2754
invNorm(0.25,36.9,13.9) = Q_1 = 27.5246
$IQR = Q_3 - Q_1$ = 18.8

b.
Find k where $P(x \geq k) = 0.40$ ("At least" translates to "greater than or equal to.")
0.40 = the area to the right.
Area to the left = 1 – 0.40 = 0.60.
The area to the left of k = 0.60.
invNorm(0.60,36.9,13.9) = 40.4215.
k = 40.4.
Forty percent of the ages that range from 13 to 55+ are at least 40.4 years.

PRACTICE QUESTIONS

For FOIT:

Introductory Statistics from OpenStax Book

Page no :(365_378)

Question no (74-80).

Version 1

Math, FOIT

Part -III

Seher Malik

Lecturer Statistics
Department of Mathematics
Faculty of Science and Technology
University of Central Punjab, Lahore

ESTIMATION

We can Estimate the unknown Population Parameter in two ways:

1. Point Estimation
2. Interval Estimation

Point Estimation:

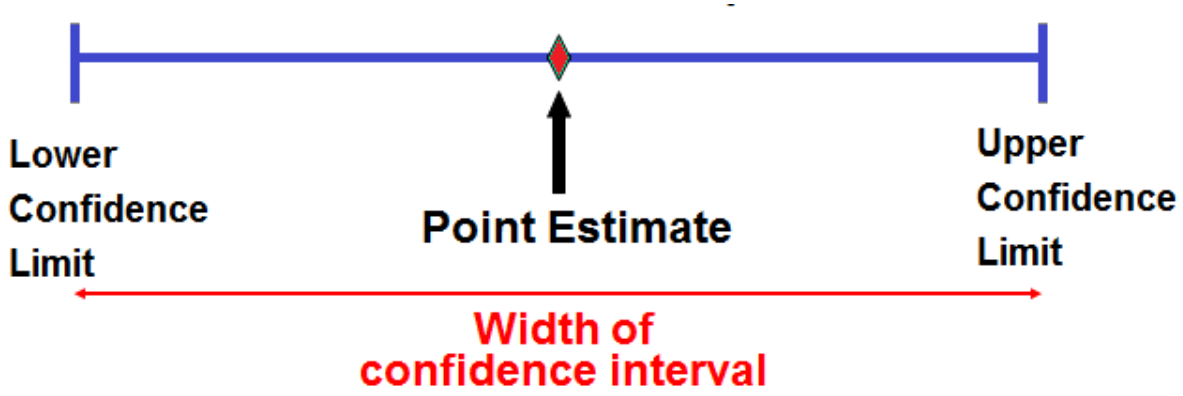
The formula for sample mean, $\bar{x}$, is the point estimator for the population mean μ and the calculated numerical answer for $\bar{x}$ is the Point Estimate for population mean μ.

Sample standard deviation '**s**' is used to estimate the population standard deviation σ. $\bar{x}$ and s are statistics.

A point estimator is a single number.

Interval Estimation:

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. It provides a range of reasonable values in which we expect the population parameter to fall.



Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

Example 23.1:

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.

2. Let σ = 3 and n = the number of students surveyed.

3. Construct the interval $\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between _____ and _____.

A Single Population Mean using the Normal Distribution

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated EBM). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ.

The confidence interval estimate will have the form:

(point estimate - error bound, point estimate + error bound) or,

In symbols, ($\bar{x}$ – EBM, $\bar{x}$ +EBM)

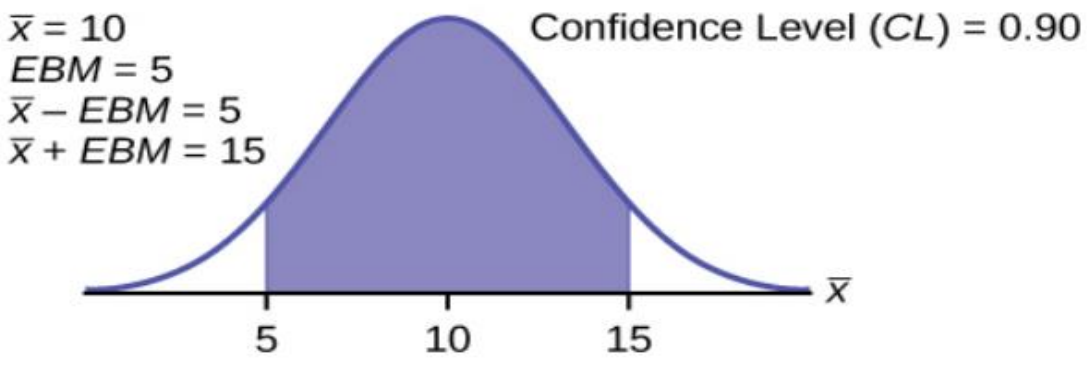
The margin of error (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions. There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically, α + CL = 1.

Example 23.2:

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5.

$\bar{x}$ = 7 and EBM = 2.5 The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5). If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x}$ = 10, and we have constructed the 90% confidence interval (5, 15) where EBM = 5. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.



To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

In summary, as a result of the central limit theorem:

• $\bar{x}$ is normally distributed, that is, $\bar{X} \sim N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$

• When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the "standard error of the mean" in order to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

Calculating the Error Bound (EBM)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

• EBM = $\left(Z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

Constructing the Confidence Interval

• The confidence interval estimate has the format ($\bar{x}$ – EBM, $\bar{x}$ + EBM) .

Example 23.3:

Suppose scores on exams in computer science are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of computer science exam scores.

Solution:

To find the confidence interval, you need the sample mean, $\bar{X}$, and the EBM.

EBM = $\left(Z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

$\bar{x} = 68$; n=36; CL=0.90; $\sigma = 3$; $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$$\frac{\alpha}{2} = 0.05;\ Z_{\frac{\alpha}{2}} = Z_{0.05}$$

The area to the right of $Z_{0.05}$is 0.05 and the area to the left of $Z_{0.05}$is 1 – 0.05 = 0.95

$$Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$$

EBM = $(1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$

$\bar{x}$ – EBM=68 - 0.8225 = 67.1775

$\bar{x}$ +EBM=68 + 0.8225 = 68.8225

The 90% confidence interval is (67.1775, 68.8225).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all computer science students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean computer science exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Example 23.4: Suppose we change the original problem in Example 23.3 by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean computer science exam score.

Solution:

To find the confidence interval, you need the sample mean, $\bar{X}$, and the EBM.

EBM = $\left(Z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

$\bar{x} = 68$;n=36; CL=0.95; $\sigma = 3$; $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$$\frac{\alpha}{2} = 0.025;\ Z_{\frac{\alpha}{2}} = Z_{0.025}$$

The area to the right of $Z_{0.025}$is 0.025 and the area to the left of $Z_{0.025}$is 1 – 0.025 = 0.975

$$Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.96$$

EBM = $(1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$

$\bar{x}$ – EBM=68 - 0.98 = 67.02

$\bar{x}$ +EBM=68 + 0.98 = 68.98

The 95% confidence interval is (67.02, 68.98).

Interpretation

We estimate with 95% confidence that the true population mean for all computer science exam scores is between 67.02 and 68.98.

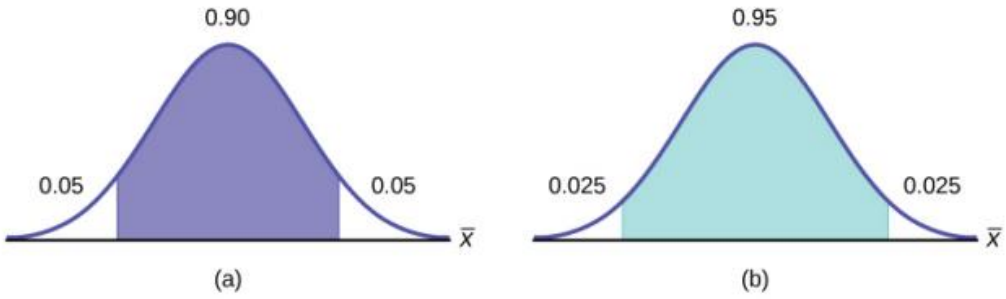
Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean computer science exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the

confidence interval actually does contain the true value of the population mean for all computer science exam scores, the confidence interval necessarily needs to be wider.



Summary: Effect of Changing the Confidence Level

• Increasing the confidence level increases the error bound, making the confidence interval wider.

• Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

For FOIT:

Introductory Statistics from OpenStax Book

Exercise (13-22), (95,96,99)

CONFIDENCE INTERVALS

Calculating Error Bound and Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backward to find both the error bound and the sample mean.

Finding the Error Bound

• From the upper value for the interval, subtract the sample mean,

• OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

• Subtract the error bound from the upper value of the confidence interval,

OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 24.1:

Suppose we know that a confidence interval is (68.18, 69.82) and we want to find the error bound. We may know that the sample mean is 69, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

• If we know that the sample mean is 69: EBM = 69.82 – 69 = 0.82.

• If we don't know the sample mean: EBM = $\frac{(69.82-68.18)}{2} = 0.82$

Calculate the Sample Mean:

• If we know the error bound: $\bar{x} = 69.82 - 0.82 = 69$

• If we don't know the error bound: $\bar{x} = \frac{(69.82+68.18)}{2} = 69$

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

EBM = $\left(Z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

The formula for sample size is n = $\frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n.

In this formula, z is $Z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 24.2: The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

From the problem, we know that σ = 15 and EBM = 2.

z = $z_{0.025}$ = 1.96, because the confidence level is 95%.

n = $\frac{z^2\sigma^2}{EBM^2} = \frac{1.96^2 15^2}{2^2}$ = 216.09 using the sample size equation.

Use n = 217: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

A Single Population Mean using the student t Distribution

Some statisticians used the normal distribution approximation for large sample sizes and used the student's t-distribution only for sample sizes of at most 30.

The notation for the Student's t-distribution (using *T* as the random variable) is:

- $T \sim t_{df}$ where $df = n - 1$.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the **error bound for a population mean** is:

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\sigma}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is:

$(\bar{x} - EBM,\ \bar{x} + EBM)$.

Example 24.3: Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was 0.78 hour. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

Solution:

To find the confidence interval, you need the sample mean, $\bar{X}$, and the EBM.

EBM = $\left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$

$\bar{x} = 7.1$;n=10; CL=0.95; $s = 0.78$; $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$$\frac{\alpha}{2} = 0.025;\ t_{\frac{\alpha}{2}} = t_{0.025}$$

The area to the right of t is 0.025 and the area to the left of $t_{0.025}$is 1 – 0.025 = 0.975

$$t_{\frac{\alpha}{2}} = t_{0.05} = 2.14$$

EBM = $(2.14)\left(\frac{0.78}{\sqrt{10}}\right) = 0.56$

$\bar{x}$ – EBM=7.1 - 0.56 = 6.54

$\bar{x}$ +EBM=7.1+ 0.56 = 7.66

The 95% confidence interval is (6.54, 7.66).

Interpretation

We estimate with 95% confidence that the true population mean for sleep time is between 6.54 and 7.66.

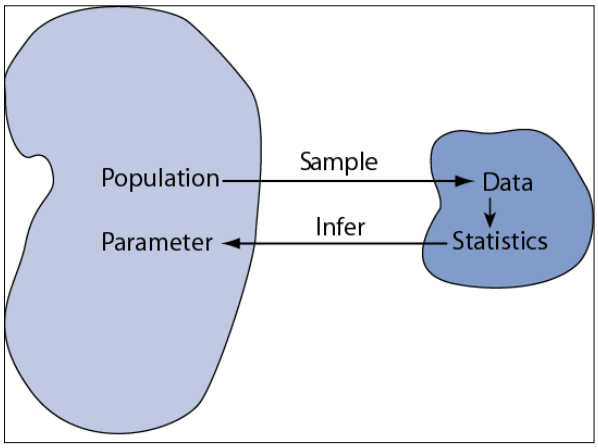
For FOIT:

Introductory Statistics from OpenStax Book

Exercise (43-48), (105-107)

HYPOTHESIS TESTING

Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter.



A statistician will make a decision about these claims. This process is called " hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

Null and Alternative Hypotheses

The actual test begins by considering two hypotheses. They are called the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints.

H_0: The null hypothesis: It is a statement of no difference between sample means or proportions or no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0.

Ha: The alternative hypothesis: It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

H_0	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Example 25.1: H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$

H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Example 25.2: We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0).

The null and alternative hypotheses are:

H_0: $\mu = 2.0$

H_a: $\mu \neq 2.0$

Example 25.3: We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

H_0: $\mu \geq 5$

H_a: $\mu < 5$

Outcomes and the Type I and Type II Errors:

ACTION	H_0 IS ACTUALLY	...
	True	False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

Example 25.4: Suppose the null hypothesis, H0, is: Frank's rock-climbing equipment is safe.

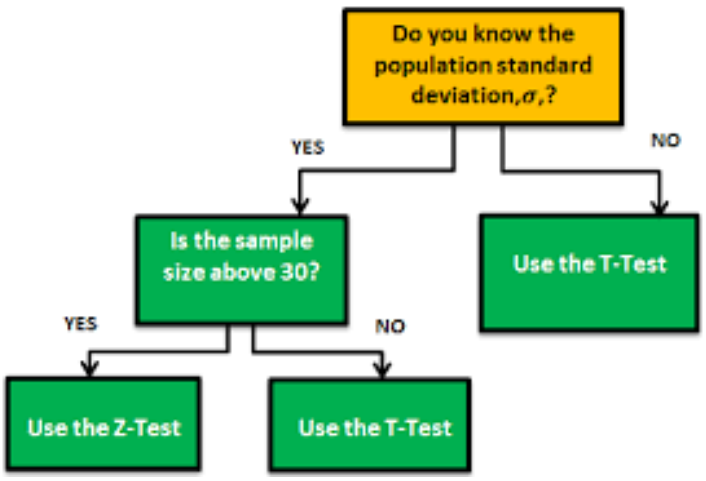
Type I error: Frank thinks that his rock-climbing equipment may not be safe when, in fact, it really is safe.

Type II error: Frank thinks that his rock-climbing equipment may be safe when, in fact, it is not safe.

α = probability that Frank thinks his rock-climbing equipment may not be safe when, in fact, it really is safe.

β = probability that Frank thinks his rock-climbing equipment may be safe when, in fact, it is not safe. Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock-climbing equipment is safe, he will go ahead and use it.)

Distribution Needed for Hypothesis Testing



If you are testing a single population mean, the distribution for the test is for means:

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \text{ or } t_{df}$$

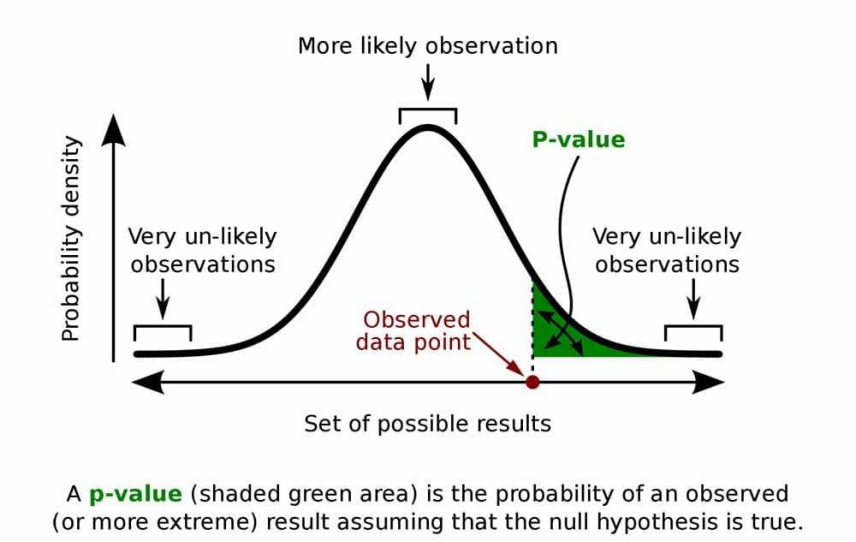
Z-Test

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

T-Test

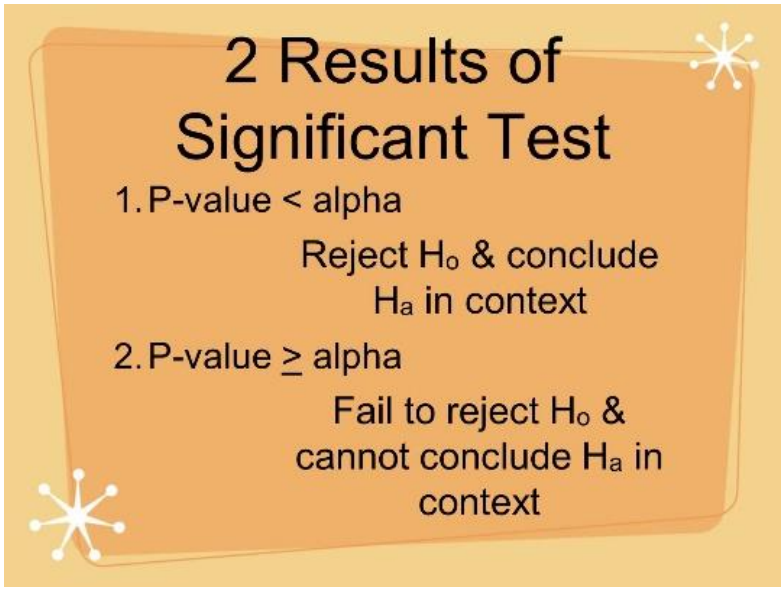
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Using the Sample to Test the Null Hypothesis



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Decision and Conclusion



Z-Test

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \leq -z_\alpha$	Reject H_0 if $z \geq z_\alpha$	Reject H_0 if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$

Example 25.5: A researcher claims that the average wind speed in a certain city is 8 miles per hour. A sample of 32 days has an average wind speed of 8.2 miles per hour. The standard deviation of the population is 0.6 mile per hour. At α= 0.05, is there enough evidence to reject the claim? Use the P-value method.

Solution:

- $H_o: \mu = 8$

  $H_1: \mu \neq 8$
- α=0.05
- Test-Statistic:

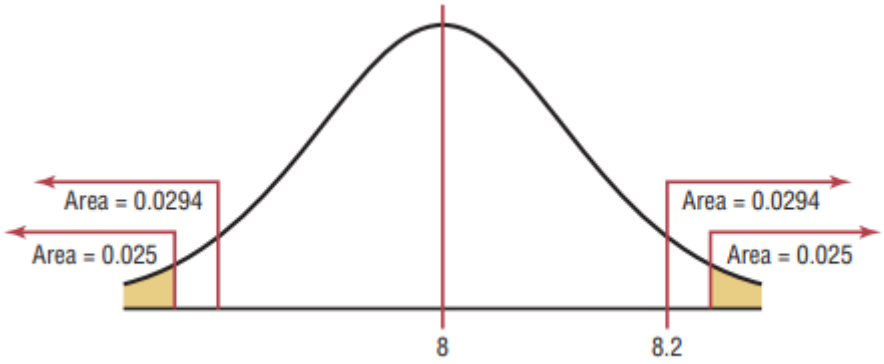
$$z = \frac{8.2 - 8}{0.6/\sqrt{32}} = 1.89$$

- **P-value:** the corresponding area for z =1.89. It is 0.9706. Subtract the value from 1.0000.

1- 0.9706= 0.0294

Since this is a two-tailed test, the area of 0.0294 must be doubled to get the P-value.

2(0.0294)= 0.0588



- **Conclusion:** The decision is to not reject the null hypothesis since the P-value is greater than 0.05. There is not enough evidence to reject the claim that the average wind speed is 8 miles per hour.

T-Test

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $t \leq -t_\alpha$	Reject H_0 if $t \geq t_\alpha$	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

Example 25.6: An educator claims that the average marks of computer science students in UCP is less than 60 marks. A random sample of eight students is selected, and their marks are shown. Is there enough evidence to support the educator's claim at a 0.10?

60 56 60 55 70 55 60 5582

Solution:

- H_o: : $\mu \geq 60$

  H_a: $\mu < 60$

- α=0.10
- Test-Statistic:

$$t = \frac{58.88 - 60}{5.08/\sqrt{8}} = -0.624$$

- **Critical-value:** The d.f. 7, the critical value is -1.415.

- **Conclusion:** Do not reject the null hypothesis since -0.624 falls in the noncritical region. There is not enough evidence to support the educator's claim that the average marks of computer science student is less than 60 marks.

For FOIT:

Introductory Statistics from OpenStax Book

Null and Alternative Hypotheses

Exercise (1-10), (35-39)

Full Hypothesis Test Examples

(74,75,76)

LINEAR REGRESSION AND CCORRELATION

Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

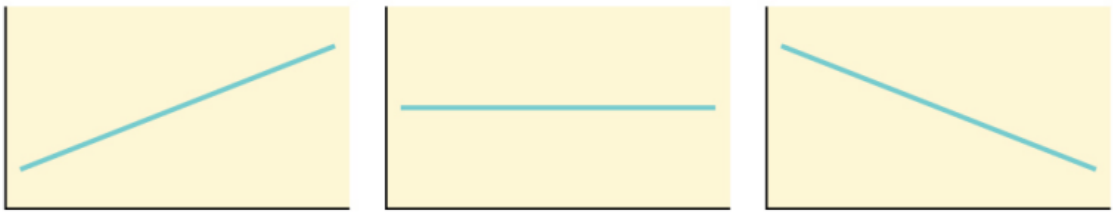
$$y = a + bx$$

where a and b are constant numbers.

The variable x is the independent variable, and y is the dependent variable. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Slope and Y-Intercept of a Linear Equation

For the linear equation y = a + bx, b = slope and a = y-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point (0, a) where the line crosses the y-axis.



Three possible graphs of y = a + bx.

(a) If b > 0, the line slopes upward to the right.

(b) If b = 0, the line is horizontal.

(c) If b < 0, the line slopes downward to the right.

Example 27.1: Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is y = 25 + 15x. What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session.

The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0).

The slope is 15 (b = 15). For each session, Svetlana earns $15 for each hour she tutors.

Scatter Plots

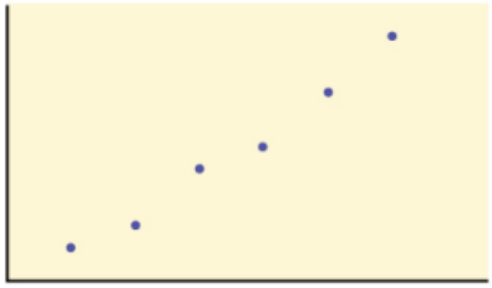
we need to examine a way to display the relationship between two variables x and y. The most common and easiest way is a scatter plot. The following example illustrates a scatter plot.

A scatter plot shows the direction of a relationship between the variables. A clear direction happens when there is either:

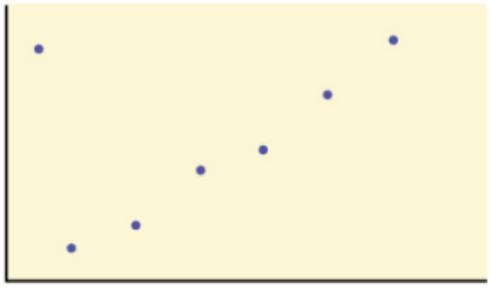
• High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.

• High values of one variable occurring with low values of the other variable.

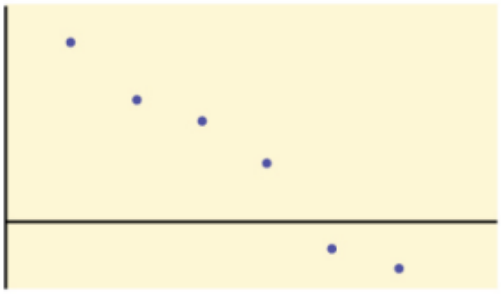
When you look at a scatterplot, you want to notice the overall pattern and any deviations from the pattern. The following scatterplot examples illustrate these concepts.



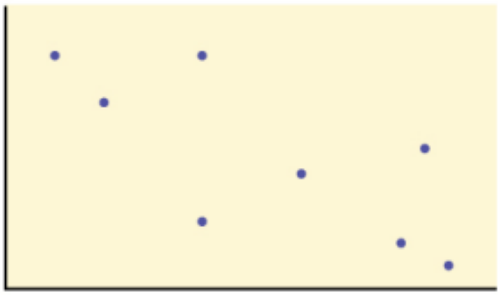
(a) Positive linear pattern (strong)



(b) Linear pattern w/ one deviation



(a) Negative linear pattern (strong)

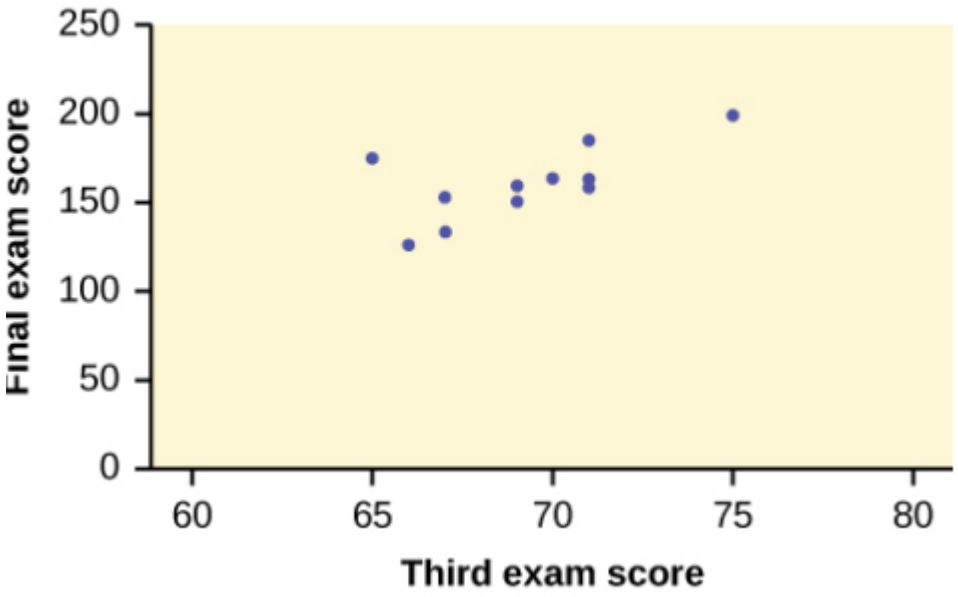


(b) Negative linear pattern (weak)

Example 27.2: A random sample of 11 computer science students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Solution:



The Regression Equation

A least-squares regression line:

$$\hat{y} = a + bx$$

$$a = \bar{y} + b\bar{x}$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2},$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Example 27.3: To illustrate the least squares method, suppose data were collected from a sample of 7 comuter science students.The number of absences and the final grades of seven randomly selected students from a class.

Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{7(3745) - (57)(511)}{7(579) - (57)^2}$$

$$b = -3.622$$

$$a = \bar{y} - b\bar{x}$$

$$a = 73 - (-3.622(8.14))$$

$$a = 102.493$$

$$\hat{y} = a + bx$$

$$\hat{y} = 102.493 - 3.622x$$

Slope

The slope of the estimated regression equation (b= -3.622) is negative, implying that as more absences a student has, the lower is his or her grade.

The Correlation Coefficient r

Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y.

The correlation coefficient, r, developed by Karl Pearson in the early 1900s, is numerical and provides a measure of the strength and direction of the linear association between the independent variable x and the dependent variable y. The correlation coefficient is calculated as

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where n = the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

• The value of r is always between –1 and +1: $-1 \leq r \leq 1$.

• The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to –1 or to +1 indicate a stronger linear relationship between x and y.

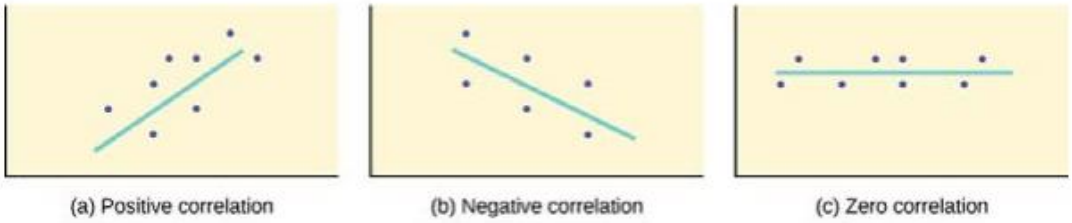
• If r = 0 there is no linear relationship between x and y (no linear correlation).

• If r = 1, there is a perfect positive correlation. If r = –1, there is a perfect negative correlation. In both these cases, all the original data points lie on a straight line. Of course, in the real world, this will not generally happen. What the SIGN of r tells us

• A positive value of r means that when x increases, y tends to increase, and when x decreases, y tends to decrease (positive correlation).

• A negative value of r means that when x increases, y tends to decrease, and when x decreases, y tends to increase (negative correlation).

• The sign of r is the same as the sign of the slope, b, of the best-fit line.



(a) A scatter plot showing data with a positive correlation. $0 < r < 1$

(b) A scatter plot showing data with a negative correlation. $-1 < r < 0$

(c) A scatter plot showing data with zero correlation. r = 0

Example 27.4:

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

$$r = \frac{7(3745) - (57)(511)}{\sqrt{[7(579) - (57)^2][7(38993) - (511)^2]}}$$

$$r = -0.944$$

The value of r suggests a strong negative relationship between a student's final grade and the number of absences a student has. That is, the more absences a student has, the lower is his or her grade.

For FOIT:

Introductory Statistics from OpenStax Book

Page no: 714-716

Exercise (10-25)

LINEAR REGRESSION AND CORRELATION

Prediction:

Example 28.1: Suppose you want to estimate, or predict, the mean final exam score of computer science students who received 73 on the third exam. The exam scores (x-values) range from 65 to 75. Since 73 is between the x-values 65 and 75, substitute x = 73 into the equation. Then:

$$\hat{Y} = -173.51 + 4.83(73) = 179.08$$

We predict that computer science students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 28.2: What would you predict the final exam score to be for a student who scored a 90 on the third exam?

$$\hat{Y} = = -173.51 + 4.83(90) = 261.19$$

The final-exam score is predicted to be 261.19. The largest final-exam score can be is 200.

Example 28.3: If we believe the least squares estimated regression equation adequately describes the relationship between x and y, it would seem reasonable to use the estimated regression equation to predict the value of y for a given value of x. For example, if we wanted to predict the final grade of students having 9 absentees we would compute

$$\hat{y} = 102.493 - 3.622(9) = 69.895$$

Outliers: In some data sets, there are values (observed data points) called outliers. Outliers are observed data points that are far from the least squares line. They have large "errors", where the "error" or residual is the vertical distance from the line to the point. Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Numerical Identification of Outliers:

In Table, the first two columns are the third-exam and final-exam data. The third column shows the predicted ŷ values calculated from the line of best fit: ŷ = –173.5 + 4.83x. The residuals, or errors, have been calculated in the fourth column of the table: observed y value−predicted y value = y − ŷ.

s is the standard deviation of all the y − ŷ = ε values where n = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{SSE}{n-2}}$$

Example 28.4:

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	35
67	133	150	-17
71	185	169	16
71	163	169	-6
66	126	145	-19
75	198	189	9
67	153	150	3
70	163	164	-1
71	159	169	-10
69	151	160	-9
69	159	160	-1

We are looking for all data points for which the residual is greater than 2s = 2(16.4) = 32.8 or less than –32.8. Compare these values to the residuals in column four of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

Covariance:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

which is known as the **covariance** and denoted by $\text{cov}(X, Y)$ or s_{xy}. For shorthand it is normally written as

$$\frac{1}{n}\Sigma\,(x - \bar{x})(y - \bar{y})$$

where the summation over i is assumed.

$$\frac{1}{n}\Sigma\,(x - \bar{x})(y - \bar{y}) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y}$$

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

$$\frac{1}{n}\Sigma\,(x - \bar{x})(y - \bar{y}) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y}$$

$$S_{xy} = \frac{1}{7}(3745) - (73)(8.14)$$

$$S_{xy} = 535 - 594.22$$

$$S_{xy} = -59.22$$

For FOIT:

Introductory Statistics from OpenStax Book

Page no: 717-718

Exercise (31-39)