Lecture Note
Introductory Statistics(Stat2161)

Zelalem T

December 20, 2017

Chapter: Introduction
Definition and Classification

Definition:

- **Statistics**: a science which deals with collection, organization, presentation, analysis and interpretation of numerical data.
- The tools of statistics are employed in many fields such as in business, education, psychology, agriculture, medicine, economics, etc.
- **Biostatistics**: is the application of statistical methods to medical, biological and public health related problems.
- Therefore, biostatistics is a scientific treatment given to medical data derived from group of individuals or patients.

Classification of Statistics:

- Statistics can be broadly classified into two categories: *Descriptive Statistics* and *Inferential Statistics*.
- **Descriptive Statistics**: deals with methods or techniques used in collection, organization, presentation and analysis of data without making any conclusions or inferences.
    - It encompasses the tabular, graphical or pictorial display of data, condensation of large data into tables, preparation of summary measures to give a concise description of complex information.
- **Inferential Statistics**: deals with the method of drawing or making conclusion about the characteristics of the population based on the results of a sample.
    - It consists of performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Stages in statistical Investigation

- Stages or steps involved in any statistical investigations include *Collection*, *Organization*, *Presentation*, *Analysis*, and *Interpretation* of data.
- **Collection of data**: the process of measuring, gathering, assembling the raw data.
    - It is the basis (foundation) of any statistical work.
    - Data can be collected from records, from surveys (either faceto-face, telephone, or postal), by direct observation or by measuring or counting.
- **Organization of data**: editing and classifying the collected data according to some characteristics to make it easier for presentation.
    - The collected data might involve irrelevant figures, incorrect facts, omissions and mistakes.
    - Errors that may occur during data collection will have to be edited.

- **Presentation of data**: organizing data in the form of tables or diagrams.
- **Analysis of data**: the process of extracting relevant information from the summarized data, mainly through the use of elementary mathematical operation.
- **Interpretation of data**: drawing conclusion on the basis of the above information.

Definition of some basic terms

- **A population**: consists of all subjects that are being studied.
- **A sample**: is a group of subjects selected from a population.
- **A census**: is a complete enumeration of every item in a population.
- **Sample Survey**: is an enumeration of units from a much smaller group of the population.
- **Parameter**: is a characteristic or measure obtained by using all the data values from a specific population.
- **Statistic**: is a characteristic or measure obtained by using the data values from a sample.

- **Sampling**: is the process of taking a sample from a population.
- **Sample size**: is the number of elements or observation to be included in the sample.
- **Variable**: is a characteristics under study that assumes different values for different elements.

Limitation of Statistics

- It does not **directly** study qualitative characteristics.
- Deals with only aggregate of facts and not with individual data items.
- Statistical data are only approximately and not mathematical correct.
- Statistics can be easily misused and therefore should be used by experts.

Type of variables and scales of measurements

Type of variables:

- Variables can be classified as *qualitative* or *quantitative*.
- **Qualitative(or categorical) variables**: are variables that can be placed into distinct categories, according to some characteristic or attribute.
    - It can not be measured/counted.
    - **Example**: gender, ethnic group, type of drug, stages of breast cancer (I, II, III, or IV), degree of pain (minimal, moderate, severe or unbearable), religious affiliation, socio-economic status(low, medium, high), place of birth, etc.
- **Quantitative (or numerical) variables**: are variables that can be measured and/or counted.

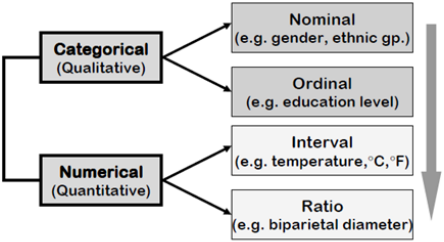
- Quantitative variables can be further classified into two groups: *discrete* and *continuous*.
- **Discrete variables**: assume distinct values.
  - They are obtained by counting.
  - **Example:** number of beds in a hospital; number of episodes of diarrhea in the first five years of life; number of HIV patients being infected by TB in a hospital.
- **Continuous variables**: assume an infinite number of values between any two specific values.
  - They are obtained by measuring.
  - **Example**: weight, height, blood pressure, age, temperature.
- **Note**: Data are individual observations or values of a variable.

Scales of measurements

- In addition to being classified as qualitative or quantitative, variables can be classified by "*how*" they are categorized, counted, or measured, and often we use the term *measurement scales*.
- Measurement scale refers to the property of value assigned to the data based on the properties of order, distance and fixed zero.
- Four common types of scales are used: *nominal*, *ordinal*, *interval*, and *ratio*.
- **Nominal scales of measurement**: classifies data into mutually exclusive (non-overlapping) categories/groups in which no order or ranking can be imposed on the data.
  - No arithmetic and relational operation can be applied.
  - **Example**: Marital status(married, single, widow, divorce); Patient ID; Sex (Male, Female); Religious affiliation.

- **Ordinal scales of measurement**: classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.
    - Arithmetic operations are not applicable but relational operations are applicable.
    - **Example**: Economic status(poor, medium, high);Rating scales (Excellent, Very good, Good, Fair, poor).
- **Interval scales of measurement**: ranks data, and precise differences between units of measure do exist; however, there is no meaningful(true) zero.
    - All arithmetic operations except division and multiplication are applicable (Thus, ratio are meaningless).
    - Relational operations are also possible.
    - **Example**: Temperature- there is a meaningful difference of 1^oF between each unit, such as 72 and 73^oF, however 0^oF does not mean no heat at all. Further more, we can **not** say 40^oF is twice as hot as 20^oF.

- **Ratio scales of measurement**: possesses all the characteristics of interval measurement, and there exists a true zero.
  - True ratios exist when the same variable is measured on two different members of the population.
  - All arithmetic and relational operations are applicable.
  - **Example**: Weight, Height, Age, Number of patients.
- The scales of measurements together with the type of variables stated in terms of increasing information content, are summarized as follows .



Methods of data collection

- According to source, data can be classified as *Primary data* and *Secondary data*.
- **Primary data**:- data measured or collected by the investigator or the user directly from the source.
    - Primary data can be collected in a variety of ways; One of the most common methods is through the use of **surveys**.
    - Three of the most common methods of doing survey are *telephone survey*, *mailed questionnaire*, and *personal interview*.
- **Secondary data**:- data gathered or compiled from published and unpublished sources or files.
- **Note:** Data which are primary for one may be secondary for the other.

Chapter Two
Descriptive Statistics: Displaying the Data

- Data is collected by researcher so that they can give solutions to the research question that they started with.
- The observations made on the subjects one after the other is called **raw data** or **unorganized data**.
- Raw data becomes useful only when they are arranged and organized in a manner that we can extract information from the data and communicate it to others.
- Data can be displayed in either **tabular** form or **graphical** form.
- Tables are used to categorize and summarize data while graphs are used to provide an overall visual representation.

- The most convenient method of organizing data is to construct *a frequency distribution.*
- **A frequency distribution:** is the organization of raw data in table form, using classes and frequencies.
- **Frequency:** is the number of values in a specific class of the distribution.
- There are two types of frequency distributions that are most often used.
  - Categorical frequency distribution
  - Numerical frequency distribution

- **Categorical frequency distribution**: is used for data that can be placed in specific categories, such as nominal- or ordinal-level data.
  - **Example 2.1**: The following data are taken from the medical records department at a certain hospital. The data include the blood type and gender(in bracket) of patients.

A(M)	B(F)	B(F)	AB(M)	O(M)
O(F)	O(F)	B(F)	AB(F)	B(M)
B(F)	B(M)	O(M)	A(M)	O(M)
A(M)	O(M)	O(M)	O(F)	AB(M)
AB(F)	A(F)	O(M)	B(F)	A(F)

Construct a frequency distribution for the variable blood type.

Solution: Since the data are categorical, specific classes can be used. There are four blood types to be used as the classes for the distribution: A, B, O, and AB.

Thus, the frequency distribution of blood types of the patients would be:

Class	Tally	Frequency	Percent(%)
A	//// /	5	20
B	//// ///	7	28
O	//// //// /	9	36
AB	////	4	16
Total		25	100

Note: Percentages and Tallies are not normally part of a frequency distribution.

- Such type of tabulation which takes only **one variable** for classification is called one-way table. When two variables are involved, the table is referred to as cross tabulation or two-way table.

- **Example 2.2**: Using the above data given in Example 2.1, create a two-way table that we could use to compare gender to blood type.
    - **Solution:** The distribution of Gender and blood type of patients is given by:

Gender	Blood Type				
	A	B	O	AB	Total
M	3	2	6	2	13
F	2	5	3	2	12
Total	5	7	9	4	25

- **Numerical frequency distribution**: is a frequency distribution of a variable that can be expressed in terms of number or numerics.
  - It can be either ***grouped*** or ***un-grouped*** depending on the range of the data values.
- **Ungrouped frequency distribution:** is a numerical frequency distribution in which a single data value is used to represent one class.
  - It is often constructed for data on discrete variable when the range of the data values is relatively small.
- **Example 2.3**: The following data give the numbers of visitors during visiting hours on a given evening for each of the 20 randomly selected patients at a hospital.

3	0	1	4	2	0	4	0	2	1
1	1	3	4	2	2	2	1	3	0

  Construct an ungrouped frequency distribution for the number of visitors.

Solution: Frequency distribution of number of visitors:

Number of Visitors	Tally	Frequency
0	////	4
1	//// /	5
2	//// /	5
3	///	3
4	///	3
Total		20

- **Grouped frequency distribution**: is a frequency distribution when several numbers are grouped in one class.
    - When the range of the data is large, the data must be grouped in to classes that are more than one unit in width.

Definitions

- **Class limits:** separates one class in a grouped frequency distribution from another.
  - The limits could actually appear in the data and have gaps between the upper limits of one class and lower limit of the next.
- **Unit of measurement (U)**: the distance between two possible consecutive measures. It is usually taken as 1, 0.1, 0.01, 0.001,... .
- **Class boundaries**: are used to separate the classes with no gaps in the frequency distribution.
  - **Note:** As a rule of thumb, the class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value.
  - The lower class boundary is obtained by subtracting U/2 from the corresponding lower class limit and the upper class boundary is found by adding U/2 to the corresponding upper class limit.

- **Class width:** the difference between the upper and lower class boundaries of any class. It is also the difference between the lower limits of any two consecutive classes or the difference between any two consecutive class marks.
- **Class mark (Mid points)**: it is the average of the lower and upper class limits or the average of upper and lower class boundary.
- **Cumulative frequency above:** it is the total frequency of all values greater than or equal to the lower class boundary of a given class.
- **Cumulative frequency below:** it is the total frequency of all values less than or equal to the upper class boundary of a given class.

Guidelines for classes:

- There should be between 5 and 20 classes.
- The classes must be mutually exclusive. Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes.
- The classes must be continuous. Even if there are no values in a class, the class must be included in the frequency distribution.
- The classes must be exhaustive. There should be enough classes to accommodate all the data.
- The classes must be equal in width. The exception here is the first or last class.
  - Afrequency distribution with an open-ended class is called an **open-ended** distribution.

Steps for constructing Grouped frequency Distribution

- Compute the Range(R) = Maximum - Minimum
- Select the number of classes desired, usually between 5 and 20 or to be specific use Sturges rule $k = 1 + 3.322 * log(n)$, where k is number of classes desired and n is total number of observation.
- Find the class width by dividing the range by the number of classes and rounding up, not off. i.e, $w = \frac{R}{k}$.
- Pick a suitable starting point less than or equal to the minimum value. The starting point is called the lower limit of the first class. Continue to add the class width to this lower limit to get the rest of the lower limits.
- To find the upper limit of the first class, subtract U from the lower limit of the second class. Then continue to add the class width to this upper limit to find the rest of the upper limits.

- Find the boundaries by subtracting U/2 units from the lower limits and adding U/2 units from the upper limits. The boundaries are also half-way between the upper limit of one class and the lower limit of the next class. !may not be necessary to find the boundaries.
- Tally the data and find the frequencies.

Example 2.4: The following data are on the number of minutes to travel from home to work for a group of automobile workers.

28	25	48	37	41	19	32	26	16	23	23	29	36
31	26	21	32	25	31	43	35	42	38	33	28	

Construct a grouped frequency distribution for this data.

Solution:

- Range = 48 - 16 = 32
- k = 1 + 3.322 log(25) = 5.64 = 6
- W = R/k = 32/6 = 5.33 = 6
- U=1 (data are placed without any decimal point)
- The final frequency distribution is:

Time (in Minutes)	**Number of workers**
16-21	3
22-27	6
28-33	8
34-39	4
40-45	3
46-51	1
Total	25

Graphical Presentation of Data

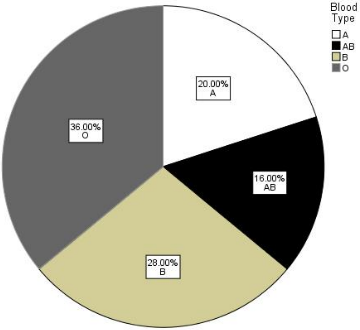
- Since it is difficult for a layman to understand complex distribution of data in tabular form, graphical presentation of data is better understood and appreciated by humans.
- Thus, the reason for displaying data graphically:
    - Investigators can have a better look at the information collected and the distribution of data.
    - To communicate this information to others quickly.
- Graphical representation brings out the hidden pattern and trends of the complex data sets.
- The most commonly used diagrammatic presentation for discrete as well as qualitative data include *Pie chart* and *Bar charts*. For continuous data, *Histogram*, *Frequency polygon* and *Cumulative frequency polygon* are the most common graphical representation.

- **Pie chart**: a circle in which the angle at the center is equal to its proportion multiplied by 360 (or, more easily, its percentage multiplied by 360 and divided by 100).
    - It is best when the total categories are between 2 to 6. If there are more than 6 categories, the diagram becomes too overcrowded.
    - **Example 2.5:** Draw a pie diagram using the frequency distribution given in Example 2.1.
      Solution: The angle of the sector could be obtained using:

      $$Angle\ of\ a\ sector = \frac{Frequency}{Total\ number\ of\ observation} * 360^0 C$$

      Thus, the angles for the individual categories and the corresponding pie chart are given as follows.

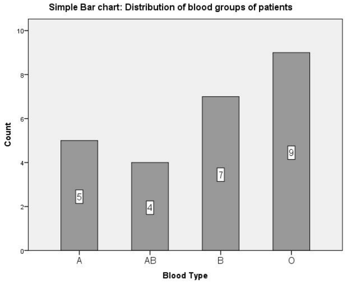
- $Blood\ group\ A = \frac{5}{25} * 360 = \frac{20}{100} * 360 = 72^{o}C$
- $Blood\ group\ B = \frac{7}{25} * 360 = \frac{28}{100} * 360 = 100.8^{o}C$
- $Blood\ group\ O = \frac{9}{25} * 360 = \frac{36}{100} * 360 = 129.6^{o}C$
- $Blood\ group\ AB = \frac{4}{25} * 360 = \frac{16}{100} * 360 = 57.6^{o}C$

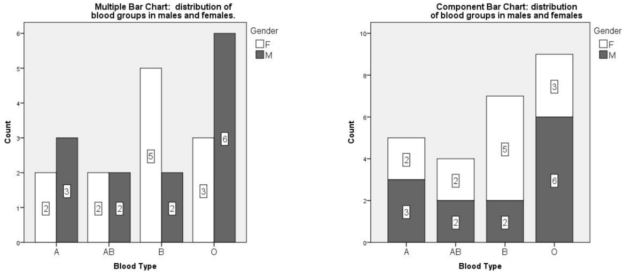


- **Bar chart**: represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies or percentage of the data.
    - The width of the bars is kept constant for all the categories and the space between the bars also remains constant throughout.
    - When we draw bar charts with only **one variable** or a single group it is called as **simple bar** chart and when **two variables** or two groups are considered it is called as **multiple bar** chart.
    - In multiple bar chart the two bars representing two variables are drawn adjacent to each other and equal width of the bars is maintained.
    - Another type of bar chart is the **component bar** chart wherein we have two qualitative variables which are further segregated into different categories or components. In this case the total height of the bar corresponding to one variable is further sub-divided into different components or categories of the other variable.

- **Example 2.6:** Using the data given in Example 2.1, draw (a) a simple bar chart showing the distribution of blood groups of patients; (b) a multiple bar chart showing the distribution of blood groups in males and females; and (c) a component bar chart showing the distribution of blood groups in males and females.

  Solution:





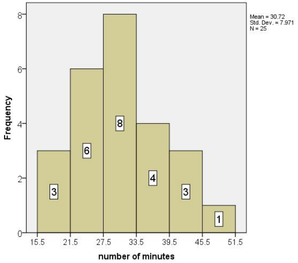
(a) Multiple Bar Chart (b) Component Bar Chart

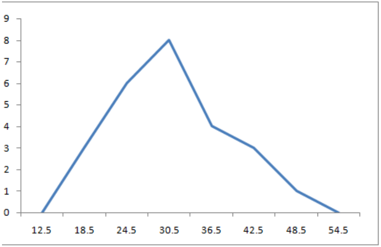
- **Histogram**: is used for quantitative continuous type of data where, on the X-axis, we plot the quantitative exclusive type of class intervals (or Class boundaries) and on the Y-axis we plot the frequencies.
  - In histogram, there are no gaps between the bars since the data measured on a continuous scale.

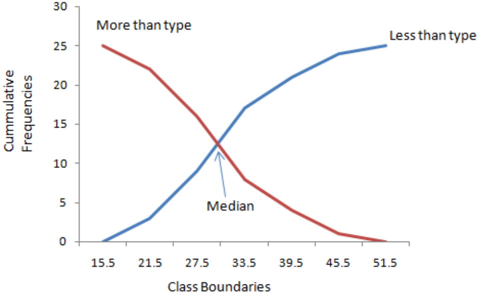
- **Frequency polygon/curve**: is a line/curve graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.
  - Conventionally, we consider one imaginary value immediately preceding the first value and one succeeding the last value and plot them with frequency = 0.
- **Ogive (cumulative frequency polygon)**: is a graph showing the cumulative frequency (less than or more than type) plotted against upper or lower class boundaries respectively.
  - i.e, class boundaries are plotted along the horizontal axis and the corresponding cumulative frequencies are plotted along the vertical axis. The points are joined by a free hand curve.

Example 2.7: Construct a histogram, frequency polygon and cumulative frequency polygon (less than and more than type) for the frequency distribution given in Example 2.4.

Solution:






30
25
20
15
10
5
0
More than type
Less than type
Cummulative Frequencies
Median
15.5
21.5
27.5
33.5
39.5
45.5
51.5
Class Boundaries

Chapter three:
Summarising the Data: Measures of Central Tendency

- When we want to make comparison between groups of numbers, it is good to have a single value that is considered to be a good representative of each group.
- Descriptive measures may be computed from the data of a sample or the data of a population.
- A descriptive measure computed from the data of a sample is called a **statistic**. Where as a descriptive measure computed from the data of a population is called a **parameter**.
- One way of finding a single value for describing the data set meaningfully is to use the summary measures such as measure of central tendencies.

- Measures of central tendency (average) helps to condense a mass of data into a single representative value.
- It gives the centrality measure of the data set which indicates where the observations are concentrated.
- A good average should posses the following:
    - It should be rigidly defined.
    - It should be based on all observation.
    - It should be as little as affected by extreme observations.
    - It should be capable of further algebraic treatment.
    - It should be as little as affected by fluctuations of sampling.
    - It should be ease to calculate and simple to understand.
- The three most commonly used measures of central tendency are the mean, the median, and the mode.

Arithmetic Mean

- It is the descriptive measure most people have in mind when they speak of the average.
- It is obtained by adding all the values and dividing by the number of values that are added.
- For raw data, it is calculated as:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Example 3.1**: The serum cholesterol level (mg/dl) of 10 subjects were given as follows:

192 242 203 212 175 284 256 218 182 228

The mean serum cholesterol level will be:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{192 + 242 + 203 + ... + 228}{10} = 219.2mg/dl$$

- For frequency distribution, the arithmetic mean can be computed as:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n}$$

where X_i is the class value (for ungrounded frequency distribution) or the class mark (for grouped frequency distribution); f_i is the frequency of the class; k is the number of classes and $n = \sum_{i=1}^{k} f_i$ is the total number of observations.

- **Example 3.2:** Consider the following frequency distribution of 9 subjects on serum cholesterol level (mg/dl):

Serum Cholestrol level	Mid-point(X_i)	Number of subjects (f_i)	$f_i X_i$
175-199	187	2	374
200-224	212	3	636
225-249	237	2	474
250-274	262	1	262
275-299	287	1	287
Total		9	2033

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n} = \frac{2033}{9} = 225.8889$$

Merits and Demerits of Arithmetic Mean

- **Merits:**
  - It is rigidly defined.
  - It is based on all observation and is unique.
  - It is suitable for further mathematical treatment.
  - It is stable average, i.e. it is not affected by fluctuations of sampling to some extent.
  - It is easy to calculate and simple to understand.

- **Demerits:**
    - It is affected by extreme observations.
    - It can not be used in the case of open end classes.
    - It can not be used when dealing with qualitative characteristics, such as intelligence, honesty, beauty.
    - It can be a number which does not exist in a serious.

Mode

- Mode is the most common value that repeats itself in the data set.
- The mode may be used for describing qualitative data.
- If all the values are different, there will not be a mode; on the other hand, a set of values may have more than one mode.

  Example 3.3:
    - **Unimodal**: 5, 3, 5, 8, 9
    - **Bimodal**: 8, 9, 9, 7, 8, 2, 5
    - **No mode**: 4, 12, 3, 6, 7

- In case of un-grouped frequency distribution, the value having the maximum frequency is the modal value.
- For grouped frequency distribution, the mode is defined as:

  $$\hat{X} = L_{mod} + \frac{\Delta_1}{\Delta_1 + \Delta_2} * w$$

  where, L_{mod} is the lower class boundary of the modal class; w is the width of the modal class; $\Delta_1 = f_{mod} - f_1$; $\Delta_2 = f_{mod} - f_2$; f_{mod} is the frequency of the modal class; f_1 is frequency of the class preceding the modal class; and f_2 is frequency of the class following the modal class.
  Note: The modal class is a class with the highest frequency.
- **Example 3.4:** Consider the previous frequency distribution of 10 subjects on serum cholesterol level given in Example 3.2.

- Modal class: [175-199]
- $\Delta_1 = 3 - 2 = 1, \Delta_2 = 3 - 2 = 1, L_{mod} = 174.5, and\ w = 25$
- $\hat{X} = 174.5 + \frac{1}{1+1} * 25 = 187$

Merits and Demerits of Mode

Merits:

- It is not affected by extreme observations.
- Easy to calculate and simple to understand.
- It can be calculated for distribution with open end class.

Demerits:

- It is not rigidly defined.
- It is not based on all observations
- It is not a stable average, i.e. it is affected by fluctuations of sampling to some extent.
- Often its value is not unique.

Median

- Median is a locative measure which is the middlemost observation after all the values are arranged in ascending or descending order.
    - That means, it is a value which divides the entire data set into 2 equal parts, when the data set is ordered in an ascending (or descending) fashion.
- For raw data or un-grouped frequency distribution, the median after arranging the data either ascending or descending order will be:

$$\tilde{X} = \begin{cases} (\frac{n+1}{2})^{th}\ obs & when\ n\ is\ odd \\ \frac{(n/2)^{th}\ \ obs + (n/2+1)^{th}\ \ obs}{2} & when\ n\ is\ even \end{cases}$$

- **Example:** Find the median of serum cholesterol levels considered in Example 3.1.

In the first step, we will order the data set in an ascending order as follows :

$$175, 182, 192, 203, 212, 218, 228, 242, 256, 284$$

Since n is 10 (even) we have two middle most observations as 212 and 218 (i.e. the 5th and 6th value).
Therefore, $\tilde{X} = \frac{212+218}{2} = 215$

- For grouped data, the median can be computed as:

$$\tilde{X} = L_{med} + \frac{w}{f_{med}}(n/2 - c)$$

  where, L_{med}=the lower class boundary of the median class; w=the size of the median class; f_{med}=frequency of the median class; c=the cumulative frequency preceding the median class, and n=the total number of observation.

  - **Note:** The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to $n/2$.

Exercise: Compute the median for the grouped frequency distribution given in Example 3.2.

Merits and Demerits of Median

- **Merits**:
  - Median is a positional average and hence not influenced by extreme observations.
  - Can be calculated in the case of open end intervals.
  - Median can be located even if the data are incomplete.
- **Demerits:**
  - It is not a good representative of data if the number of items is small.
  - It is not amenable to further algebraic treatment.
  - It is susceptible to sampling fluctuations.

Quantiles: Quartiles, Deciles and Percentiles (Reading Assignment).

Chapter 4

Measures of Dispersion

- Knowledge of central tendency alone is not sufficient for complete understanding of distribution.
- For example, if we have three series having the same mean, then it alone does not throw light on the composition of the data, hence to supplement it we need a measure which will tell us regarding the spread of the data.
- In contrast to measures of central tendency which describes the center of the data set, measures of variability describes the variability or spreadness of the observation from the center of the data.

- Objectives of measuring Variation:
    - To judge the reliability of measures of central tendency
    - To control variability itself.
    - To compare two or more groups of numbers in terms of their variability.
    - To make further statistical analysis.
- The measures of dispersion which are expressed in terms of the original unit of a series are termed as absolute measures.
- Such measures are not suitable for comparing the variability of two distributions which are expressed in different units of measurement and different average size.
- Relative measures of dispersions are a ratio or percentage of a measure of absolute dispersion to an appropriate measure of central tendency and are thus pure numbers independent of the units of measurement.

Range

- The range is the difference between the largest and smallest value in a set of observations.
- It is computed as: $\mathbf{R} = \mathbf{X_L} - \mathbf{X_S}$, where $X_L\ and\ X_S$ are respectively the largest and smallest values.
- It is the poorest measure of dispersion as it takes into account only two values(i.e, the largest and smallest values).
- It is highly affected by extreme observations.
- It is affected by the fluctuations of sampling.
- The relative measure of range, also called coefficient of range, is given by:

$$RR = \frac{X_L - X_S}{X_L - X_S}$$

Quartile Deviation

- Quartile deviation is halve of the inter quartile range.
- Interquartile range is the difference between the values of the two extreme quartiles.
- Thus, $QD = \frac{Inter\ quartile\ range}{2} = \frac{Q_3 - Q_1}{2}$
- It includes only the middle 50% of the observations.
- The relative measure of quartile deviation, also called coefficient of quartile deviation(CQD), is given by:

$$CQD = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation

- Mean deviation is the mean of the absolute deviations of observations from a given constant "A".
- For raw data, it is given by:

$$MD(A) = \frac{\sum |X_i - A|}{n}$$

  where A may be mean, median, mode or any constant.
- For frequency distribution(grouped and ungrouped), mean deviation could be computed using:

$$MD(A) = \frac{\sum_{i=1}^{k} f_i |X_i - A|}{n}$$

- The relative measure of MD, also known as Coefficient of mean deviation(CMD), is given by: $CMD(A) = \dfrac{MD(A)}{A}$.
- Mean deviation ignores the algebraic signs and hence to overcome this drawback we have another measure of variability called as **Variance**.

Variance

- Variance is the average of the squared deviations of each of the individual value from the mean.
- This variance is known as the **population variance** and is given by:

$$\sigma^2 = \begin{cases} \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} & for\ raw\ data \\ \frac{\sum_{i=1}^{k} f_i(X_i-\mu)^2}{N} & for\ frequency\ distribution \end{cases}$$

- In practice the population variance is unknown and hence, we will estimate it by the statistic is called sample variance.
- The sample variance can be computed as:

$$s^2 = \begin{cases} \frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n-1} & for\ raw\ data \\ \frac{\sum_{i=1}^{k} f_i(X_i-\bar{x})^2}{n-1} & for\ frequency\ distribution \end{cases}$$

- Using the short-cut formula, we can compute the sample variance as:

$$s^2 = \begin{cases} \frac{\sum_{i=1}^{n} X_i^2 - n\bar{x}^2}{n-1} & for\ raw\ data \\ \frac{\sum_{i=1}^{k} f_i X_i^2 - n\bar{x}^2}{n-1} & for\ frequency\ distribution \end{cases}$$

- The reason for dividing by n-1 rather than n, as we might have expected, is the theoretical consideration referred to as **degrees of freedom**.
- The variance represents squared units and hence, we should return back to the original unit as in the data by taking the square root of the variance to get an appropriate measure of dispersion.
- The square root of the variance is called the **standard deviation**.

The Coefficient of Variation(CV)

- Whenever two samples have the same units of measure, the variance and standard deviation for each sample can be compared directly.
- A statistic that allows us to compare standard deviations when the units are different is called the coefficient of variation.
- It is thus the relative measure of standard deviation.
- The coefficient of variation, denoted by CV, is given by:

$$CV = \frac{s}{\bar{x}} * 100\%.$$

- The distribution having **less C.V** is said to be less variable or more consistent(more homogeneous).

Example: The mean of the waiting times in an emergency room is 80.2 minutes with a standard deviation of 10.5 minutes for people who are admitted for additional treatment. The mean waiting time for patients who are discharged after receiving treatment is 120.6 minutes with a standard deviation of 18.3 minutes. Which times are more variable?

Solution:

Group	$\bar{x}$	**s**	$\mathbf{CV=\frac{s}{\bar{(x)}} * 100\%}$
Additional Treatment	80.2	10.5	13.092%
Discharched	120.6	18.3	15.174%

Thus, The waiting times of patients who are discharged after receiving treatment are more variable since they have large CV.

Standard Scores (Z-scores)

- If X is a measurement from a distribution with mean μ and standard deviation σ, then its value in standard units is:

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

- For the sample with mean $\bar{X}$ and standard deviation S, the standard score is given by:

$$Z = \frac{X - \bar{X}}{S}$$

- Z gives the deviations from the mean in units of standard deviation. That means, it gives the number of standard deviation a particular observation lie above or below the mean.
- It is used to compare two observations coming from different groups.

Example: Two groups of people were trained to perform a certain task and tested to find out which group is faster to learn the task. For the two groups the following information was given:

Value	**Group one**	**Group two**
Mean	10.4 min	11.9 min
Standard deviation	1.2 min	1.3 min

- **Relatively speaking:**
  a. Which group is more consistent in its performance
  b. Suppose a person A from group one take 9.2 minutes while person B from Group two take 9.3 minutes, who was faster in performing the task? Why?

Solution:

a. Use coefficient of variation.

$$CV_1 = \frac{S_1}{\bar{X}_1} * 100\% = \frac{1.2}{10.4} * 100\% = 11.54\%$$

$$CV_2 = \frac{S_2}{\bar{X}_2} * 100\% = \frac{1.3}{10.9} * 100\% = 10.92\%$$

Since $CV_2 < CV_1$, group 2 is more consistent.

b. Calculate the standard score of A and B.

$$Z_A = \frac{X_A - \bar{X}_A}{S_A} = \frac{9.2 - 10.4}{1.2} = -1$$

$$Z_A = \frac{X_A - \bar{X}_A}{S_A} = \frac{9.3 - 11.9}{1.3} = -2$$

- Child B is faster because the time taken by child B is two standard deviation shorter than the average time taken by group 2 while, the time taken by child A is only one standard deviation shorter than the average time taken by group 1.

Skewness and Kurtosis

Skewness

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- It is concerned with the shape of the curve not size.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

$$Mean - Mode = 3(Mean - Median)$$

- Among the various measures of skewness, the most common is based on the pearsonian coefficient of skewness. It can be given by:

$$\alpha_3 = \frac{Mean - Mode}{Standard\ deviation} = \frac{\bar{X} - \hat{X}}{S}$$

Interpretation:

- If $\alpha_3 > 0$, then the distribution is positively skewed.
- If $\alpha_3 = 0$, then the distribution is symmetric
- If $\alpha_3 < 0$, then the distribution is negatively skewed.

Remark:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average. Thus, the distribution will have a longer tail to the right of the central maximum than to the left.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average. In this case, the distribution has a longer tail to the left of the central maximum than to the right.

Kurtosis

- Kurtosis is the degree of peakdness/flattedness of a distribution, usually taken relative to a normal distribution.
- A distribution having relatively high peak is called **leptokurtic**.
- If a curve representing a distribution is flat topped, it is called **platykurtic**.
- The normal distribution which is not very high peaked or flat topped is called **mesokurtic**.
- The most common measure of skewness is based on moments. It is given by:

$$\alpha_4 = \frac{M_4}{M_2^2} = \frac{M_4}{\sigma^2}$$

- M_r, the r^{th} central moment from the mean, is defined as:

$$M_r = \frac{\sum (X_i - \mu)^r}{N}$$

- For the sample data, the r^{th} central moment from the mean could be obtained as:

$$M_r = \frac{\sum (X_i - \bar{X})^r}{n} = \frac{n-1}{n} \frac{\sum (X_i - \bar{X})^r}{n-1}$$

- In particular,

$$M_2 = \frac{n-1}{n} \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{n-1}{n} S^2$$

- **Interpretation:**
  - If $\alpha_4 > 3$, then the distribution is leptokurtic.
  - If $\alpha_4 = 3$, then the distribution is mesokurtic.
  - If $\alpha_4 < 3$, then the distribution is platikurtic. skewed.

- **Probability** as a general concept can be defined as the chance of an event occurring.
- It is the basis or foundation for inferential statistics. For example: predictions and hypothesis testings are based on probability.
- The concept of probability is not foreign to health workers and is frequently encountered in everyday communication.
  - **Example:**
    1. Physician may say that she is 95 percent certain that a patient has a particular disease.
    2. A public health nurse may say that nine times out of ten a certain client will break an appointment.

Basic Concepts:

- A **probability experiment** is a chance process that leads to well-defined results called outcomes.
- An **outcome** is the result of a single trial of a probability experiment.
- A **sample space** is the set of all possible outcomes of a probability experiment.
- An **event** consists of a set of outcomes of a probability experiment.
- An event can hold one outcome or more than one outcome. An event with one outcome is called a **simple event**. Whereas an event with two or more outcomes is called **compound events**.
- **Equally Likely Events/outcomes:** are events/outcomes that have the same chance of occurring.

- **Intersection of two events($A \cap B$)**: is an event containing the common outcomes of the two events (i.e, A and B).
- **Mutually exclusive events:** are events with no common outcomes.
- **Union of two events ($A \cup B$)**: is an event containing all outcomes of A or B or both.
- The **complement of an event A**, often denoted by $\bar{A}$, is the set of outcomes in the sample space that are not included in the outcomes of event A.
- **Independent Events**: Two events are said to be independent if the occurrence of one does not affect the probability of the other occurrence.
- When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be **dependent events**.

Example: Suppose a family planned to have three children. The different possible orderings of boy (B) and girl (G) in the three sequential births will be:

BBB BBG BGB BGG
GBB GBG GGB GGG

Thus,

- $S = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$ is the sample space.
- Let $A = \{BBG, BGB, BGG\}$, $B = \{BBB, BBG, BGB\}$, $C = \{GBG, GGB\}$, $D = \{GGB\}$, and $E = \{\}$ then:
  - D and E are respectively simple and an empty/null events.
  - A, B and C are compound events.
  - $A \cap B = \{BBG, BGB\}$
  - $A \cup B = \{BBG, BGB, BGG, BBB\}$
  - $\bar{A} = \{BBB, GBB, GBG, GGB, GGG\}$
  - A and C are mutually exclusive events.

Counting Techniques

- To compute the probability of an event occurring, we need to know the number of outcomes in that event and also in the sample space.
- There are different counting rules to determine the number of possible outcomes. Some of the counting rules are: multiplication, permutation, and combination.
- We can use a tree diagram to list down the outcomes of the sequence of events.

Multiplication

- Suppose we have k set of elements in which set 1 has n_1 elements, set 2 has n_2 elements,..., set k has n_k element. If we form a sample of k elements by taking one element out of each set, then the number of different samples that can be formed is equal to:

$$n_1 * n_2 * n_3 * ... * n_k$$

- **Example:** On a hospital staff, there are 4 dermatologists, 7 surgeons, 5 general practitioners, 3 psychiatrists, and 3 orthopedic specialists. In how many ways can a hospital administrator form a team by taking one from each field?

$$4 * 7 * 5 * 3 * 3 = \mathbf{1260} \; ways \; of \; forming \; a \; team.$$

Permutation

- It is an arrangement of n objects in a specific **order**.
- The arrangement of n distinct objects in a specific order using r objects at a time is given by:

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

- The number of permutations of n objects in which n_1 are alike, n_2 are alike, ..., n_k are alike is:

$$_nP_r = \frac{n!}{n_1!n_2!...n_k!}$$

Example: In how many ways can a city health department inspector visit 5 restaurants in a city with 10 restaurants?

$$_{10}P_5 = \frac{10!}{(10-5)!} = 30240\ different\ ways$$

Example: How many different permutations can be made from the letters in the word "CORRECTION"?
Here n=10 of which 2 are C, 2 are O, 2 are R, 1 E, 1 T, 1 I, and 1 N. Therefore, there are:

$$_{10}P_7 = 10!/(2!2!2!1!1!1!1!) = 453600\ permutations$$

Partitions

- Suppose there is a single set of N different elements. You want to partition this set into k sets with n_1 elements in set 1, n_2 in set 2, . . . , n_k in set k. The number of different partitions:

$$\frac{N!}{n_1!n_2!...n_k!}$$

Example: Suppose you have 12 system analyst and you want to assign three to job 1, four to job 2 and five to job 3. In how many ways can you assign these people?

$$\frac{12!}{3!4!5!} = 27720$$

Combination

- A selection of distinct objects without regard to order is called a **combination**.
- Combinations are used when the order or arrangement is not important, as in the selecting process.
- The number of combinations of r objects selected from n objects, denoted by ${}_nC_r$, is given by:

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

Example: In how many different ways can a researcher select 5 rats from 20 rats and assign each to a different test?

$$_{20}C_9 = \frac{20!}{9!(20-9)!} = 1,860,480 \ ways$$

Example: A certain study found that out of 10 alcoholic patients, 8 had elevated cholesterol levels, and of 20 nonalcoholic patients, 4 had elevated cholesterol levels. In how many ways of selecting 4 patients for further investigation can be made if:

a. there is no restriction:

b. 2 alcoholic and 2 nonalcoholic patients are selected.

c. exactly 2 patients with elevated cholesterol levels are selected.

d. only patients with elevated cholesterol levels are selected.

Solution:

a. ${}_{30}C_4 = \binom{30}{4} = \dfrac{30!}{4!(30-4)!} = 27405\ ways$

b. $\binom{10}{2}\binom{20}{2} =$

c. $\binom{8+4}{2}\binom{2+16}{2} =$

d. $\binom{8+4}{4}\binom{2+16}{0} =$

Exercise: A new employee has a choice of 5 health care plans, 3 retirement plans, and 2 different expense accounts. If a person selects 1 of each option, how many different options does s/he have?

Type of Probability

1. Classical (Objective) Probability
    - It uses sample spaces to determine the numerical probability that an event will happen.
    - It assumes that all outcomes in the sample space are equally likely to occur.
    - The probability of any event E is:

$$P(E) = \frac{number\ of\ outcomes\ in\ E}{number\ of\ outcomes\ in\ the\ sample\ space} = \frac{n(E)}{n(S)}$$

Example: If a family has three children, find the probability that two of the three children are girls.

Solution The sample space for the gender of the children for a family that has three children has eight outcomes, that is,

BBB, BBG, BGB, GBB, GGG, GGB, GBG, BGG.

Since there are three ways to have two girls, namely, GGB, GBG, and BGG, $\mathbf{P}(\textbf{two girls}) = \mathbf{\frac{3}{8}}$.

2. **Empirical Probability**
   - It relies on **actual experience** to determine the likelihood of outcomes (i.e, based on the relative frequencies or observations).
   - Given a frequency distribution, the probability of an event E being in a given class is:

$$P(E) = \frac{frequency\ of\ the\ class}{total\ frequency\ in\ the\ distribution} = \frac{f}{n}$$

Example: In a sample of 50 people, 21 had type O blood, 22 had type A blood, 5 had type B blood, and 2 had type AB blood. Find the probabilities that a randomly chosen person has type O blood.

Solution: $P(O) = \frac{f}{n} = \dfrac{21}{50}$.

Exercise: Hospital records indicated that knee replacement patients stayed in the hospital for the number of days shown in the distribution.

Number of Days stayed	3	4	5	6	7
Frequency	15	32	56	19	5

- Find the probability that:
  a. A patient stayed exactly 5 days?
  b. A patient stayed less than 6 days?
  c. A patient stayed at most 4 days?

3. **Subjective Probability**
   - It uses a probability value based on an educated guess or estimate, employing opinions and inexact information.
   - **Example:** A physician might say that, on the basis of her diagnosis, there is a **30%** chance the patient will need an operation.

Axioms of Probability

1. **Range of Probability:** For any event A, $0 \leq P(A) \leq 1$.
2. **Probability of a sample space**: $P(S) = 1$.
3. **Probability of an empty event**: $P(\emptyset) = 0$.
4. **Probability of compliment of an event**:
$$P(\bar{A}) = 1 - P(A)$$
5. **Probability of A and B**: $P(A \cap B) = \frac{n(A \cap B)}{n(S)}$
   - For mutually exclusive events A and B: $P(A \cap B) = 0$, since $n(A \cap B) = 0$.
6. **Probability of A or B**:
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
   - For mutually exclusive events A and B:
$$P(A \cup B) = P(A) + P(B), \text{ since } P(A \cap B) = 0.$$

Conditional Probabilities

- The **conditional probability** of an event B in relationship to an event A is the probability that event B occurs after event A has already occurred.
- For any two events, the conditional probability of B given A is:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

  - If A and B are independent, then $P(B/A) = P(A)$.

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

Example: In a certain high school class, consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses. If a student is picked at random from this class,

a. What is the probability that the student wears eyeglasses?

b. What is the probability that the student wears eyeglasses and being a boy?

c. What is the probability that a student wears eyeglasses, given that the student is a boy?

Solution: Let E be an event that the student wears an eyeglass and B be an event that the student is a boy.

a. $P(E) = \frac{number\ of\ students\ wearing\ eyeglass}{total\ number\ of\ students} = \frac{24+16}{60+40} = \mathbf{0.4}$

b. $P(E \cap B) = \frac{n(E \cap B)}{total\ number\ of\ students} = \frac{16}{100} = \mathbf{0.16}$

c. $P(E/B) = \frac{P(E \cap B)}{P(B)} = \frac{16/100}{40/100} = \mathbf{0.4}$

Exercise: In a certain population of hospital patients the probability that a randomly selected patient have heart disease is .35. The probability that a heart diseased patient being a smoker is .86. What is the probability that a patient randomly selected from the population will be a smoker and have heart disease?

- **Note**: When a small sample is selected from a large population and the subjects are not replaced, the probability of the event occurring will changes slightly(i.e, the change in probability will be negligible).

Example: In a certain high-risk group, the chances of a person having suffered a heart attack are 55%. If 6 people are chosen, find the probability that at least 1 will have had a heart attack.

Solution: Let A_i be an event that the i^{th} person will have had a heart attack, then:

- P(A_i)=0.55 for all i=1,2,...,6, since the population is large.
- Thus,
  $$P(A_1 \; or \; A_2 \; or \; ... \; or \; A_6) = 1 - P(A_1^c \; and \; A_2^c \; and \; ... \; and \; A_6^c)$$
  $$= 1 - P(A_1^c)P(A_2^c)...P(A_6^c)$$
  $$= 1 - (P(A_i^c))^6$$
  $$= 1 - (1 - 0.55))^6 = \mathbf{0.9916962}$$

Exercise: A medication is 75% effective against a bacterial infection. Find the probability that if 12 people take the medication, at least 1 persons infection will not improve.

Law of total probability

- Let A_1 and A_2 be a partition of the sample space and E be an event, then:

$$P(E) = P(A_1 \cap E) + P(A_2 \cap E)$$

$$= P(E/A_1)P(A_1) + P(E/A_2)P(A_2)$$

Bayes's rule

- Let A_1 and A_2 be a partition of the sample space and E be an event with $P(E) > 0$, then for i=1,2:

$$P(A_i/E) = \frac{P(A_i \cap E)}{P(E)}$$

$$= \frac{P(E/A_i)P(A_i)}{P(E/A_1)P(A_1) + P(E/A_2)P(A_2)}$$

Medical Application:

- Assume that there is a disease D. Let D^+ (resp. D^-) represent the event that a patient has (resp. has not) the disease D. Suppose there exists a test T to diagnose this disease and we denote a positive (resp. negative) test result by T^+ (resp. T^-). Then,
  - **Sensitivity** $= P(T^+/D^+)$ = the probability that an ill person has a positive test result.
  - **Specificity** $= P(T^-/D^-)$ = the probability that a non-ill person has a negative test result.
  - $P(T^-/D^+)$= the probability of a false negative result
  - $P(T^+/D^-)$= the probability of a false positive result
  - Here, we are interested in the probability that a person is ill when he/she has a positive test result. This is the **predictive positive value**, i.e,

$$P(D^+/T^+) = \frac{P(T^+/D^+)P(D^+)}{P(T^+/D^+)P(D^+) + P(T^+/D^-)P(D^-)}$$

Chapter 6
Probability distributions

6.1 Definition of random variables and probability distributions

- A variable is defined as a characteristic or attribute that can assume different values.
- When the variables are associated with probability(i.e, when the values of a variable can't be predicted in advance), they are called **random variables**.
- Therefore, a random variable is a variable whose values are determined by **chance**.
- Specifically, a random variable is a numerical valued function defined on sample space, usually denoted by capital letters.

  Example: Suppose a family planned to have 3 children. Let X be the number of boys. Using the sample space, the possible values of X are 0, 1, 2, and 3.

- **Probability distribution** is a method used to determine the relationship between the values of a random variable and the probabilities of their occurrence.
- Depending on the type random variables, we can classify probability distributions as **discrete** and **continuous probability distribution**.
- The **discrete probability distribution** is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.
- If we denote the discrete probability distribution by $p(x) = p(X = x)$, then:
  1. $p(x) \geq 0$, for all x.
  2. $\sum_x p(X = x) = 1$, for all x.

- A **continuous probability distribution**(or sometimes called **density function**) is the probability distribution of a continuous random variable.
- If we represent the continuous probability distribution by $f(x)$, then:
    1. $f(x) \geq 0$, for all x.
    2. $\int_x f(x)dx = 1$, for all x.
- For a continuous random variable, probability means **area** under the curve. Thus, the probability of any specific value of the random variable is **zero**.

Note:

- If X is a continuous random variable then:

  $P(a < X < b) = \int_a^b f(x)dx$

  Note: $P(a < X < b) = P(a < X \leq b) =$

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

- If X is a discrete random variable then:

  $P(a < X < b) = \sum_{x=a+1}^{b-1} P(x)$

  $P(a \leq X < b) = \sum_{x=a}^{b-1} P(x)$

  $P(a < X \leq b) = \sum_{x=a+1}^{b} P(x)$

  $P(a \leq X \leq b) = \sum_{x=a}^{b} P(x)$

Example: A fair coin is tossed three times. Let X be a random variable representing the number of heads appears from the 3 tosses.

i. Construct the probability distribution of X.
ii. Find the probability that exactly 2 heads will appear.
iii. Find the probability that at least 1 heads will appear.

Example: For the r.v. X with p.d.f.:

$$f(X) = \begin{cases} kx & for\ 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

i. Determine the value of k.
ii. Compute $P(1 < X < \frac{1}{2})$

Expectation

- The expectation or mean value of the random variable X is denoted by $E(X)$ and is defined by:

$$E(X) = \begin{cases} \sum_x xP(x) & when\ X\ is\ discrete \\ \int_x xf(x)dx & when\ X\ is\ continous \end{cases}$$

- Further more, the expectation or mean value of the function g(.) of the random variable X, denoted by $E(g(X))$, is defined by:

$$E(X) = \begin{cases} \sum_x g(x)P(x) & when\ X\ is\ discrete \\ \int_x g(x)f(x)dx & when\ X\ is\ continous \end{cases}$$

- Let X and Y are random variables and k be a constant, then expectation will have the following properties.
  - $E(k) = k$
  - $E(kX) = kE(X)$
  - $E(X + Y) = E(X) + E(Y)$

  Mean and Variance of a random variable

  Mean: $\mu = E(X)$.

  Variance: $\sigma^2 = V(X) = E(X - E(X))^2$

  $$= E(X^2) - [E(X)]^2$$

  Note: Let X be random variable and k be a constant. Then:
  - $Var(k) = 0$
  - $Var(kX) = k^2Var(X)$

Common Discrete Probability Distributions
Binomial Distribution

- Many types of probability problems have only two possible outcomes or can be reduced to two outcomes.

  Example:

  - A medical treatment can be classified as effective or ineffective, depending on the results.
  - A person can be classified as having normal or abnormal blood pressure, depending on the measure of the blood pressure gauge.

- When a random process or experiment can result in only one of two mutually exclusive outcomes, the trial is called a Bernoulli trial.

- A sequence of Bernoulli trials forms a Bernoulli process or a binomial experiment.

- A binomial experiment is a probability experiment that satisfies the following four requirements:
  1. There must be a fixed number of trials.
  2. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. The outcomes are usually classified as successes or failures.
  3. The outcomes of each trial must be independent of one another.
  4. The probability of a success must remain the same for each trial.
- The outcomes of a binomial experiment and the corresponding probabilities of these outcomes are called a **binomial distribution**.

- In a binomial experiment, the probability of exactly x successes in n trials is:

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

where p=the probability of success and

$q = 1 - p$=the probability of failure.

- The mean, and variance of a variable that has the binomial distribution can be computed by using:

Mean: $E(X) = np$

Variance: $V(X) = npq$

Note: Strictly speaking, the binomial distribution is applicable in situations where sampling is from an infinite population or from a finite population with replacement.

Example: Suppose it is known that 10 percent of a certain population is color blind. If a random sample of 4 people is drawn from this population, find the probability that:

(a) at most two people will be color blind.
(b) at least two people will be color blind.
(c) none of them will be color blind.

Poisson Distribution

- This distribution has been used extensively as a probability model in biology and medicine.
- A random variable X is said to have a Poisson distribution if its probability distribution is given by:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{n!}$$

where λ=average number per unit time.

- If X is a Poisson random variable with parameters λ then:
  Mean: $E(X) = \lambda$
  Variance: $V(X) = \lambda$
- The Poisson distribution is used as a distribution of rare events, such as:
  - Number of misprints.
  - Natural disasters like earth quake.
  - Accidents.
  - Hereditary.
  - Arrivals

  Example: In the study of a certain aquatic organism, a large number of samples were taken from a pond, and the number of organisms in each sample was counted. The average number of organisms per sample was found to be two. Assuming that the number of organisms follows a Poisson distribution, find the probability that the next sample taken will contain:

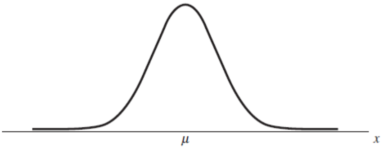
  (a) one or fewer organisms
  (a) exactly three organisms

Common Continuous Probability Distributions
Normal Distribution

- It is the most important distribution in all of statistics.
- A random variable X is said to have a normal distribution if its probability density function is given by:

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

  where $\mu = E(X) - mean$ and $\sigma^2 = E(X-\mu)^2 - variance$ are the two parameters of the distribution.
- The graph of the normal distribution produces the familiar bell-shaped curve like as follows.



Properties of Normal Distribution

1. It is symmetrical about its mean, μ.
2. The mean, the median, and the mode are all equal.
3. The total area under the curve above the x-axis is one square unit.
4. The normal distribution is completely determined by the parameters μ and σ.

   Note: μ is often referred to as a location parameter and σ is often referred to as a shape parameter.

- The normal distribution is really a family of distributions in which one member is distinguished from another on the basis of the values of μ and σ.
- The most important member of this family is the standard normal distribution or unit normal distribution.
- It may be obtained from the above distribution by creating a random variable $z = (x - \mu)/\sigma$.
- The equation for the standard normal distribution is written:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2} z^2}$$

- The standard normal distribution is sometimes called unit normal distribution because it has a mean of 0 and a standard deviation of 1.

- It is useful to facilitate the computations in the examples and applications that follow.
- Using the standard normal, to find the area between z_0 and z_1 directly, we would need to evaluate the following integral:

$$\int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2}$$

- However, there is no a closed-form solution for the integral, and hence a numerical methods of calculus can be used to approximate the desired areas.
- Fortunately, we do not have to concern ourselves with such matters as there are tables available that provide the results of any integration in which we might be interested.

- Areas under the standard normal distribution curve have been tabulated in various ways.
- The most common ones are the areas between 0 and positive value of z.
- Given a normally distributed random variable X with Mean μ and standard deviation σ, then:

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

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

Examples

1. Given the standard normal distribution, find the area under the curve:
   a) between z=0 and z=1.43.
   b) between z=-1.43 and z=0.
   c) to the right of -0.55.
   d) to the left of -0.55.
   e) between z=-1.43 and z=0.75.
2. Given the following probabilities, find z_1:
   a) $P(Z < z_1) = .0055$
   b) $P(Z > z_1) = .0384$
   c) $P(-2.67 < Z < z_1) = .9718$
   d) $P(z_1 < Z < 2.982) = .1117$
   e) $P(-z_1 < Z < z_1) = .8132$

Sampling

- Most researchers come to a conclusion of their study by studying a small sample from the huge population or universe.
- To draw conclusions about population from sample, there are two major requirements for a sample:
  1. the sample size should be adequately large.
  2. the sample has to be selected appropriately so that it will be a representative of the population.
- **Sampling techniques** is concerned with the selection of representative sample, especially for the purposes of statistical inference.

- **Some definitions:**
  - **Target population** (reference population): is the population about which an investigator wishes to draw a conclusion.
  - **Sampled population** (population sampled): a population from which the actual sample was drawn and about which a conclusion can be made.
  - **Sampling unit:** the ultimate unit to be sampled or elements of the population to be sampled.
  - **Sampling frame:** is the list of all elements in a population.
  - **Sampling errors:** are errors arising due to drawing inferences about the population on the basis of few observations. Thus, it is the discrepancy between the population value and sample value.
    - It involved in the collection, processing and analysis of a data.
    - It may arise due to inappropriate sampling techniques.

- **Non Sampling errors**: are errors that arise at the stages of observations, compilation and analysis of data.
  - It can happen in both sample surveys as well as complete population enumeration survey. Thus, the sample survey would be subject to both the sampling errors as well as non-sampling errors.
  - Non sampling errors occur at every stage of planning and execution because of faulty planning, errors in response by the respondents, compilation errors etc.

Reasons for Sampling:

1. Reduced cost; Greater speed; Greater accuracy
2. Greater scope
3. Avoids destructive test

Sometimes taking a census makes more sense than using a sample. Some of the reasons include:

- Universality; Qualitativeness
- Detailedness; Non-representativeness

Methods of sampling/Sampling techniques

- Sampling can be classified into two categories, namely, *probability sampling* and *non-probability sampling*.
- **Probability sampling**: is a method of sampling in which all elements in the population have a pre-assigned non zero probability to be included in to the sample. That is, sampling units are selected on the basis of chance.
- **Non probability sampling**: is a sampling technique in which the choice of individuals for a sample depends on the basis of convenience, personal choice or interest.
- The most common examples of probability sampling include Simple random sampling, stratified random sampling, cluster sampling,systematic sampling and multistage sampling. However, Judgment sampling, Convenience sampling and Quota Sampling are some examples for non probability sampling.

Probability Sampling

- **Simple random sampling(SRS)**
  - In simple random sampling, each unit in the population has equal chance or probability to be selected in the sample.
  - There are two types: SRS with replacement and SRS without replacement.
  - In SRS with replacement, the selected unit is replaced back to the population and again has the chance of getting selected.Whereas in SRS without replacement, which is the usual method in medical research, the selected unit is not put back in the population and hence the population size reduces by one at each selection.
  - Random samples can be drawn by lottery method or by using random number tables(Reading Assignment).
  - It is applied when the population is homogeneous.

- **Stratified random sampling**
  - It is preferred when the population is heterogeneous with respect to characteristic under study.
  - In this method, the complete population is divided into homogenous sub groups called "Strata" and then a stratified sample is obtained by independently selecting a separate simple random sample from each population stratum.
  - Some of the criteria for dividing a population into strata are: Sex (male, female); Age (under 18, 18 to 28, 29 to 39); Occupation (blue-collar, professional, other).
  - Random samples taken within a stratum will have much less variability than a random sample taken across all strata. This is true because sample units within each stratum tend to have characteristics that are similar.

- **Systematic Random Sampling**
  - Systematic sampling is a commonly employed technique, when complete and up to date list of sampling units is available.
  - A systematic random sample is obtained by selecting one unit on a random basis and then choosing additional units at evenly spaced intervals until the desired number of sample size is obtained.
  - Let N=population size; n=sample size and $k = N/n$ is sampling interval.

    Then choose randomly a number between 1 and k. Suppose the randomly chosen number is j $(1 \leq j \leq k)$.
  - The j^{th} unit is selected at first and then $(j+k)^{th}$, $(j+2k)^{th}$, $(j+3k)^{th}$...,etc until the required sample size is reached.

- **Cluster sampling**
  - It is obtained by selecting clusters from the population on the basis of simple random sampling so that each and every units in the selected clusters will be included in the sample.
  - Clusters are formed by grouping units on the basis of their geographical locations.Thus, elements **within a cluster** are heterogeneous.
  - The advantage of cluster sampling is that sampling frame is not required and in practice when complete lists are rarely available, cluster sampling is suitable.
- **Multistage Sampling**
  - In this method, the whole population is divided in first stage sampling units from which a random sample is selected.
  - The selected first stage is then subdivided into second stage units from which another sample is selected. Third and fourth stage sampling is done in the same manner if necessary. For example, in an urban survey in a state, a sample of towns may be taken first and then in each of the selected towns, a second stage sample of households may be taken.

Sampling distribution

- The distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the **sampling distribution** of that statistic.
- Sampling distributions may be constructed empirically when sampling from a discrete, finite population.
- To construct a sampling distribution we proceed as follows:
    - From a finite population of size N, randomly draw all possible samples of size n.
    - Compute the statistic of interest for each sample.
    - List in one column the different distinct observed values of the statistic, and in another column list the corresponding frequency/probability of occurrence of each distinct observed value of the statistic.

- There are commonly three properties of interest for a given sampling distribution.
  - Its Mean
  - Its Variance
  - Its Functional form.
- **Example:** Suppose we have a population of size N = 5, consisting of the age of five children: 6, 8, 10, 12, and 14.

$$Population\ mean = \mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{(6+8+10+12+14)}{5} = 10$$

$Population\ variance = \sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = 8$
Take samples of size 2 with **replacement** and construct sampling distribution of the sample mean.

Solution: Since the sampling is with replacement, there is $N^n = 5^2$ possible ways of getting a sample of size 2. Thus, the possible samples and the corresponding sample means are presented in open and closed braces respectively as follows.

	6	8	10	12	14
6	(6,6) [6]	(6,8)[7]	(6,10)[8]	(6,12)[9]	(6,14[10])
8	(8,6)[7]	(8,8)[8]	(8,10)[9]	(8,12)[10]	(8,14)[11]
10	(10,6)[8]	(10,8)[9]	(10,10)[10]	(10,12)[11]	(10,14)[12]
12	(12,6)[9]	(12,8)[10]	(12,10)[11]	(12,12)[12]	(12,14)[13]
14	(14,6)[10]	(14,8)[11]	(14,10)[12]	(14,12)[13]	(14,14)[14]

Therefore, the sampling distribution of the mean will be constructed by listing the different values in one column and their probability/frequency of occurrence like as follows.

Sample mean($\bar{X}$)	6	7	8	9	10	11	12	13	14
P($\bar{X} = \bar{x}$)	1/25	2/25	3/25	4/25	5/25	4/25	3/25	2/25	1/25

- We are usually interested in the functional form of a sampling distribution, its mean, and its variance. To illustrate these characteristics, lets we again consider the sampling distribution of the sample mean($\bar{X}$).
  - **Mean**:

$$E(\bar{X}) = \sum P(\bar{X}_i = \bar{x}_i)\bar{X}_i = 1/25{*}6{+}2/25{*}7{+}...{+}1/25{*}14 = 10$$

$$=> \mu_{\bar{X}} = E(\bar{X}) = \mu$$

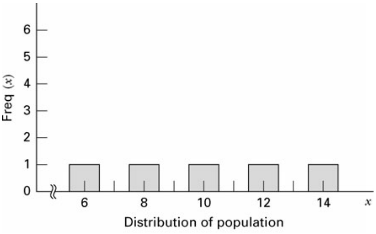
  - **Variance**:

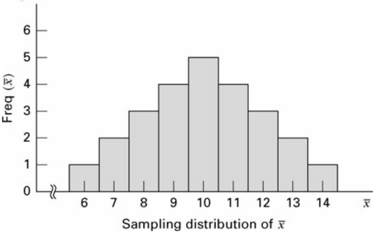
$$Var(\bar{X}) = E(\bar{X}-\mu)^2 = \sum P(\bar{X}_i = \bar{x}_i)(\bar{X}_i-\mu)^2 = 1/25{*}(6-10)^2+$$

$$2/25 * (7-10)^2 + ... + 1/25 * (14-10)^2 = 4 = 8/2$$

$$=> Var(\bar{X}) = \sigma^2/n$$

  - **Functional form:** the distribution of the sample mean plotted as a histogram, along with the distribution of the population, both of which are shown as follows.





From the plots we can observe that the parent population is uniformly distributed, while the sampling distribution of the mean gradually rises to a peak and then drops off with perfect symmetry.

Remarks:

- In any case (i.e, sampling with and without replacement), the sample mean is unbiased estimator of the population mean.

$$=> E(\bar{X}) = \mu$$

- For sampling with replacement: $Var(\bar{X}) = \sigma^2/n$
- For sampling without replacement: $Var(\bar{X}) = \frac{\sigma^2}{n} * \dfrac{N-n}{N-1}$

  Note:

  - The square root of the variance of the sampling distribution is called the standard error of the mean or, simply, the standard error.
  - When sampling is from an infinite population, the standard errors under both sampling with and without replacement will close each other.

- When sampling is from a normally distributed population, the distribution of the sample mean will also be normal with mean μ and variance σ^2/n. That means,

$$X \sim N(\mu, \sigma^2) => \bar{X} \sim N(\mu, \sigma^2/n)$$

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

  - **Example:** If the uric acid values in normal adult males are approximately normally distributed with mean 5.7 mgs and standard deviation 1mg, find the probability that a sample of size 9 will yield a mean: (1) greater that 6; (2) between 5 and 6 (3) less than 5.2.

    Solution: Let X is the amount of uric acid in normal adult males. Thus, $X \sim N(\mu, \sigma^2) = N(5.7, 1)$.

1. $P(\bar{X} > 6) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{6 - 5.7}{1/\sqrt{9}}) = P(Z > 0.9) = \mathbf{0.1841}$
2. $P(5 < \bar{X} < 6) = P(\frac{5 - 5.7}{1/\sqrt{9}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{6 - 5.7}{1/\sqrt{9}}) =$ $P(-2.1 < Z < 0.9) = \mathbf{0.7981}$
3. $P(\bar{X} > 5.2) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{5.2 - 5.7}{1/\sqrt{9}}) = P(Z < -1.5) =$ $\mathbf{0.0668}$

Central Limit Theorem

- Given a population of any nonnormal functional form with a mean and finite variance, the sampling distribution of $\bar{X}$ computed from samples of size n from this population, will have mean μ and variance σ^2/n and will be approximately normally distributed when the **sample size is large**. i.e,

$$If \;\; n \; is \; large \;\; then \;\; \bar{X} \sim N(\mu, \sigma^2/n)$$

- **Example:** If the mean and standard deviation of serum iron values for healthy men are 120 and 15 micro-grams per 100 ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 micro-grams per 100 ml?
  Solution: The functional form of the population of serum iron values is not specified, but since we have a sample size greater than 30, we make use of the central limit theorem.
  Thus, $P(115 < \bar{X} < 125) = P(\frac{115-120}{1/\sqrt{50}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{125-120}{1/\sqrt{50}}) = P(-2.36 < Z < 2.36) = \mathbf{0.9818}$

Chapter 8
Inferential Statistics

- Inference is the process of making interpretations or drawing conclusions from sample data to the population.
- Any researcher collects information with the aim to draw valid conclusions regarding the research question.
- In statistics, inference can be done in two ways:
  - Statistical estimation
  - Statistical hypothesis testing.

Statistical Estimation

- **Estimation**: is one way of making inference about the population parameter where the investigator does not have any prior notion about values or characteristics of the population parameter.

- There are two methods of making estimation: *Point Estimation* and *Interval Estimation*.
- **Point Estimation**: is a procedure that results in a single value as an estimate for a parameter.
- **Interval estimation**: is the procedure that results in the interval of values as an estimate for a parameter.
  - It deals with identifying the upper and lower limits of a parameter.
- **Estimator Vs Estimate**
  - **Estimator**: is the rule or random variable that helps us to approximate a population parameter.
  - **Estimate**: is the different possible values in which an estimator can assume. Thus, A point estimate is a specific numerical value estimate of a parameter.

- Properties of a Good Estimator:
  - The estimator should be an *unbiased* estimator. That is, the expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.
  - The estimator should be *consistent*. For a consistent estimator, as sample size increases, the value of the estimator approaches the value of the parameter estimated.
  - The estimator should be a *relatively efficient* estimator. That is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance.
- **Point estimation of the population mean**: μ
  - The *sample mean* is a better estimator of the *population mean* than the sample median or sample mode. That is, $(\bar{X}) = \dfrac{\sum_{i=1}^{n} X_i}{n}$ is a point estimator of the population mean μ.

- **Example:** An investigator is interested in finding out the mean duration of hospital stay by patients undergoing cesarean section. Ideally the investigator should go through the case details of all patients who have undergone cesarean section. But the investigator decides to examine a sample of these patients from which he computes the average duration of hospital stay.
- **Interval Estimation**
  - Due to sampling error, there may be some doubt on the accuracy of point estimates as there is no way of knowing how close a particular point estimate is to the population mean.
  - Consequently, statisticians prefer another type of estimate, called an *interval estimate*.
  - **An interval estimate** of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.

- The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.
- A **confidence interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.
- **Interval Estimation of Mean**
  - To calculate confidence interval we make use of the knowledge of sampling distributions.
  - **Assumption:** Either the population is normally distributed or $n >= 30$.

- In constructing confidence interval, knowledge of population variance (σ^2) from which the sample is taken is required.
- **Case 1:** σ^2 is known
  - A $100(1-\alpha)\%$ confidence interval for mean:

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

  - **Example:** A dentist wished to estimate with 95% confidence, the mean marginal displacement in the teeth taking place by applying a particular treatment modality. He assumes that the marginal displacement values are normally distributed with a mean of 6.2 units after studying 100 people. The population variance is 9 units.
    Solution: $\bar{x} = 6.2$, $\sigma = \sqrt{9} = 3$, n=100 and $z_{\dfrac{\alpha}{2}} = 1.96$.

    Thus, 95% confidence interval for μ is:

$$\bar{X} \mp Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 6.2 \mp 1.96 * \frac{3}{\sqrt{100}} = \mathbf{(5.61,\ 6.79)}$$

Interpretation: In repeated sampling from the study population, we are 95% confident that the mean marginal displacement for population lies between 5.61 and 6.79.

- **Case 2:** σ^2 is unknown
  - When sample size is large (i.e. $n > 30$), we use sample standard deviation as a replacement for the unknown population standard deviation even if it is from non normal distribution by virtue of central limit theorem. That is,

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

Example: Suppose a researcher interested in finding the serum TSH in healthy adult females; studied 100 subjects and found that the mean serum TSH was 2 units with a standard deviation (SD) of 0.2. Calculate 95% confidence interval.
Solution: $\bar{x} = 2$, $s = 0.2$, n=100 and $z_{\overline{\frac{\alpha}{2}}} = 1.96$.

Thus, a 95% confidence interval for population mean is given by:

$$\bar{X} \mp Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 2 \mp 1.96 * \frac{0.2}{\sqrt{100}} = \mathbf{(1.9608\ ,\ 2.0392)}$$

Interpretation: We are 95% confident that the reality or truth that exists in the total population (of millions of adult healthy females) would be that mean serum TSH would be between 1.9608 to 2.0392.

- When sample size is small but the population is normal, less than 30, the procedure remains same except that we use t distribution with n-1 degrees of freedom instead of standard normal distribution z.
  The confidence interval for population mean in such cases is given by:

$$\bar{X} - t_{\frac{\alpha}{2};(n-1)} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2};(n-1)} \frac{S}{\sqrt{n}}$$

- **Note:** If the population is not normal and the sample size is small (i.e, n< 30), then we can not construct the interval due to the fact that the assumption is not met.

Confidence Interval for Population Proportion

- Many times the researcher is interested not in the mean value but a proportion value.
- Fore example, the proportion of patients who survive/recover after administrating a particular drug or what is the proportion of side effects and so on.
- For large sample size, the sampling distribution of proportion is normal distribution with standard error as $\sqrt{\frac{p(1-p)}{n}}$.
- Thus, a 100(1-α)% CI for the population proportion π will be:

$$p \mp Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Example: A researcher is interested in knowing the proportion of young men who are seropositive for HIV infection. He took a sample of, say, 1000 young men and did a sero study. He found that out of 1000 young men, 5 were positive. Calculate the 95 confidence interval for population proportion. **Solution:** The sample proportion value = p= 5/1000 = 0.005; thus, q = 1-p = (1-0.005) = 0.995, and n = total sample = 1000. The 95% Confidence interval for population proportion is given by:

$$p \mp Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Thus, 95% CI= 0.005 $(1.96 * \sqrt{0.005(1-0.005)/10000})$
$= 0.005 \mp 0.0044 = (0.0006\ ,\ 0.0094)$

- Our interpretation would be that though we are not aware of what exactly the real seropositivity in the entire province would be, we are 95% sure that whatever this reality is, it will be somewhere between 6 per 10,000 to 94 per 10,000 in the total population.

Hypothesis testing

- **Hypothesis testing** is also one way of making inference about population parameter, where the investigator has *prior notion* about the value of the parameter.
- The researcher states a **hypothesis** to be tested, formulates an analysis plan, analyzes sample data according to the plan, and accepts or rejects the hypothesis, based on results of the analysis.
- A statistical hypothesis is a conjecture about a population parameter. This conjecture may or may not be true.
- There are two types of statistical hypotheses: *null hypothesis* and *alternative hypothesis*.
- **Null hypothesis (H_0)**
    - It is the hypothesis to be tested.
    - It is the hypothesis that often states "there is *no difference* between a parameter and a specific value", or that "there is *no difference* between two parameters".

- **Alternative hypothesis** (H_1 *or* H_a)
  - It is the hypothesis available when the null hypothesis has to be rejected.
  - It is the hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters.
- A **statistical test** uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected.
- The numerical value obtained from a statistical test is called the **test statistic** value.

- **Example:** Suppose we are interested to study the effect of a new drug in reducing cholesterol levels.
    - The research question is formally converted into a formal scientific hypothesis, which has two parts: the null hypothesis and the alternative hypothesis.
    - In the settings where two treatments (new drug and placebo) are administered to two different samples, the null hypothesis would be there is no difference between cholesterol levels in the two groups i.e. ”Persons treated with new drug will have same cholesterol levels as persons not treated with new drug”.
    - If the null hypothesis gets rejected then the hypothesis that gets accepted is called ”Alternate hypothesis”.
    - Thus, the alternate hypothesis would be phrased as, Persons treated with a new drug have different (higher or lower) cholesterol levels than persons not treated with new drug.

Types and size of errors:

- In reality, the null hypothesis **may or may not** be true, and a decision made to reject or not reject it is on the basis of the sample data which may involve sampling and non sampling errors.
- In hypothesis-testing, there are four possible outcomes as shown below:

		Decision	
		Reject H_0	**Don't Reject H_0**
Truth	H_0	**Type I Error**	Correct Decision
	H_1	Correct Decision	**Type II Error**

  - A **type I error** occurs when you reject the **true** null hypothesis.
  - A **type II error** occurs when you fail to reject the **false** null hypothesis.

Exercise: For each of the following situations, identify the type I and type II errors and the correct actions.

H_0: "A new treatment is not more effective than the traditional one".

- Adopt the new treatment when the new one is more effective.
- Continue with the traditional treatment when the new one is more effective.
- Continue with the traditional treatment when the new one is not more effective.
- Adopt the new treatment when the new one is not more effective.

- The **level of significance**, denoted by α, is the maximum probability of committing a type I error. i.e,
$$P(Type\ I\ error) \leq \alpha.$$
- The probability of making type II error often denoted by β. i.e, $P(Type\ II\ error) = \beta$.
- It is natural to aim first for a test whose type I and type II error probabilities are minimum. However, Type I error and Type II error have an inverse relationship and therefore, can not be minimized at the same time.
- The **most powerful test** is a test that fixes the level of significance and minimizes the probability of type II error.
- **Power of a test** is defined as the probability of rejecting the null hypothesis when it is actually false. i.e,
$$power = 1 - \beta$$

- **Note:** Type I error is often considered to be more serious, and therefore more important to avoid, than a type II error.
- **Steps in hypothesis testing:**
  - State hypotheses.
  - Select test statistics.
  - Determine distribution of test statistics.
  - State decision rule or obtain the critical/table value.
  - Calculate test statistics
  - Make statistical decision (Reject or do not reject H_0) by comparing test statistic value and critical value.
  - Draw conclusion

Testing a single population mean(μ):

- **Hypothesis:**

$$H_0 : \mu = \mu_0 \text{ Vs } \begin{array}{ll} H_1 : \mu \neq \mu_0 & (1) \\ H_1 : \mu > \mu_0 & (2) \\ H_1 : \mu < \mu_0 & (3) \end{array}$$

  - **Note:** (1) is for a two sided test whereas (2) and (3) are for a one sided test.

- Here we consider two situations about a population mean:
  - When population variance is known (σ^2 is known)
  - When population variance is unknown.

  Situation 1: When the population variance is known, the test statistic would be:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1)$$

  Note: The sample size should be large (i.e, $n \geq 30$) when sampling is not from a normally distributed population.

The decision rule will depends on the type of test (i.e, one sided or two sided), and is summarized as follows.

	Reject H_0	Do not reject H_0	Inconclusive
$H_1 : \mu \neq \mu_0$	$\lvert z_{cal} \rvert > Z_{\alpha/2}$	$\lvert z_{cal} \rvert < Z_{\alpha/2}$	$\lvert z_{cal} \rvert = Z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$z_{cal} > Z_\alpha$	$z_{cal} < Z_\alpha$	$z_{cal} = Z_\alpha$
$H_1 : \mu < \mu_0$	$z_{cal} < -Z_\alpha$	$z_{cal} > -Z_\alpha$	$z_{cal} = -Z_\alpha$

Situation 2: When the population variance is unknown
In this case, we use the sample standard deviation (s) as an estimate of the population variance (σ^2). However, this will adds another element of uncertainty to our inference.

To account the additional uncertainty that comes from estimating the population variance, we use a modification of Z called t-distribution.

Note: t distributions are similar to z distribution, but have broader tails and less peaked at the center. As n increases, t distribution approaches normal distribution.

Thus, the test statistic would be:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_(n-1)$$

The decision rule will also again depends on the type of test like summarized as follows.

	Reject H_0	Do not reject H_0	Inconclusive
$H_1 : \mu \neq \mu_0$	$\lvert t_{cal} \rvert > t_{\alpha/2;n-1}$	$\lvert t_{cal} \rvert < t_{\alpha/2;n-1}$	$\lvert t_{cal} \rvert = t_{\alpha/2}$
$H_1 : \mu > \mu_0$	$t_{cal} > t_{\alpha;n-1}$	$t_{cal} < t_{\alpha;n-1}$	$t_{cal} = t_{\alpha;n-1}$
$H_1 : \mu < \mu_0$	$t_{cal} < -t_{\alpha;n-1}$	$t_{cal} > -t_{\alpha;n-1}$	$t_{cal} = -t_{\alpha;n-1}$

Note: When the sample size is large (i.e, $n > 30$),

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)} \approx N(0,1)$$

- **Example**: Researchers claim that the mean age of population having a certain disease A is 35 years. To prove their claim, a researcher collected information from a random sample of 20 individuals drawn from a normally distributed population. Population variance is known and is equal to 25 and the study found that the mean age of 20 individuals is as 29. Test the claim at 5% level of significance.

  Solution

  - **Hypothesis:** $H_0 : \mu = 35 \; Vs \; H_1 : \mu \neq 35$, (i.e, $\mu_0 = 35$).
  - **Test Statistic :** Since population variance is known, our statistic will be given by: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Thus, the test statistic value will be:

$$z_{cal} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{(29 - 35)}{5/\sqrt{20}} = -5.36$$

-
  - **Critical/Table value**: $Z_{\alpha/2} = 1.96$
  - **Decision:** Since $|Z_{cal}| = 5.36 > 1.96$, reject H_0.
  - **Conclusion:** We conclude that the mean age of the population with a specific disease ”A” is not equal to 35 years.
- **Exercise:** A research team is willing to assume that systolic blood pressures in a certain population of males are approximately normally distributed with a standard deviation of 16. A simple random sample of 64 males from the population had a mean systolic blood pressure reading of 133. At the .05 level of significance, do these data provide sufficient evidence for us to conclude that the population mean is greater than 130?

- **Example:** A study was made of a sample of 25 records of patients seen at a chronic disease hospital on an outpatient basis. The mean number of outpatient visits per patient was 4.8, and the sample standard deviation was 2. Can it be concluded from these data that the population mean is greater than four visits per patient? Let the probability of committing a type I error be .05. What assumptions are necessary?

  Solution:

  - **Assumption:** Since the sample size is small, it is necessary to assume the samples are drawn from a normal population.
  - **Hypothesis**: $H_0 : \mu = 4 \; Vs \; H_1 : \mu \neq 4$, (i.e, $\mu_0 > 4$)
  - **Test statistic Value:**

$$t_{cal} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{(4.8 - 4)}{2/\sqrt{25}} = 2$$

- **Table value**: $t_{\alpha;n-1} = t_{0.05,24}$=1.71
- **Decision:** Reject H_0 since $t_{cal} > t_{\alpha,n-1}$.
- **Conclusion**: Yes, we can conclude from the data that the population mean greater than four visits per patients.

- **Exercise:** A sample of eight patients admitted to a hospital with a diagnosis of biliary cirrhosis had a mean IgM level of 160.55 units per milliliter. The sample standard deviation was 50. Do these data provide sufficient evidence to indicate that the population mean is greater than 150? Use 5% level of significance. What assumption is required? Determine the p value.

Paired Comparison(paired t-test)

- The t test we have described above deals with situations in which there are two independent samples whose means are to be compared.
- Often in medical research, we may have either only one sample which gives us two sets of readings (before and after readings) or else.
- If the same subjects are used in two sets of readings, the samples are related or dependent.
- To account the dependency or correlation between the two sets of data, we will use "Paired t-test".
-

Assumption: the population(s) must be normally distributed or the sample size(s) are greater than 30.

The statistical procedure in such situations will be given as follows.

- **Hypotheses:**

$$H_0 : \mu_d = 0 \text{ Vs } \begin{array}{ll} H_1 : \mu_d \neq 0 & (1) \\ H_1 : \mu_d > 0 & (2) \\ H_1 : \mu_d < 0 & (3) \end{array}$$

- **Test statistic:** The test statistic for testing $H_0 : \mu_d = 0$ (i.e, the mean difference between the before and after value $= 0$) would be:

$$T = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} \sim t_{n-1}$$

  where, $d_i = x_{1i} - x_{2i};\ \bar{d} = \sum d_i/n;\ S_d = \sqrt{\frac{\sum(d_i - \bar{d})}{n-1}}$

- The decision will be the same as in testing one population mean.

Example: To study the efficacy of a drug in reducing the Serum Cholesterol level, we took 10 healthy adult males and measured their Serum Cholesterol levels. These 10 subjects are then given the drug for 1 month and the Serum Cholesterol levels were again measured thereafter. The Serum Cholesterol levels of the 10 subjects before and after the drug are given as follows. Test the efficacy of the drug at 5% level of significant.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Before	306	254	198	242	236	228	286	274	208	188
After	280	242	204	238	228	202	264	258	209	198
d_i	26	12	-6	4	8	26	22	16	-1	-10

Solution:
Assumption: The observed differences constitute a random sample from a normally distributed population

- **Hypotheses:** $H_0 : \mu_d = 0 \; Vs \; H_1 : \mu_d \neq 0$
- **Test Statistic:** $\bar{d} = 9.7; \; S_d = 12.96; n = 10.$
  Thus, the test statistic value is:

$$T = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} = \frac{9.7 - 0}{12.96/\sqrt{10}} = 2.37$$

- **Decision:** the t-table value at α= 0.05 and df = 9 is 2.26. Since the calculated value is greater than the tabulated one, we conclude that our results are significant i.e. we reject our null hypothesis.
- **Conclusion:** At 5% level of significance, our sample of 10 subjects shows that there is a reduction on an average by 9.7 mg/dl due to the drug.

Inferences with Population Proportion(s).

- In clinical trials one may count the number of times an event occurs such as number of successful outcomes, number of failures or number of patients recovered after administration of drug etc.
- Fore instance, patients in one group may receive new treatment drug and another independent group may receive existing conventional treatment. We may be interested in comparing the proportion of patients attacked by disease after administration of the treatment in the two populations.

Hypothesis Testing : A Single Population Proportion

- **Hypothesis**:

$$H_0 : \pi = \pi_0 \text{ Vs } \begin{array}{ll} H_1 : \pi \neq \pi_0 & (1) \\ H_1 : \pi > \pi_0 & (2) \\ H_1 : \pi < \pi_0 & (3) \end{array}$$

- **Test statistic**
  - In testing a single population proportion denoted by π against a hypothesized value of π_0, approximate normality assumptions holds true if the sample size is large.
  - If the sample size is large, then the test statistic to be used in this procedure will be:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

- The decision will be the same like in testing one population mean.

Example: In clinical studies of an anti-allergy drug, 70 of 781 subjects experienced drowsiness. A competitor claims that 8% of users of his drug experience drowsiness. Use a 0.05 significance level to test this claim.

Solution:

- **Assumption:** The random sample is drawn from a normally distributed population.
- **Hypotheses:** $H_0 : \pi = 0.08 \; Vs \; H_1 : \pi \neq 0.08$
- **Test statistic:** The data obtained on drug says 70 out of 781 subjects experienced drowsiness. Hence, $\frac{70}{781} = 0.089$. Therefore, The test statistic value is:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.089 - 0.08}{\sqrt{\frac{0.08 * (1 - 0.08)}{781}}} = 0.9271$$

- **Decision:** At $\alpha = 0.05$, the standard normal table value is 1.96. Since our calculated test statistic value is less than the table value, we fail to reject the null hypothesis.
- **Conclusion:** There is not sufficient evidence to warrant rejection of the claim that drowsiness will be less among users of the competitors drug.