

1. INTRODUCTION

Definition:

We can define statistics in two ways.

1. Plural sense (lay man definition).

It is an aggregate or collection of numerical facts.

2. Singular sense (formal definition)

Statistics is defined as the science of collecting, organizing, presenting, analyzing and interpreting numerical data for the purpose of assisting in making a more effective decision.

Classifications:

Depending on how data can be used statistics is sometimes divided in to two main areas or branches.

1. **Descriptive Statistics:** is concerned with summary calculations, graphs, charts and tables.

2. **Inferential Statistics:** is a method used to generalize from a sample to a population.

For example, the average income of all families (the population) in Ethiopia can be estimated from figures obtained from a few hundred (the sample) families.

- It is important because statistical data usually arises from sample.
- Statistical techniques based on probability theory are required.

Stages in Statistical Investigation

There are five stages or steps in any statistical investigation.

1. Collection of data: the process of measuring, gathering, assembling the raw data up on which the statistical investigation is to be based.
 - Data can be collected in a variety of ways; one of the most common methods is through the use of survey. Survey can also be done in different methods, three of the most common methods are:
 - Telephone survey
 - Mailed questionnaire
 - Personal interview.

Exercise: discuss the advantage and disadvantage of the above three methods with respect to each other.

2. Organization of data: Summarization of data in some meaningful way, e.g table form
3. Presentation of the data: The process of re-organization, classification, compilation, and summarization of data to present it in a meaningful form.
4. Analysis of data: The process of extracting relevant information from the summarized data, mainly through the use of elementary mathematical operation.
5. Inference of data: The interpretation and further observation of the various statistical measures through the analysis of the data by implementing those methods by which conclusions are formed and inferences made. tatistical techniques based on probability theory are required.

Definitions of some terms

- a. **Statistical Population:** It is the collection of all possible observations of a specified characteristic of interest (possessing certain common property) and being under study. An example is all of the students in AAU 3101 course in this term.
- b. **Sample:** It is a subset of the population, selected using some sampling technique in such a way that they represent the population.
- c. **Sampling:** The process or method of sample selection from the population.
- d. **Sample size:** The number of elements or observation to be included in the sample.
- e. **Census:** Complete enumeration or observation of the elements of the population. Or it is the collection of data from every element in a population
- f. **Parameter:** Characteristic or measure obtained from a population.
- g. **Statistic:** Characteristic or measure obtained from a sample.
- h. **Variable:** It is an item of interest that can take on many different numerical values.

Types of Variables or Data:

1. **Qualitative Variables** are nonnumeric variables and can't be measured. Examples include gender, religious affiliation, and state of birth.
2. **Quantitative Variables** are numerical variables and can be measured. Examples include balance in checking account, number of children in family. Note that quantitative variables are either discrete (which can assume only certain values, and there are usually "gaps" between the values, such as the number of bedrooms in your house) or continuous (which can assume any value within a specific range, such as the air pressure in a tire.)

Applications of statistics:

- In almost all fields of human endeavor.
- Almost all human beings in their daily life are subjected to obtaining numerical facts e.g. about price.
- Applicable in some process e.g. invention of certain drugs, extent of environmental pollution.
- In industries especially in quality control area.

Uses of statistics:

The main function of statistics is to enlarge our knowledge of complex phenomena. The following are some uses of statistics:

1. It presents facts in a definite and precise form.
2. Data reduction.
3. Measuring the magnitude of variations in data.
4. Furnishes a technique of comparison
5. Estimating unknown population characteristics.
6. Testing and formulating of hypothesis.
7. Studying the relationship between two or more variable.
8. Forecasting future events.

Limitations of statistics

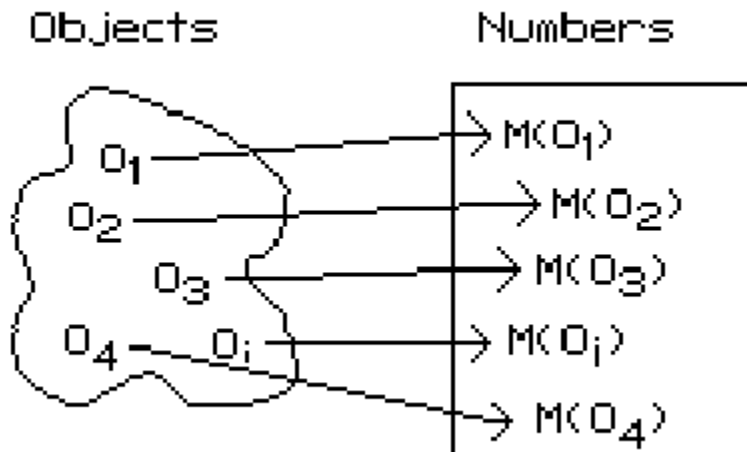
As a science statistics has its own limitations. The following are some of the limitations:

- Deals with only quantitative information.
- 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.

Scales of measurement

Proper knowledge about the nature and type of data to be dealt with is essential in order to specify and apply the proper statistical method for their analysis and inferences.

Measurement scale refers to the property of value assigned to the data based on the properties of order, distance and fixed zero. In mathematical terms measurement is a functional mapping from the set of objects $\{O_i\}$ to the set of real numbers $\{M(O_i)\}$.



The goal of measurement systems is to structure the rule for assigning numbers to objects in such a way that the relationship between the objects is preserved in the numbers assigned to the objects. The different kinds of relationships preserved are called properties of the measurement system.

Order

The property of order exists when an object that has more of the attribute than another object, is given a bigger number by the rule system. This relationship must hold for all objects in the "real world". The property of ORDER exist. When for all i, j if $O_i > O_j$, then $M(O_i) > M(O_j)$.

Distance

The property of distance is concerned with the relationship of differences between objects. If a measurement system possesses the property of distance it means that the unit of measurement means the same thing throughout the scale of numbers. That is, an inch is an inch, no matters were it falls - immediately ahead or a mile downs the road.

More precisely, an equal difference between two numbers reflects an equal difference in the "real world" between the objects that were assigned the numbers. In order to define the property of distance in the mathematical notation, four objects are required: O_i , O_j , O_k , and O_l . The difference between objects is represented by the "-" sign; $O_i - O_j$ refers to the actual "real world" difference between object i and object j , while $M(O_i) - M(O_j)$ refers to differences between numbers. The property of DISTANCE exists, for all i, j, k, l If $O_i - O_j \geq O_k - O_l$ then $M(O_i) - M(O_j) \geq M(O_k) - M(O_l)$.

Fixed Zero

A measurement system possesses a rational zero (fixed zero) if an object that has none of the attribute in question is assigned the number zero by the system of rules. The object does not need to really exist in the "real world", as it is somewhat difficult to visualize a "man with no height". The requirement for a rational zero is this: if objects with none of the attribute did exist would they be given the value zero. Defining O_0 as the object with none of the attribute in question, the definition of a rational zero becomes: The property of FIXED ZERO exists if $M(O_0) = 0$. The property of fixed zero is necessary for ratios between numbers to be meaningful.

SCALE TYPES

Measurement is the assignment of numbers to objects or events in a systematic fashion. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio and each possessed different properties of measurement systems.

Nominal Scales

Nominal scales are measurement systems that possess none of the three properties above.

- Level of measurement which classifies data into mutually exclusive, all-inclusive categories in which no order or ranking can be imposed on the data.
- No arithmetic and relational operation can be applied.

Examples:

- Political party preference (Republican, Democrat, or Other,)
- Sex (Male or Female.)
- Marital status(married, single, widow, divorce)
- Country code
- Regional differentiation of Ethiopia.

Ordinal Scales

Ordinal Scales are measurement systems that possess the property of order, but not the property of distance. The property of fixed zero is not important if the property of distance is not satisfied.

- Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist.
- Arithmetic operations are not applicable but relational operations are applicable.
- Ordering is the sole property of ordinal scale.

Examples:

- Letter grades (A, B, C, D, F).
- Rating scales (Excellent, Very good, Good, Fair, poor).
- Military status.

Interval Scales

Interval scales are measurement systems that possess the properties of Order and distance, but not the property of fixed zero.

- Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.
- All arithmetic operations except division are applicable.
- Relational operations are also possible.

Examples:

- IQ
- Temperature in °F.

Ratio Scales

Ratio scales are measurement systems that possess all three properties: order, distance, and fixed zero. The added power of a fixed zero allows ratios of numbers to be meaningfully interpreted; i.e. the ratio of Bekele's height to Martha's height is 1.32, whereas this is not possible with interval scales.

- Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.
- All arithmetic and relational operations are applicable.

Examples:

- Weight

- Height
- Number of students
- Age

Exercise: classify the different measurement systems into one of the four types of scales.

1. Your checking account number as a name for your account. nominal
2. Your checking account balance as a measure of the amount of money you have in that account. ratio
3. The order in which you were eliminated in a spelling bee as a measure of your spelling ability. ordinal
4. Your score on the first statistics test as a measure of your knowledge of statistics. interval
5. Your score on an individual intelligence test as a measure of your intelligence. interval
6. The distance around your forehead measured with a tape measure as a measure of your intelligence. interval
7. A response to the statement Abortion is a woman's right where, Strongly Disagree = 1, Disagree = 2, No Opinion = 3, Agree = 4, and Strongly Agree = 5, as a measure of attitude toward abortion. nominal
8. Times for swimmers to complete a 50-meter race. ratio
9. Months of the year Meskerm, Tikimit...nominal
10. Socioeconomic status of a family classified as low, middle and upper classes. ordinal
11. Blood type of individuals, A, B, AB and O. nominal
12. Pollen counts provided as numbers between 1 and 10 where 1 implies there is almost no pollen and 10 that it is rampant, but for which the values do not represent an actual counts of grains of pollen. ordinal
13. Regions numbers of Ethiopia (1, 2, 3 etc.) nominal
14. The number of students in a college. ratio
15. the net wages of a group of workers. ratio
16. The height of the men in the same town. Ratio

INTRODUCTION TO METHODS OF DATA COLLECTION

There are two sources of data:

1. Primary Data

- Data measured or collect by the investigator or the user directly from the source.
- Two activities involved: planning and measuring.

a) Planning:

- Identify source and elements of the data.
- Decide whether to consider sample or census.
- If sampling is preferred, decide on sample size, selection method,... etc
- Decide measurement procedure.
- Set up the necessary organizational structure.

b) Measuring: there are different options.

- Focus Group
- Telephone Interview
- Mail Questionnaires
- Door-to-Door Survey
- Mall Intercept
- New Product Registration
- Personal Interview and
- Experiments are some of the sources for collecting the primary data.

2. Secondary Data

- Data gathered or compiled from published and unpublished sources or files.
- When our source is secondary data check that:
 - The type and objective of the situations.
 - The purpose for which the data are collected and compatible with the present problem.
 - The nature and classification of data is appropriate to our problem.
 - There are no biases and misreporting in the published data.

Note: Data which are primary for one may be secondary for the other.

METHODS OF DATA PRESENTATION

Having collected and edited the data, the next important step is to organize it. That is to present it in a readily comprehensible condensed form that aids in order to draw inferences from it. It is also necessary that the like be separated from the unlike ones.

The presentation of data is broadly classified in to the following two categories:

- Tabular presentation
- Diagrammatic and Graphic presentation.

The process of arranging data in to classes or categories according to similarities technically is called classification. Classification is a preliminary and it prepares the ground for proper presentation of data.

- Raw data: recorded information in its original collected form, whether it be counts or measurements, is referred to as raw data.
- Frequency: is the number of values in a specific class of the distribution.
- Frequency distribution: is the organization of raw data in table form using classes and frequencies.

There are three basic types of frequency distributions

- Categorical frequency distribution
- Ungrouped frequency distribution
- Grouped frequency distribution

There are specific procedures for constructing each type.

1. Categorical frequency Distribution:

Data can be place in specific categories such as nominal or ordinal. e.g. marital status.

Example: a social worker collected the following data on marital status for 25 persons.

(M=married, S=single, W=widowed, D=divorced)

M	S	D	W	D
S	S	M	M	M
W	D	S	M	M
W	D	D	S	S
S	W	W	D	D

Solution:

Since the data are categorical, discrete classes can be used. There are four types of marital status M, S, D, and W. These types will be used as class for the distribution. We follow procedure to construct the frequency distribution.

Step 1: Make a table as shown.

Class	Frequency	Percent
(1)	(2)	(3)
M		
S		
D		
W		

Step 2: Find the percentages of values in each class by using;

$$\% = \frac{f}{n} * 100 \quad \text{Where } f = \text{frequency of the class, } n = \text{total number of value.}$$

Percentages are not normally a part of frequency distribution but they can be added since they are used in certain types diagrammatic such as pie charts.

Step 5: Find the total for column (2) and (3).

Combing all the steps one can construct the following frequency distribution.

Class	Frequency	Percent
(1)	(3)	(4)
M	5	20
S	7	28

D	7	28
W	6	24

2. Ungrouped frequency Distribution:

Is a table of all the potential raw score values that could possibly occur in the data along with the number of times each actually occurred. Is often constructed for small set or data on discrete variable.

Constructing ungrouped frequency distribution:

- First find the smallest and largest raw score in the collected data.
- Arrange the data in order of magnitude and count the frequency.
- To facilitate counting one may include a column of tallies.

Example: The following data represent the mark of 20 students.

80	76	90	85	80
70	60	62	70	85
65	60	63	74	75
76	70	70	80	85

Construct a frequency distribution, which is ungrouped.

Solution:

Step 1: Find the range, $\text{Range} = \text{Max} - \text{Min} = 90 - 60 = 30$.

Step 2: Make a table as shown

Step 3: Compute the frequency.

Mark	60	62	63	65	70	74	75	76	80	85	90
Frequency	2	1	1	1	4	1	2	1	3	3	1

Each individual value is presented separately, that is why it is named ungrouped frequency distribution.

3. Grouped frequency Distribution:

When the range of the data is large, the data must be grouped in to classes that are more than one unit in width.

- **Grouped Frequency Distribution:** a frequency distribution when several numbers are grouped in one class.
- **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.
- **Units of measurement (U):** the distance between two possible consecutive measures. It is usually taken as 1, 0.1, 0.01, 0.001, -----.
- **Class boundaries:** Separates one class in a grouped frequency distribution from another. The boundaries have one more decimal places than the row data and therefore do not appear in the data. There is no gap between the upper boundary of one class and lower boundary of the next class. The lower class boundary is found 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:** is the number of observations less than/more than or equal to a specific value.
- **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 blow:** it is the total frequency of all values less than or equal to the upper class boundary of a given class.

- **Cumulative Frequency Distribution (CFD):** it is the tabular arrangement of class interval together with their corresponding cumulative frequencies. It can be more than or less than type, depending on the type of cumulative frequency used.
- **Relative frequency (rf):** it is the frequency divided by the total frequency.
- **Relative cumulative frequency (rcf):** it is the cumulative frequency divided by the total frequency.

Guidelines for classes

1. There should be between 5 and 20 classes.
2. The classes must be mutually exclusive. This means that no data value can fall into two different classes
3. The classes must be all inclusive or exhaustive i.e. all data values must be included.
4. The classes must be continuous. There are no gaps in a frequency distribution.
5. The classes must be equal in width. The exception here is the first or last class. It is possible to have "below ..." or "... and above" class. This is often used with ages.

Steps for constructing Grouped frequency Distribution

1. Find the largest and smallest values
2. Compute the Range(R) = Maximum - Minimum
3. Select the number of classes desired, usually between 5 and 20 or use Sturges rule $k = 1 + 3.322 \log n$ where k is number of classes desired and n is total number of observation.
4. Find the class width by dividing the range by the number of classes and rounding up, not off. $w = \frac{R}{k}$.
5. 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.
6. 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.

7. 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. It may not be necessary to find the boundaries.
8. Find the frequencies.
9. Find the cumulative frequencies. Depending on what you're trying to accomplish, it may not be necessary to find the cumulative frequencies.
10. If necessary, find the relative frequencies and/or relative cumulative frequencies

Example: Construct a frequency distribution for the following data.

11 29 6 33 14 31 22 27 19 20
18 17 22 38 23 21 26 34 39 27

Solutions:

Step 1: Find the highest and the lowest value $H=39$, $L=6$

Step 2: Find the range; $R=H-L=39-6=33$

Step 3: Select the number of classes desired using Sturges formula;

$$k = 1 + 3.322 \log n = 1 + 3.322 \log (20) = 5.32 = 6 \text{ (rounding up)}$$

Step 4: Find the class width; $w=R/k=33/6=5.5=6$ (rounding up)

Step 5: Select the starting point, let it be the minimum observation.

- 6, 12, 18, 24, 30, 36 are the lower class limits.

Step 6: Find the upper class limit; e.g. the first upper class $=12-U=12-1=11$

- 11, 17, 23, 29, 35, 41 are the upper class limits.

So combining step 5 and step 6, one can construct the following classes.

Class limits	6 – 11	12 – 17	18 – 23	24 – 29	30 – 35	36 – 41
--------------	--------	---------	---------	---------	---------	---------

Step 7: Find the class boundaries;

E.g. for class 1 Lower class boundary $=6-U/2=5.5$

Upper class boundary $=11+U/2=11.5$

- Then continue adding w on both boundaries to obtain the rest boundaries. By doing so one can obtain the following classes.

Class boundary	5.5 – 11.5	11.5 – 17.5	17.5 – 23.5	23.5 – 29.5	29.5 – 35.5	35.5 – 41.5
----------------	------------	-------------	-------------	-------------	-------------	-------------

Step 8: Write the numeric values for the tallies in the frequency column.

Step 9: Find cumulative frequency.

Step 10: Find relative frequency or/and relative cumulative frequency.

The complete frequency distribution follows:

Class limit	Class boundary	Class Mark	Freq.	Cf (less than type)	Cf (more than type)	rf.	rcf (less than type)
6 – 11	5.5 – 11.5	8.5	2	2	20	0.10	0.10
12 – 17	11.5 – 17.5	14.5	2	4	18	0.10	0.20
18 – 23	17.5 – 23.5	20.5	7	11	16	0.35	0.55
24 – 29	23.5 – 29.5	26.5	4	15	9	0.20	0.75
30 – 35	29.5 – 35.5	32.5	3	18	5	0.15	0.90
36 – 41	35.5 – 41.5	38.5	2	20	2	0.10	1.00

Diagrammatic and Graphic presentation of data.

These are techniques for presenting data in visual displays using geometric and pictures.

Importance:

- They have greater attraction.
- They facilitate comparison.
- They are easily understandable.

-Diagrams are appropriate for presenting discrete data.

-The two most commonly used diagrammatic presentation for discrete as well as qualitative data are:

- Pie charts
- Bar charts

Pie chart

- A pie chart is a circle that is divided in to sections or wedges according to the percentage of frequencies in each category of the distribution. The angle of the sector is obtained using:

$$\text{Angle of sector} = \frac{\text{Value of the part}}{\text{the whole quantity}} * 100$$

Example: Draw a suitable diagram to represent the following population in a town.

Men Women Girls Boys

Lecture notes on Statistics & Probability

2500 2000 4000 1500

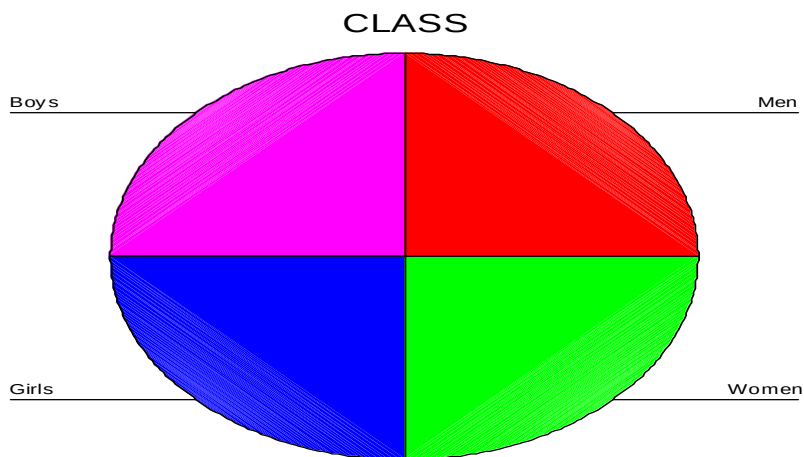
Solutions:

Step 1: Find the percentage.

Step 2: Find the number of degrees for each class.

Step 3: Using a protractor and compass, graph each section and write its name corresponding percentage.

Class	Frequency	Percent	Degree
Men	2500	25	90
Women	2000	20	72
Girls	4000	40	144
Boys	1500	15	54



Bar Charts:

- A set of bars (thick lines or narrow rectangles) representing some magnitude over time space.
- They are useful for comparing aggregate over time space.
- Bars can be drawn either vertically or horizontally.
- There are different types of bar charts. The most common being :
 - Simple bar chart
 - Deviation or two way bar chart
 - Broken bar chart
 - Component or sub divided bar chart.

- Multiple bar charts.

Simple Bar Chart

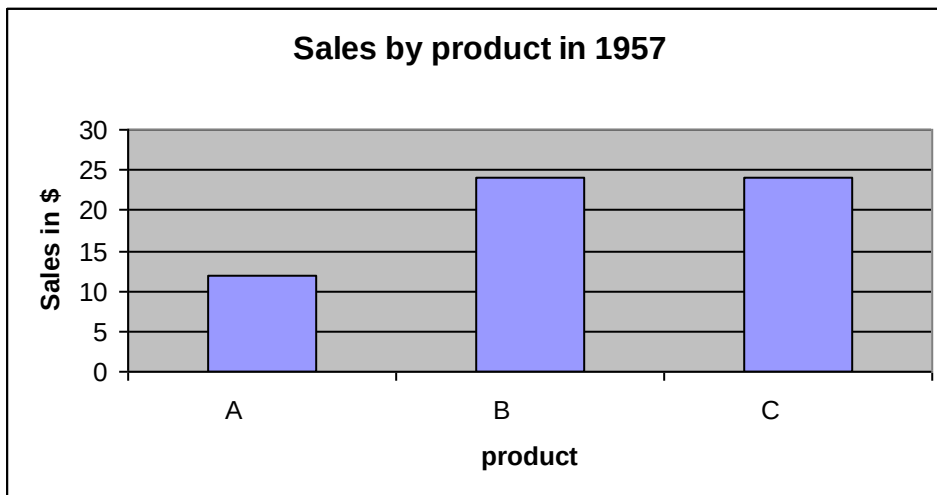
-Are used to display data on one variable.

-They are thick lines (narrow rectangles) having the same breadth. The magnitude of a quantity is represented by the height /length of the bar.

Example: The following data represent sale by product, 1957- 1959 of a given company for three products A, B, C.

Product	Sales(\$)	Sales(\$)	Sales(\$)
	In 1957	In 1958	In 1959
A	12	14	18
B	24	21	18
C	24	35	54

Solutions:



Component Bar chart

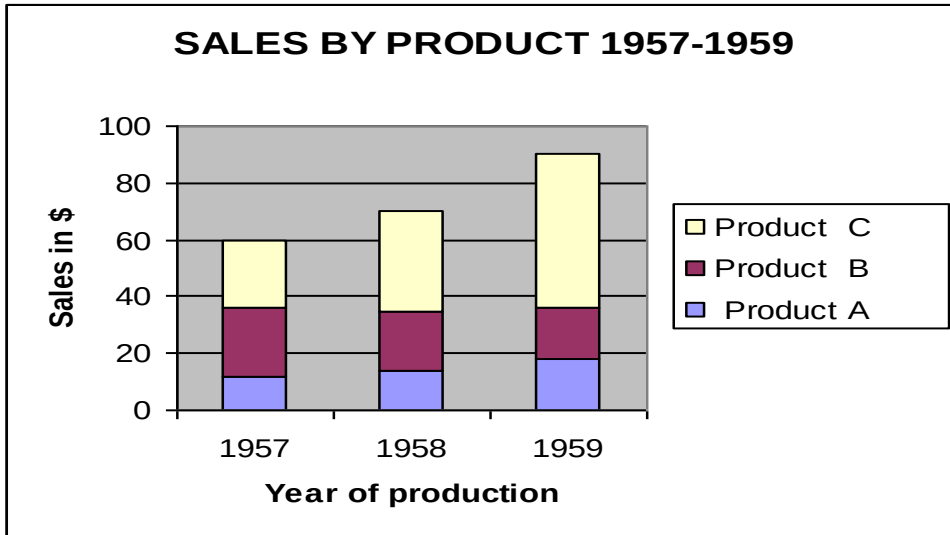
-When there is a desire to show how a total (or aggregate) is divided in to its component parts, we use component bar chart.

-The bars represent total value of a variable with each total broken in to its component parts and different colours or designs are used for identifications

Example:

Draw a component bar chart to represent the sales by product from 1957 to 1959.

Solutions:



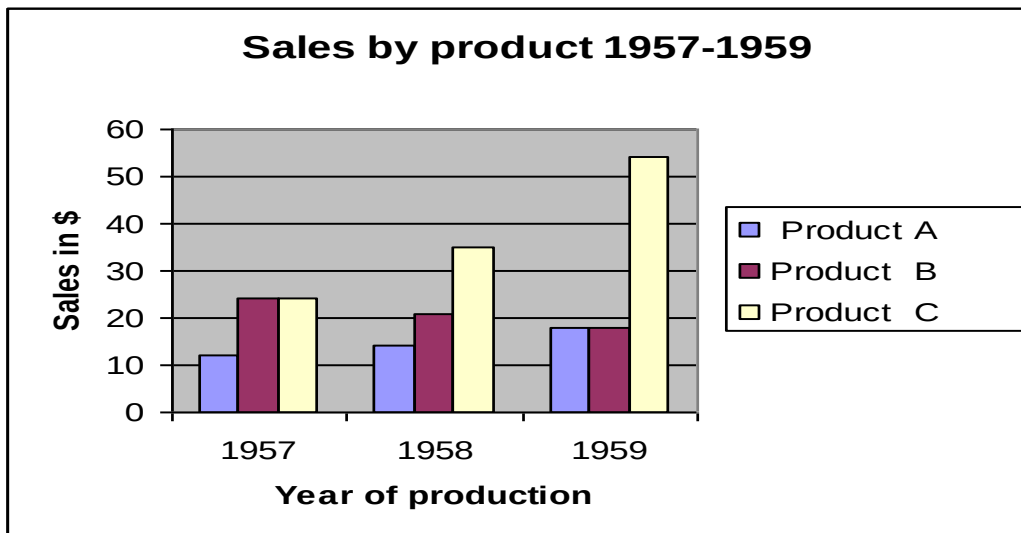
Multiple Bar charts

- These are used to display data on more than one variable.
- They are used for comparing different variables at the same time.

Example:

Draw a component bar chart to represent the sales by product from 1957 to 1959.

Solutions:



Graphical Presentation of data

- The histogram, frequency polygon and cumulative frequency graph or ogive is most commonly applied graphical representation for continuous data.

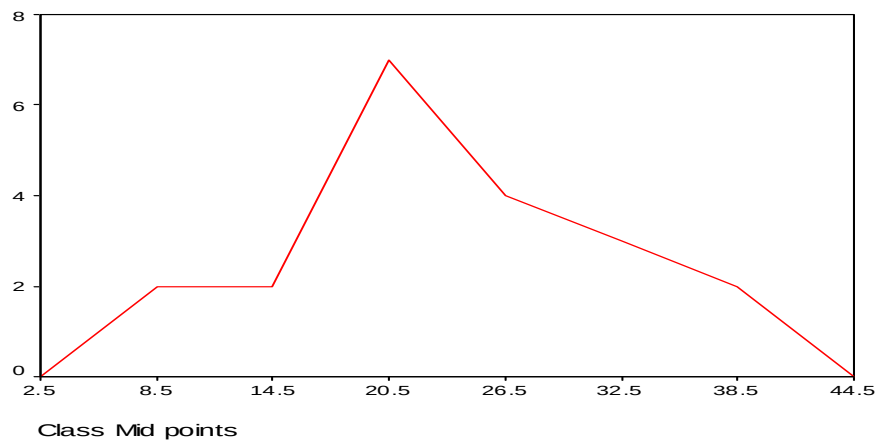
Procedures for constructing statistical graphs:

- Draw and label the X and Y axes.
- Choose suitable scale for frequencies or cumulative frequencies and label the Y axes.
- Represent the class boundaries for the histogram or ogive or the mid points for the frequency polygon on the X axes.
- Plot the points.
- Draw the bars or lines to connect the points.

Histogram: A graph which displays the data by using vertical bars of various heights to represent frequencies. Class boundaries are placed along the horizontal axes. Class marks and class limits are sometimes used as quantity on the X axes.

Frequency Polygon: A line graph. The frequency is placed along the vertical axis and classes mid points are placed along the horizontal axis. It is customer to the next higher and lower class interval with corresponding frequency of zero, this is to make it a complete polygon.

Example: Draw a frequency polygon for the above data (example *).



Ogive (cumulative frequency polygon)

- A graph showing the cumulative frequency (less than or more than type) plotted against upper or lower class boundaries respectively. That is 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.

2. Summarizing of 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. This single value is called the **average** of the group. Averages are also called measures of central tendency.
- An average which is representative is called typical average and an average which is not representative and has only a theoretical value is called a descriptive average. A typical average should possess the following:
 - It should be rigidly defined.
 - It should be based on all observation under investigation.
 - 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.

Objectives:

- ☞ To comprehend the data easily.
- ☞ To facilitate comparison.
- ☞ To make further statistical analysis

The Summation Notation:

- Let $X_1, X_2, X_3 \dots X_N$ be a number of measurements where N is the total number of observation and X_i is i^{th} observation.
- Very often in statistics an algebraic expression of the form $X_1 + X_2 + X_3 + \dots + X_N$ is used in a formula to compute a statistic. It is tedious to write an expression like this very often, so mathematicians have developed a shorthand notation to represent a sum of scores, called the summation notation.
- The symbol $\sum_{i=1}^N X_i$ is a mathematical shorthand for $X_1 + X_2 + X_3 + \dots + X_N$

$$\sum_{i=1}^N X_i = X_1 + X_2 + \dots + X_N$$

The expression is read, "the sum of X sub i from i equals 1 to N ." It means "add up all the numbers."

Example: Suppose the following were scores made on the first homework assignment for five students in the class: 5, 7, 7, 6, and 8. In this example set of five numbers, where $N=5$, the summation could be written:

$$\sum_{i=1}^5 X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 5 + 7 + 7 + 6 + 8 = 33$$

Sometimes if the summation notation is used in an expression and the expression must be written a number of times, as in a proof, then a shorthand notation for the shorthand notation is employed. When the summation sign $\sum$ is used without additional notation, then " $i=1$ " and " N " are assumed.

For example:

$$\sum X = \sum_{i=1}^N X_i = X_1 + X_2 + \dots + X_N$$

PROPERTIES OF SUMMATION

1. $\sum_{i=1}^n k = nk$ where k is any constant
2. $\sum_{i=1}^n kX_i = k \sum_{i=1}^n X_i$ where k is any constant
3. $\sum_{i=1}^n (a + bX_i) = na + b \sum_{i=1}^n X_i$ where a and b are any constant
4. $\sum_{i=1}^n (X_i + Y_i) = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$

The sum of the product of the two variables could be written:

$$\sum_{i=1}^N (X_i * Y_i) = (X_1 * Y_1) + (X_2 * Y_2) + \dots + (X_N * Y_N)$$

Example: considering the following data determine

X	5	7	7	6	8
Y	6	7	8	7	8

a) $\sum_{i=1}^5 X_i$

d) $\sum_{i=1}^5 (X_i + Y_i)$

g) $\sum_{i=1}^5 X_i^2$

b) $\sum_{i=1}^5 Y_i$

e) $\sum_{i=1}^5 (X_i - Y_i)$

h) $(\sum_{i=1}^5 X_i)(\sum_{i=1}^5 Y_i)$

c) $\sum_{i=1}^5 10$

f) $\sum_{i=1}^5 X_i Y_i$

Solutions:

a) $\sum_{i=1}^5 X_i = 5 + 7 + 7 + 6 + 8 = 33$

b) $\sum_{i=1}^5 Y_i = 6 + 7 + 8 + 7 + 8 = 36$

c) $\sum_{i=1}^5 10 = 5 * 10 = 50$

d) $\sum_{i=1}^5 (X_i + Y_i) = (5 + 6) + (7 + 7) + (7 + 8) + (6 + 7) + (8 + 8) = 69 = 33 + 36$

e) $\sum_{i=1}^5 (X_i - Y_i) = (5 - 6) + (7 - 7) + (7 - 8) + (6 - 7) + (8 - 8) = -3 = 33 - 36$

f) $\sum_{i=1}^5 X_i Y_i = 5 * 6 + 7 * 7 + 7 * 8 + 6 * 7 + 8 * 8 = 241$

g) $\sum_{i=1}^5 X_i^2 = 5^2 + 7^2 + 7^2 + 6^2 + 8^2 = 223$

h) $(\sum_{i=1}^5 X_i)(\sum_{i=1}^5 Y_i) = 33 * 36 = 1188$

Types of measures of central tendency

There are several different measures of central tendency; each has its advantage and disadvantage.

- The Mean (Arithmetic, Geometric and Harmonic)
- The Mode
- The Median
- Quantiles (Quartiles, Deciles and Percentiles)

The choice of these averages depends up on which best fit the property under discussion.

The Arithmetic Mean

- Is defined as the sum of the magnitude of the items divided by the number of items.
- The mean of $X_1, X_2, X_3 \dots X_n$ is denoted by A.M ,m or $\bar{X}$ and is given by:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$
$$\Rightarrow \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- If X_1 occurs f_1 times
- If X_2 occurs f_2 times
- .
- .
- If X_n occurs f_n times

Then the mean will be $\bar{X} = \frac{\sum_{i=1}^k f_i X_i}{\sum_{i=1}^k f_i}$, where k is the number of classes and

$$\sum_{i=1}^k f_i = n$$

Example: Obtain the mean of the following number

2, 7, 8, 2, 7, 3, 7

Solution:

X_i	f_i	$X_i f_i$
2	2	4
3	1	3

7	3	21
8	1	8
Total	7	36

$$\bar{X} = \frac{\sum_{i=1}^4 f_i X_i}{\sum_{i=1}^4 f_i} = \frac{36}{7} = 5.15$$

Arithmetic Mean for Grouped Data

If data are given in the shape of a continuous frequency distribution, then the mean is obtained as follows:

$$\bar{X} = \frac{\sum_{i=1}^k f_i X_i}{\sum_{i=1}^k f_i}, \text{ Where } X_i = \text{the class mark of the } i^{\text{th}} \text{ class and } f_i = \text{the frequency of the } i^{\text{th}}$$

class

Example: calculate the mean for the following age distribution.

Class	6- 10	11- 15	16- 20	21- 25	26- 30	31- 35
frequency	35	23	15	12	9	6

Solutions:

- First find the class marks
- Find the product of frequency and class marks
- Find mean using the formula.

Class	6- 10	11- 15	16- 20	21- 25	26- 30	31- 35	Total
f_i	35	23	15	12	9	6	100
X_i	8	13	18	23	28	33	
$X_i f_i$	280	299	270	276	252	198	1575

$$\bar{X} = \frac{\sum_{i=1}^6 f_i X_i}{\sum_{i=1}^6 f_i} = \frac{1575}{100} = 15.75$$

Exercises:

1. Marks of 75 students are summarized in the following frequency distribution:

Marks	40-44	45-49	50-54	55-59	60-64	65-69	70-74
No. of students	7	10	22	f_4	f_5	6	3

If 20% of the students have marks between 55 and 59

- i. Find the missing frequencies f_4 and f_5 .
- ii. Find the mean.

Special properties of Arithmetic mean

1. The sum of the deviations of a set of items from their mean is always zero.

$$\text{i.e. } \sum_{i=1}^n (X_i - \bar{X}) = 0.$$

2. If $\bar{X}_1$ is the mean of n_1 observations

If $\bar{X}_2$ is the mean of n_2 observations

·
·

If $\bar{X}_k$ is the mean of n_k observations

Then the mean of all the observation in all groups often called the combined mean is given by:

$$\bar{X}_c = \frac{\bar{X}_1 n_1 + \bar{X}_2 n_2 + \dots + \bar{X}_k n_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k \bar{X}_i n_i}{\sum_{i=1}^k n_i}$$

Example: In a class there are 30 females and 70 males. If females averaged 60 in an examination and boys averaged 72, find the mean for the entire class.

Solutions:

Females

$$\bar{X}_1 = 60$$

$$n_1 = 30$$

Males

$$\bar{X}_2 = 72$$

$$n_2 = 70$$

$$\bar{X}_c = \frac{\bar{X}_1 n_1 + \bar{X}_2 n_2}{n_1 + n_2} = \frac{\sum_{i=1}^2 \bar{X}_i n_i}{\sum_{i=1}^2 n_i}$$

$$\Rightarrow \bar{X}_c = \frac{30(60) + 70(72)}{30 + 70} = \frac{6840}{100} = 68.40$$

3. If a wrong figure has been used when calculating the mean the correct mean can be obtained without repeating the whole process using:

$$\text{CorrectMean} = \text{WrongMean} + \frac{(\text{CorrectValue} - \text{WrongValue})}{n}$$

Where n is total number of observations.

Example: An average weight of 10 students was calculated to be 65. Later it was discovered that one weight was misread as 40 instead of 80 k.g. Calculate the correct average weight.

Solutions:

$$\text{CorrectMean} = \text{WrongMean} + \frac{(\text{CorrectValue} - \text{WrongValue})}{n}$$

$$\text{CorrectMean} = 65 + \frac{(80 - 40)}{10} = 65 + 4 = 69 \text{ k.g.}$$

4. The effect of transforming original series on the mean.
- If a constant k is added/ subtracted to/from every observation then the new mean will be the old mean $\pm$ k respectively.
 - If every observation is multiplied by a constant k then the new mean will be k*old mean

Example:

- The mean of n Tetracycline Capsules $X_1, X_2, \dots, X_n$ are known to be 12 gm. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ ($i = 1, 2, \dots, n$) then what will be the mean of the new set of capsules

Solutions:

$$\text{NewMean} = 2 * \text{OldMean} - 0.5 = 2 * 12 - 0.5 = 23.5$$

- The mean of a set of numbers is 500.

- a) If 10 is added to each of the numbers in the set, then what will be the mean of the new set?
- b) If each of the numbers in the set are multiplied by -5, then what will be the mean of the new set?

Solutions:

a). $\text{New Mean} = \text{Old Mean} + 10 = 500 + 10 = 510$

b). $\text{New Mean} = -5 * \text{Old Mean} = -5 * 500 = -2500$

Weighted Mean

- ☞ When a proper importance is desired to be given to different data a weighted mean is appropriate.
- ☞ Weights are assigned to each item in proportion to its relative importance.
- ☞ Let $X_1, X_2, \dots, X_n$ be the value of items of a series and $W_1, W_2, \dots, W_n$ their corresponding weights, then the weighted mean denoted $\bar{X}_w$ is defined as:

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

Example: A student obtained the following percentage in an examination: English 60, Biology 75, Mathematics 63, Physics 59, and chemistry 55. Find the student's weighted arithmetic mean if weights 1, 2, 1, 3, 3 respectively are allotted to the subjects.

Solutions:

$$\bar{X}_w = \frac{\sum_{i=1}^5 X_i W_i}{\sum_{i=1}^5 W_i} = \frac{60*1 + 75*2 + 63*1 + 59*3 + 55*3}{1 + 2 + 1 + 3 + 3} = \frac{615}{10} = 61.5$$

Merits of Arithmetic Mean

- It is rigidly defined.
- It is based on all observation.
- 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 of Arithmetic Mean

- It is affected by extreme observations.
- It can not be used in the case of open end classes.
- It can not be determined by the method of inspection.
- 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 series.
- Some times it leads to wrong conclusion if the details of the data from which it is obtained are not available.
- It gives high weight to high extreme values and less weight to low extreme values.

The Geometric Mean

- ☞ The geometric mean of a set of n observation is the n^{th} root of their product.
- ☞ The geometric mean of $X_1, X_2, X_3 \dots X_n$ is denoted by G.M and given by:

$$G.M = \sqrt[n]{X_1 * X_2 * \dots * X_n}$$

- ☞ Taking the logarithms of both sides

$$\begin{aligned}\log(G.M) &= \log(\sqrt[n]{X_1 * X_2 * \dots * X_n}) = \log(X_1 * X_2 * \dots * X_n)^{\frac{1}{n}} \\ \Rightarrow \log(G.M) &= \frac{1}{n} \log(X_1 * X_2 * \dots * X_n) = \frac{1}{n} (\log X_1 + \log X_2 + \dots + \log X_n) \\ \Rightarrow \log(G.M) &= \frac{1}{n} \sum_{i=1}^n \log X_i\end{aligned}$$

The logarithm of the G.M of a set of observation is the arithmetic mean of their logarithm.

$$\Rightarrow G.M = \text{Antilog}\left(\frac{1}{n} \sum_{i=1}^n \log X_i\right)$$

Example: Find the G.M of the numbers 2, 4, 8.

Solutions:

$$G.M = \sqrt[3]{X_1 * X_2 * \dots * X_n} = \sqrt[3]{2 * 4 * 8} = \sqrt[3]{64} = 4$$

Remark: The Geometric Mean is useful and appropriate for finding averages of ratios.

The Harmonic Mean

The harmonic mean of $X_1, X_2, X_3 \dots X_n$ is denoted by H.M and given by:

$$\boxed{H.M = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}}}, \text{ This is called simple harmonic mean.}$$

In a case of frequency distribution:

$$\boxed{H.M = \frac{n}{\sum_{i=1}^k \frac{f_i}{X_i}}}, \text{ } n = \sum_{i=1}^k f_i$$

If observations $X_1, X_2, \dots X_n$ have weights $W_1, W_2, \dots W_n$ respectively, then their harmonic mean is given by

$$\boxed{H.M = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n w_i / X_i}}, \text{ This is called Weighted Harmonic Mean.}$$

Remark: The Harmonic Mean is useful and appropriate in finding average speeds and average rates.

Example: A cyclist pedals from his house to his college at speed of 10 km/hr and back from the college to his house at 15 km/hr. Find the average speed.

Solution: Here the distance is constant

→ The simple H.M is appropriate for this problem.

$X_1 = 10 \text{ km/hr}$ $X_2 = 15 \text{ km/hr}$

$$H.M = \frac{2}{\frac{1}{10} + \frac{1}{15}} = 12 \text{ km/hr}$$

The Mode

- Mode is a value which occurs most frequently in a set of values
- The mode may not exist and even if it does exist, it may not be unique.

- In case of discrete distribution the value having the maximum frequency is the modal value.

Examples:

1. Find the mode of 5, 3, 5, 8, and 9. Mode = 5
2. Find the mode of 8, 9, 9, 7, 8, 2, and 5. It is a bimodal Data: 8 and 9
3. Find the mode of 4, 12, 3, 6, and 7. No mode for this data.

- The mode of a set of numbers $X_1, X_2, \dots, X_n$ is usually denoted by $\hat{X}$.

Mode for Grouped data

If data are given in the shape of continuous frequency distribution, the mode is defined as:

$$\hat{X} = L_{mo} + w \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)$$

Where:

$\hat{X}$ = the mode of the distribution

w = the size of the modal class

$\Delta_1 = f_{mo} - f_1$

$\Delta_2 = f_{mo} - f_2$

f_{mo} = frequency of the modal class

f_1 = frequency of the class preceding the modal class

f_2 = frequency of the class following the modal class

Note: The modal class is a class with the highest frequency.

Example: Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farms	5-15	15-25	25-35	35-45	45-55	55-65	65-75
No. of farms	8	12	17	29	31	5	3

Solutions:

45 – 55 is the modal class, since it is a class with the highest frequency.

$$L_{mo} = 45$$

$$w = 10$$

$$\Delta_1 = f_{mo} - f_1 = 2$$

$$\Delta_2 = f_{mo} - f_2 = 26$$

$$f_{mo} = 31$$

$$f_1 = 29$$

$$f_2 = 5$$

$$\begin{aligned}\Rightarrow \hat{X} &= 45 + 10 \left(\frac{2}{2 + 26} \right) \\ &= 45.71\end{aligned}$$

Merits of Mode

- 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 of Mode

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

Note: being the point of maximum density, mode is especially useful in finding the most popular size in studies relating to marketing, trade, business, and industry. It is the appropriate average to be used to find the ideal size.

The Median

- In a distribution, median is the value of the variable which divides it into two equal halves.
 - In an ordered series of data median is an observation lying exactly in the middle of the series.
- It is the middle most value in the sense that the number of values less than the median is equal to the number of values greater than it.

-If $X_1, X_2, \dots, X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]}, \dots, X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value.

$$\Rightarrow X_{[1]} < X_{[2]} < \dots < X_{[n]}$$

-Median is denoted by $\hat{X}$.

Median for ungrouped data

$$\tilde{X} = \begin{cases} X_{[(n+1)/2]} & , \text{If } n \text{ is odd.} \\ \frac{1}{2}(X_{[n/2]} + X_{[(n/2)+1]}), & \text{If } n \text{ is even} \end{cases}$$

Example: Find the median of the following numbers.

a) 6, 5, 2, 8, 9, 4.

b) 2, 1, 8, 3, 5, 8.

Solutions:

a) First order the data: 2, 4, 5, 6, 8, 9

Here $n=6$

$$\begin{aligned} \tilde{X} &= \frac{1}{2}(X_{[\frac{n}{2}]} + X_{[\frac{n}{2}+1]}) \\ &= \frac{1}{2}(X_{[3]} + X_{[4]}) \\ &= \frac{1}{2}(5 + 6) = 5.5 \end{aligned}$$

b) Order the data :1, 2, 3, 5, 8

Here $n=5$

$$\begin{aligned} \tilde{X} &= X_{[\frac{n+1}{2}]} \\ &= X_{[3]} \\ &= 3 \end{aligned}$$

Median for grouped data

If data are given in the shape of continuous frequency distribution, the median is defined

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

Where:

- as:
- L_{med} = lower class boundary of the median class.
 - w = the size of the median class
 - n = total number of observations.
 - c = the cumulative frequency (less than type) preceeding the median class.
 - f_{med} = the frequency of the median class.

Remark: The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{n}{2}$.

Example: Find the median of the following distribution.

Class	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Frequency	7	10	22	15	12	6	3

Solutions:

- First find the less than cumulative frequency.
- Identify the median class.
- Find median using formula.

Class	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Frequency	7	10	22	15	12	6	3
Cumu.Freq (less than type)	7	17	39	54	66	72	75

$$\frac{n}{2} = \frac{75}{2} = 37.5$$

39 is the first cumulative frequency to be greater than or equal to 37.5
 $\Rightarrow 50 - 54$ is the median class.

$$\begin{aligned} L_{\text{med}} &= 49.5, & w &= 5 \\ n &= 75, & c &= 17, & f_{\text{med}} &= 22 \end{aligned}$$

$$\begin{aligned} \Rightarrow \tilde{X} &= L_{\text{med}} + \frac{w}{f_{\text{med}}} \left(\frac{n}{2} - c \right) \\ &= 49.5 + \frac{5}{22} (37.5 - 17) \\ &= 54.16 \end{aligned}$$

Merits of Median

- 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 of Median

- 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

When a distribution is arranged in order of magnitude of items, the median is the value of the middle term. Their measures that depend up on their positions in distribution quartiles, deciles, and percentiles are collectively called quantiles.

Quartiles:

- Quartiles are measures that divide the frequency distribution in to four equal parts.
- The value of the variables corresponding to these divisions are denoted Q_1 , Q_2 , and Q_3 often called the first, the second and the third quartile respectively.
- Q_1 is a value which has 25% items which are less than or equal to it. Similarly Q_2 has 50% items with value less than or equal to it and Q_3 has 75% items whose values are less than or equal to it.
- To find Q_i ($i=1, 2, 3$) we count $\frac{iN}{4}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

$$Q_i = L_{Q_i} + \frac{w}{f_{Q_i}} \left(\frac{iN}{4} - c \right), i = 1, 2, 3$$

Where:

L_{Q_i} = lower class boundary of the quartile class.

w = the size of the quartile class

N = total number of observations.

c = the cumulative frequency (less than type) preceeding the quartile class.

f_{Q_i} = the frequency of the quartile class.

Remark: The quartile class (class containing Q_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{iN}{4}$.

Deciles:

- Deciles are measures that divide the frequency distribution in to ten equal parts.
- The values of the variables corresponding to these divisions are denoted $D_1, D_2, \dots, D_9$ often called the first, the second, ..., the ninth decile respectively.
- To find D_i ($i=1, 2, \dots, 9$) we count $\frac{iN}{10}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

$$D_i = L_{D_i} + \frac{w}{f_{D_i}} \left(\frac{iN}{10} - c \right), i = 1, 2, \dots, 9$$

Where:

L_{D_i} = lower class boundary of the decile class.

w = the size of the decile class

N = total number of observations.

c = the cumulative frequency (less than type) preceeding the decile class.

f_{D_i} = the frequency of the decile class.

Remark: The decile class (class containing D_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{iN}{10}$.

Percentiles:

- Percentiles are measures that divide the frequency distribution in to hundred equal parts.

- The values of the variables corresponding to these divisions are denoted $P_1, P_2, \dots, P_{99}$ often called the first, the second, ..., the ninety-ninth percentile respectively.
- To find P_i ($i=1, 2, \dots, 99$) we count $\frac{iN}{100}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

$$P_i = L_{p_i} + \frac{w}{f_{p_i}} \left(\frac{iN}{100} - c \right), i = 1, 2, \dots, 99$$

Where:

L_{p_i} = lower class boundary of the percentile class.

w = the size of the percentile class

N = total number of observations.

c = the cumulative frequency (less than type) preceeding the percentile class.

f_{p_i} = the frequency of the percentile class.

Remark: The percentile class (class containing P_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{iN}{100}$.

Example: Considering the following distribution

Calculate:

- a) All quartiles.
- b) The 7th decile.
- c) The 90th percentile.

Values	140-150	150-160	160-170	170-180	180-190	190-200	200-210	210-220	220-230	230-240	240-250
Frequency	17	29	42	72	84	107	49	34	31	16	12

Solutions:

- First find the less than cumulative frequency.
- Use the formula to calculate the required quantile.

Values	140-150	150-160	160-170	170-180	180-190	190-200	200-210	210-220	220-230	230-240	240-250
Frequency	17	29	42	72	84	107	49	34	31	16	12
Cum.Freq(less than type)	17	46	88	160	244	351	400	434	465	481	493

a) Quartiles:

i. Q_1

- determine the class containing the first quartile.

$$\frac{N}{4} = 123.25$$

$\Rightarrow 170 - 180$ is the class containing the first quartile.

$$\begin{aligned} L_{Q_1} &= 170, & w &= 10 \\ N &= 493, & c &= 88, & f_{Q_1} &= 72 \end{aligned}$$

$$\begin{aligned} \Rightarrow Q_1 &= L_{Q_1} + \frac{w}{f_{Q_1}} \left(\frac{N}{4} - c \right) \\ &= 170 + \frac{10}{72} (123.25 - 88) \\ &= \underline{\underline{174.90}} \end{aligned}$$

ii. Q_2

- determine the class containing the second quartile.

$$\frac{2 * N}{4} = 246.5$$

$\Rightarrow 190 - 200$ is the class containing the second quartile

$$\begin{aligned} L_{Q_2} &= 190, & w &= 10 \\ N &= 493, & c &= 244, & f_{Q_2} &= 107 \end{aligned}$$

$$\begin{aligned} \Rightarrow Q_2 &= L_{Q_2} + \frac{w}{f_{Q_2}} \left(\frac{2 * N}{4} - c \right) \\ &= 190 + \frac{10}{107} (246.5 - 244) \\ &= \underline{\underline{190.23}} \end{aligned}$$

iii. Q_3

- determine the class containing the third quartile.

$$\frac{3 * N}{4} = 369.75$$

$\Rightarrow 200 - 210$ is the class containing the third quartile

$$\begin{aligned} L_{Q_3} &= 200, & w &= 10 \\ N &= 493, & c &= 351, & f_{Q_3} &= 49 \end{aligned}$$

$$\begin{aligned} \Rightarrow Q_3 &= L_{Q_3} + \frac{w}{f_{Q_3}} \left(\frac{3 * N}{4} - c \right) \\ &= 200 + \frac{10}{49} (369.75 - 351) \\ &= \underline{\underline{203.83}} \end{aligned}$$

b) D_7

- determine the class containing the 7th decile.

$$\frac{7 * N}{10} = 345.1$$

$\Rightarrow 190 - 200$ is the class containing the seventh decile.

$$\begin{aligned} L_{D_7} &= 190, & w &= 10 \\ N &= 493, & c &= 244, & f_{D_7} &= 107 \end{aligned}$$

$$\begin{aligned} \Rightarrow D_7 &= L_{D_7} + \frac{w}{f_{D_7}} \left(\frac{7 * N}{10} - c \right) \\ &= 190 + \frac{10}{107} (345.1 - 244) \\ &= \underline{\underline{199.45}} \end{aligned}$$

c) P_{90}

- determine the class containing the 90th percentile.

$$\frac{90 * N}{100} = 443.7$$

$\Rightarrow 220 - 230$ is the class containing the 90th percentile

$$\begin{aligned} L_{P_{90}} &= 220, & w &= 10 \\ N &= 493, & c &= 434, & f_{P_{90}} &= 31 \end{aligned}$$

$$\begin{aligned} \Rightarrow P_{90} &= L_{P_{90}} + \frac{w}{f_{P_{90}}} \left(\frac{90 * N}{100} - c \right) \\ &= 220 + \frac{10}{31} (443.7 - 434) \\ &= \underline{\underline{223.13}} \end{aligned}$$

3. Measures of Dispersion (Variation)

Introduction and objectives of measuring Variation

-The scatter or spread of items of a distribution is known as dispersion or variation. In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.

-Measures of dispersions are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

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.

Absolute and Relative Measures of Dispersion

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. For comparing the variability of two distributions (even if they are measured in the same unit), we compute the relative measure of dispersion instead of absolute measures of dispersion.

Types of Measures of Dispersion

Various measures of dispersions are in use. The most commonly used measures of dispersions are:

Absolute measure of Variation

1. Range
2. Quartile deviation
3. Mean deviation
4. Standard deviation

Relative measure of Variation

1. Relative range
2. Coefficient of Quartile deviation
3. Coefficient of Mean deviation
4. Coefficient of variation.

The Range (R)

The range is the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

Distribution 1: 32 35 36 36 37 38 40 42 42 43 43 45

Distribution 2: 32 32 33 33 33 34 34 34 34 34 35 45

For this reason, among others, the range is not the most important measure of variability.

$$R = L - S \quad , L = \text{largest observation}, S = \text{smallest observation}$$

Range for grouped data:

If data are given in the shape of continuous frequency distribution, the range is computed as:

$$R = UCL_k - LCL_1, \quad \begin{array}{l} UCL_k \text{ is upper class limit of the last class.} \\ UCL_1 \text{ is upper class limit of the first class.} \end{array}$$

This is sometimes expressed as:

$$R = X_k - X_1, \quad \begin{array}{l} X_k \text{ is class mark of the last class.} \\ X_1 \text{ is class mark of the first class.} \end{array}$$

Merits of range

- It is rigidly defined.
- It is easy to calculate and simple to understand.

Demerits of range

- It is not based on all observation.
- It is highly affected by extreme observations.
- It is affected by fluctuation in sampling.
- It is not liable to further algebraic treatment.
- It cannot be computed in the case of open end distribution.
- It is very sensitive to the size of the sample.

Relative Range (RR)

-it is also sometimes called coefficient of range and given by:

$$RR = \frac{L - S}{L + S} = \frac{R}{L + S}$$

Example:

1. Find the relative range of the above two distribution.(exercise!)
2. If the range and relative range of a series are 4 and 0.25 respectively. Then what is the value of:
 - a) Smallest observation
 - b) Largest observation

Solutions :(2)

$$R = 4 \Rightarrow L - S = 4 \text{ _____ (1)}$$

$$RR = 0.25 \Rightarrow L + S = 16 \text{ _____ (2)}$$

Solving (1) and (2) at the same time, one can obtain the following value

$$L = 10 \text{ and } S = 6$$

The Quartile Deviation (Semi-inter quartile range), Q.D

The inter quartile range is the difference between the third and the first quartiles of a set of items and semi-inter quartile range is half of the inter quartile range.

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Coefficient of Quartile Deviation (C.Q.D)

$$C.Q.D = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{2 * Q.D}{Q_3 + Q_1} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

➤ It gives the average amount by which the two quartiles differ from the median.

Example: Compute Q.D and its coefficient for the following distribution

Values	140- 150	150- 160	160- 170	170- 180	180- 190	190- 200	200- 210	210- 220	220- 230	230- 240	240- 250
Frequency	17	29	42	72	84	107	49	34	31	16	12

Solutions:

In the previous chapter we have obtained the values of all quartiles as:

$$Q_1 = 174.90, \quad Q_2 = 190.23, \quad Q_3 = 203.83$$

$$\Rightarrow Q.D = \frac{Q_3 - Q_1}{2} = \frac{203.83 - 174.90}{2} = 14.47$$

$$C.Q.D = \frac{2 * Q.D}{Q_3 + Q_1} = \frac{2 * 14.47}{203.83 + 174.90} = 0.076$$

Remark: Q.D or C.Q.D includes only the middle 50% of the observation.

The Mean Deviation (M.D):

The mean deviation of a set of items is defined as the arithmetic mean of the values of the absolute deviations from a given average. Depending up on the type of averages used we have different mean deviations.

a) Mean Deviation about the mean

- Denoted by $M.D(\bar{X})$ and given by

$$M.D(\bar{X}) = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

- For the case of frequency distribution it is given as:

$$M.D(\bar{X}) = \frac{\sum_{i=1}^k f_i |X_i - \bar{X}|}{n}$$

Steps to calculate M.D ($\bar{X}$):

1. Find the arithmetic mean, $\bar{X}$
2. Find the deviations of each reading from $\bar{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

b) Mean Deviation about the median.

- Denoted by $M.D(\tilde{X})$ and given by

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^n |X_i - \tilde{X}|}{n}$$

- For the case of frequency distribution it is given as:

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^k f_i |X_i - \tilde{X}|}{n}$$

Steps to calculate M.D ($\tilde{X}$):

1. Find the median, $\tilde{X}$
 2. Find the deviations of each reading from $\tilde{X}$.
 3. Find the arithmetic mean of the deviations, ignoring sign.
- c) Mean Deviation about the mode.

- Denoted by M.D($\hat{X}$) and given by

$$M.D(\hat{X}) = \frac{\sum_{i=1}^n |x_i - \hat{x}|}{n}$$

- For the case of frequency distribution it is given as:

$$M.D(\hat{X}) = \frac{\sum_{i=1}^k f_i |X_i - \hat{X}|}{n}$$

Steps to calculate M.D ($\hat{X}$):

1. Find the mode, $\hat{X}$
2. Find the deviations of each reading from $\hat{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

Examples:

1. The following are the number of visit made by ten mothers to the local doctor's surgery.
8, 6, 5, 5, 7, 4, 5, 9, 7, 4. Find mean deviation about mean, median and mode.

Solutions: First calculate the three averages

$$\bar{X} = 6, \tilde{X} = 5.5, \hat{X} = 5$$

Then take the deviations of each observation from these averages.

X_i	4	4	5	5	5	6	7	7	8	9	Total
$ X_i - 6 $	2	2	1	1	1	0	1	1	2	3	14
$ X_i - 5.5 $	1.5	1.5	0.5	0.5	0.5	0.5	1.5	1.5	2.5	3.5	14
$ X_i - 5 $	1	1	0	0	0	1	2	2	3	4	14

$$\Rightarrow M.D(\bar{X}) = \frac{\sum_{i=1}^{10} |X_i - 6|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^{10} |X_i - 5.5|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{10} |X_i - 5|}{10} = \frac{14}{10} = 1.4$$

2. Find mean deviation about mean, median and mode for the following distributions.(exercise)

Class	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Frequency	7	10	22	15	12	6	3

Remark: Mean deviation is always minimum about the median.

Coefficient of Mean Deviation (C.M.D)

$$C.M.D = \frac{M.D}{\text{Average about which deviations are taken}}$$

$$\Rightarrow C.M.D(\bar{X}) = \frac{M.D(\bar{X})}{\bar{X}}$$

$$C.M.D(\tilde{X}) = \frac{M.D(\tilde{X})}{\tilde{X}}$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}}$$

Example: calculate the C.M.D about the mean, median and mode for the data in example 1 above.

Solutions:

$$C.M.D = \frac{M.D}{\text{Average about which deviations are taken}}$$

$$\Rightarrow C.M.D(\bar{X}) = \frac{M.D(\bar{X})}{\bar{X}} = \frac{1.4}{6} = 0.233$$

$$C.M.D(\tilde{X}) = \frac{M.D(\tilde{X})}{\tilde{X}} = \frac{1.4}{5.5} = 0.255$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}} = \frac{1.4}{5} = 0.28$$

Population Variance

If we divide the variation by the number of values in the population, we get something called the population variance. This variance is the "average squared deviation from the mean".

$$\text{Population Varince} = \sigma^2 = \frac{1}{N} \sum (X_i - \mu)^2, \quad i = 1, 2, \dots, N$$

For the case of frequency distribution it is expressed as:

$$\text{Population Varince} = \sigma^2 = \frac{1}{N} \sum f_i (X_i - \mu)^2, \quad i = 1, 2, \dots, k$$

Sample Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$\text{Sample Varince} = S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2, \quad i = 1, 2, \dots, n$$

For the case of frequency distribution it is expressed as:

$$\text{Sample Varince} = S^2 = \frac{1}{n-1} \sum f_i (X_i - \bar{X})^2, \quad i = 1, 2, \dots, k$$

We usually use the following short cut formula.

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}, \text{ for raw data.}$$

$$S^2 = \frac{\sum_{i=1}^k f_i X_i^2 - n\bar{X}^2}{n-1}, \text{ for frequency distribution.}$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

Populations tan dard deviation = $\sigma = \sqrt{\sigma^2}$

Samples tan dard deviation = $s = \sqrt{S^2}$

The following steps are used to calculate the sample variance

Class	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Frequency	7	10	22	15	12	6	3

1. Find the arithmetic mean.
2. Find the difference between each observation and the mean.
3. Square these differences.
4. Sum the squared differences.
5. Since the data is a sample, divide the number (from step 4 above) by the number of observations minus one, i.e., $n-1$ (where n is equal to the number of observations in the data set).

Examples: Find the variance and standard deviation of the following sample data: 5, 17, 12, 10.

1. The data is given in the form of frequency distribution.

Solutions: $\bar{X} = 11$

X_i	5	10	12	17	Total
$(X_i - \bar{X})^2$	36	1	1	36	74

$$\Rightarrow S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{74}{3} = 24.67.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{24.67} = 4.97.$$

$$\bar{X} = 55$$

$X_i(\text{C.M})$	42	47	52	57	62	67	72	Total
$f_i(X_i - \bar{X})^2$	1183	640	198	60	588	864	867	4400

$$\Rightarrow S^2 = \frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{n-1} = \frac{4400}{74} = 59.46.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{59.46} = 7.71.$$

Special properties of Standard deviations

$$1. \quad \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} < \sqrt{\frac{\sum (X_i - A)^2}{n-1}}, A \neq \bar{X}$$

2. For normal (symmetric distribution) the following holds.

- Approximately 68.27% of the data values fall within one standard deviation of the mean. i.e. within $(\bar{X} - S, \bar{X} + S)$
- Approximately 95.45% of the data values fall within two standard deviations of the mean. i.e. within $(\bar{X} - 2S, \bar{X} + 2S)$
- Approximately 99.73% of the data values fall within three standard deviations of the mean. i.e. within $(\bar{X} - 3S, \bar{X} + 3S)$

3. Chebyshev's Theorem

For any data set, no matter what the pattern of variation, the proportion of the values that fall within k standard deviations of the mean or $(\bar{X} - kS, \bar{X} + kS)$ will be at least

$1 - \frac{1}{k^2}$, where k is a number greater than 1. i.e. the proportion of items falling beyond

k standard deviations of the mean is at most $\frac{1}{k^2}$

Example: Suppose a distribution has mean 50 and standard deviation 6.

What percent of the numbers are:

- | | |
|----------------------|----------------------------------|
| a) Between 38 and 62 | c) Less than 38 or more than 62. |
| b) Between 32 and 68 | d) Less than 32 or more than 68. |

Solutions:

a) 38 and 62 are at equal distance from the mean, 50 and this distance is 12

$$\Rightarrow ks = 12$$

$$\Rightarrow k = \frac{12}{S} = \frac{12}{6} = 2$$

→ Applying theorem above at least $(1 - \frac{1}{k^2}) * 100\% = 75\%$ of the numbers lie between 38

and 62

b) Similarly done.

c) It is just the complement of a) i.e. at most $\frac{1}{k^2} * 100\% = 25\%$ of the numbers lie less than 32 or more than 62.

d) Similarly done.

Exercise 2: The average score of a special test of knowledge of wood refinishing has a mean of 53 and standard deviation of 6. Find the range of values in which at least 75% the scores will lie.

4. If the standard deviation of $X_1, X_2, \dots, X_n$ is S , then the standard deviation of

a) $X_1 + k, X_2 + k, \dots, X_n + k$ will also be S

b) $kX_1, kX_2, \dots, kX_n$ would be $|k|S$

c) $a + kX_1, a + kX_2, \dots, a + kX_n$ would be $|k|S$

Examples:

1. The mean and standard deviation of n Tetracycline Capsules $X_1, X_2, \dots, X_n$ are known to be 12 gm and 3 gm respectively. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ ($i = 1, 2, \dots, n$) then what will be the standard deviation of the new set of capsules
2. The mean and the standard deviation of a set of numbers are respectively 500 and 10.
 - a. If 10 is added to each of the numbers in the set, then what will be the variance and standard deviation of the new set?

- b. If each of the numbers in the set are multiplied by -5, then what will be the variance and standard deviation of the new set?

Solutions:

1. Using c) above the new standard deviation = $|k|S = 2 * 3 = 6$
2. a. They will remain the same.
- c. New standard deviation = $|k|S = 5 * 10 = 50$

Coefficient of Variation (C.V)

- Is defined as the ratio of standard deviation to the mean usually expressed as percents.

$$C.V = \frac{S}{\bar{X}} * 100$$

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

Examples:

1. An analysis of the monthly wages paid (in Birr) to workers in two firms A and B belonging to the same industry gives the following results

Value	Firm A	Firm B
Mean wage	52.5	47.5
Median wage	50.5	45.5
Variance	100	121

In which firm A or B is there greater variability in individual wages?

Solutions:

Calculate coefficient of variation for both firms.

$$C.V_A = \frac{S_A}{\bar{X}_A} * 100 = \frac{10}{52.5} * 100 = 19.05\%$$

$$C.V_B = \frac{S_B}{\bar{X}_B} * 100 = \frac{11}{47.5} * 100 = 23.16\%$$

Since $C.V_A < C.V_B$, in firm B there is greater variability in individual wages.

2. A meteorologist interested in the consistency of temperatures in three cities during a given week collected the following data. The temperatures for the five days of the week in the three cities were

City 1	25	24	23	26	17
City2	22	21	24	22	20
City3	32	27	35	24	28

Which city have the most consistent temperature, based on these data? (Exercise)

Standard Scores (Z-scores)

- If X is a measurement from a distribution with mean $\bar{X}$ and standard deviation S , then its value in standard units is

$$Z = \frac{X - \mu}{\sigma}, \text{ for population.}$$

$$Z = \frac{X - \bar{X}}{S}, \text{ for sample}$$

- Z gives the deviations from the mean in units of standard deviation
- Z 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.

Examples:

1. Two sections were given introduction to statistics examinations. The following information was given.

Value	Section 1	Section 2
Mean	78	90
Stan.deviation	6	5

Student A from section 1 scored 90 and student B from section 2 scored 95. Relatively speaking who performed better?

Solutions:

Calculate the standard score of both students.

$$Z_A = \frac{X_A - \bar{X}_1}{S_1} = \frac{90 - 78}{6} = 2$$

$$Z_B = \frac{X_B - \bar{X}_2}{S_2} = \frac{95 - 90}{5} = 1$$

➔ Student A performed better relative to his section because the score of student A is two standard deviation above the mean score of his section while, the score of student B is only one standard deviation above the mean score of his section.

2. 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
Stan.dev.	1.2 min	1.3 min

Relatively speaking:

- Which group is more consistent in its performance
- 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?

Solutions:

- Use coefficient of variation.

$$C.V_1 = \frac{S_1}{\bar{X}_1} * 100 = \frac{1.2}{10.4} * 100 = 11.54\%$$

$$C.V_2 = \frac{S_2}{\bar{X}_2} * 100 = \frac{1.3}{11.9} * 100 = 10.92\%$$

Since $C.V_2 < C.V_1$, group 2 is more consistent.

- Calculate the standard score of A and B

$$Z_A = \frac{X_A - \bar{X}_1}{S_1} = \frac{9.2 - 10.4}{1.2} = -1$$

$$Z_B = \frac{X_B - \bar{X}_2}{S_2} = \frac{9.3 - 11.9}{1.3} = -2$$

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.

Moments: If X is a variable that assume the values $X_1, X_2, \dots, X_n$ then

1. The r^{th} moment is defined as:

$$\begin{aligned}\bar{X}^r &= \frac{X_1^r + X_2^r + \dots + X_n^r}{n} \\ &= \frac{\sum_{i=1}^n X_i^r}{n}\end{aligned}$$

- For the case of frequency distribution this is expressed as: $\bar{X}^r = \frac{\sum_{i=1}^k f_i X_i^r}{n}$

- If $r=1$, it is the simple arithmetic mean, this is called the first moment.

2. The r^{th} moment about the mean (the r^{th} central moment)

- Denoted by M_r and defined as:

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

For the case of frequency distribution this is expressed as $M_r = \frac{\sum_{i=1}^k f_i (X_i - \bar{X})^r}{n}$

If $r = 2$, it is population variance, this is called the second central moment. If we assume $n-1 \approx n$, it is also the sample variance.

3. The r^{th} moment about any number A is defined as: Denoted by M_r' and

$$M_r' = \frac{\sum_{i=1}^n (X_i - A)^r}{n} = \frac{(n-1)}{n} \frac{\sum_{i=1}^n (X_i - A)^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r' = \frac{\sum_{i=1}^k f_i (X_i - A)^r}{n}$$

Example:

1. Find the first two moments for the following set of numbers 2, 3, 7
2. Find the first three central moments of the numbers in problem 1
3. Find the third moment about the number 3 of the numbers in problem 1.

Solutions:

1. Use the r^{th} moment formula.

$$\begin{aligned}\bar{X}^r &= \frac{\sum_{i=1}^n X_i^r}{n} \\ \Rightarrow \bar{X}^1 &= \frac{2+3+7}{3} = 4 = \bar{X} \\ \bar{X}^2 &= \frac{2^2 + 3^2 + 7^2}{3} = 20.67\end{aligned}$$

2. Use the r^{th} central moment formula.

$$\begin{aligned}M_r &= \frac{\sum_{i=1}^n (X_i - \bar{X})^r}{n} \\ \Rightarrow M_1 &= \frac{(2-4) + (3-4) + (7-4)}{3} = 0 \\ M_2 &= \frac{(2-4)^2 + (3-4)^2 + (7-4)^2}{3} = 4.67 \\ M_3 &= \frac{(2-4)^3 + (3-4)^3 + (7-4)^3}{3} = 6\end{aligned}$$

3. Use the r^{th} moment about A.

$$M_r = \frac{\sum_{i=1}^n (X_i - A)^r}{n}$$
$$\Rightarrow M_3 = \frac{(2-3)^3 + (3-3)^3 + (7-3)^3}{3} = 21$$

Skewness

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size.
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

$$\text{Mean} - \text{Mode} = 3 * (\text{Mean} - \text{Median})$$

Measures of Skewness

- Denoted by α_3
- There are various measures of skewness.
 1. The Pearsonian coefficient of skewness

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

2. The Bowley's coefficient of skewness (coefficient of skewness based on quartiles)

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

3. The moment coefficient of skewness

$$\alpha_3 = \frac{M_3}{M_2^{3/2}} = \frac{M_3}{(\sigma^2)^{3/2}} = \frac{M_3}{\sigma^3}, \text{ Where } \sigma \text{ is the populations standard deviation}$$

The shape of the curve is determined by the value of α_3

- 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.
- 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.

Examples:

1. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?

Solutions:

Use the Pearsonian coefficient of skewness

$$\alpha_3 = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{32 - 30.5}{10} = 0.15$$

$$\alpha_3 > 0 \Rightarrow \text{The distribution is positively skewed.}$$

2. In a frequency distribution, the coefficient of skewness based on the quartiles is given to be 0.5. If the sum of the upper and lower quartile is 28 and the median is 11, find the values of the upper and lower quartiles.

Solutions:

$$\text{Given: } \alpha_3 = 0.5, \quad \tilde{X} = Q_2 = 11$$

$$\text{Required: } Q_1, Q_3$$

$$Q_1 + Q_3 = 28 \dots\dots\dots (*)$$

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = 0.5$$

Substituting the given values, one can obtain the following

$$Q_3 - Q_1 = 12 \dots\dots\dots (**)$$

Solving () and (**) at the same time we obtain the following values*

$$Q_1 = 8 \quad \text{and} \quad Q_3 = 20$$

Exercises

1. Some characteristics of annually family income distribution (in Birr) in two regions is as follows:

Region	Mean	Median	Standard Deviation
A	6250	5100	960
B	6980	5500	940

- a) Calculate coefficient of skewness for each region
 - b) For which region is, the income distribution more skewed. Give your interpretation for this Region
 - c) For which region is the income more consistent?
3. For a moderately skewed frequency distribution, the mean is 10 and the median is 8.5. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness and the probable mode of the distribution.
 4. The sum of fifteen observations, whose mode is 8, was found to be 150 with coefficient of variation of 20%
 - (a) Calculate the pearsonian coefficient of skewness and give appropriate conclusion.
 - (b) Are smaller values more or less frequent than bigger values for this distribution?
 - (c) If a constant k was added on each observation, what will be the new pearsonian coefficient of skewness? Show your steps. What do you conclude from this?

Kurtosis

Kurtosis is the degree of peakness 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.

Measures of kurtosis

The moment coefficient of kurtosis:

- Denoted by α_4 and given by

$$\alpha_4 = \frac{M_4}{M_2^2} = \frac{M_4}{\sigma^4}$$

Where: M_4 is the fourth moment about the mean.

M_2 is the second moment about the mean.

σ is the population standard deviation

The peakness depends on the value of α_4 .

- *If $\alpha_4 > 3$ then the curve is leptokurtic.*
- *If $\alpha_4 = 3$ then the curve is mesokurtic.*
- *If $\alpha_4 < 3$ then the curve is platykurtic.*

Examples:

1. If the first four central moments of a distribution are:

$$M_1 = 0, M_2 = 16, M_3 = -60, M_4 = 162$$

- a) Compute a measure of skewness
- b) Compute a measure of kurtosis and give your interpretation.

Solutions:

$$\text{a) } \alpha_3 = \frac{M_3}{M_2^{3/2}} = \frac{-60}{16^{3/2}} = -0.94 < 0$$

$\Rightarrow$ *The distribution is negatively skewed.*

$$\text{b) } \alpha_4 = \frac{M_4}{M_2^2} = \frac{162}{16^2} = 0.6 < 3$$

$\Rightarrow$ The curve is platykurtic.

Exercise 1: The median and the mode of a mesokurtic distribution are 32 and 34 respectively. The 4th moment about the mean is 243. Compute the Pearsonian coefficient of skewness and identify the type of skewness. Assume $(n-1 = n)$.

Exercise 2: If the standard deviation of a symmetric distribution is 10, what should be the value of the fourth moment so that the distribution is mesokurtic?

4. Elementary Probability

In our daily life, it is not uncommon to hear words which express our doubts or being uncertain about the happenings of certain events. To mention some instances, “if by chance you meet them, please convey my heart-left greeting”, “probably, he might not take the class today”, etc. These statements show uncertainty about the happening of the event under question. In statistics, however, sensible numerical statements can be made about uncertainty and apply different approaches to calculate probabilities.

In this unit, there are two main points to be discussed: possibilities and probabilities, the first part about techniques of counting or the methods used to determine the number of possibilities, which are indispensable to calculate probabilities. The second part is meant to introduce some basic terms in probability, followed by different methods of measuring probabilities and finally, conditional probability and independent events will be presented.

Deterministic vs. non-deterministic (probabilistic or stochastic)

A deterministic model is one in which every set of variable states is uniquely determined by parameters in the model and by sets of previous states of these variables. Hypothesize exact relationships and it will be suitable when prediction error is negligible.

In a non-deterministic (stochastic/probabilistic) model, randomness is present, and variable states are not described by unique values, but rather by probability distributions. Hence, there will be a defined pattern or regularity appears to construct a precise mathematical model. Hypothesize two components, which is deterministic and random error.

Fundamental Principle of Counting and Multiplication

This is the first technique of counting, also called the rule for multiplication of choices, states as follows:

Definition: If an operation consists of two steps of which first can be done in n_1 ways and for each of these the second can be done in n_2 ways, then the whole operation can be done in $n_1 \cdot n_2$ ways.

Here “operation” stands for any kind of procedure, process, or method of selection.

In order to calculate probabilities, we have to know

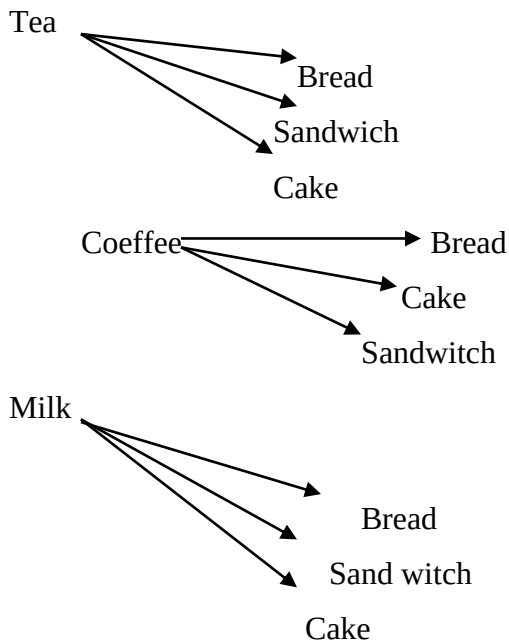
- The number of elements of an event
- The number of elements of the sample space.

That is in order to judge what is **probable**, we have to know what is **possible**.

- In order to determine the number of outcomes, one can use several rules of counting.
 - The addition rule
 - Permutation rule
 - The multiplication rule
 - Combination rule
- To list the outcomes of the sequence of events, a useful device called **tree- diagram** is used.

Example: A student goes to the nearest snack to have a breakfast. He can take tea, coffee, or milk with bread, cake and sandwich. How many possibilities does he have?

Solutions:



→ There are nine possibilities.

Definition : If a choice consists of k steps of which the first can be made in n_1 ways, the second can be made in n_2 ways..., the k^{th} can be made in n_k ways, then the whole choice can be made in $(n_1 * n_2 * \dots * n_k)$ ways.

Example: The digits 0, 1, 2, 3, and 4 are to be used in 4 digit identification card. How many different cards are possible if

a) Repetitions are permitted.

b) Repetitions are not permitted.

Solutions

a.

1 st digit	2 nd digit	3 rd digit	4 th digit
5	5	5	5

There are four steps

1. Selecting the 1st digit, this can be made in 5 ways.
2. Selecting the 2nd digit, this can be made in 5 ways.
3. Selecting the 3rd digit, this can be made in 5 ways.
4. Selecting the 4th digit, this can be made in 5 ways.

$\Rightarrow 5 * 5 * 5 * 5 = 625$ different cards are possible

b.

1 st digit	2 nd digit	3 rd digit	4 th digit
5	4	3	2

There are four steps

2. Selecting the 1st digit, this can be made in 5 ways.
3. Selecting the 2nd digit, this can be made in 4 ways.
4. Selecting the 3rd digit, this can be made in 3 ways.
5. Selecting the 4th digit, this can be made in 2 ways.

$\Rightarrow 5 * 4 * 3 * 2 = 120$ different cards are possible

Note: In selection with replacement (repetition) an object will have more than one chance to be included in the selection process. But in selection without repetition (replacement) an object have will have one chance to be included in the selection process.

Factorial notation: The symbol “n!” which is read as “n factorial” denotes the product of all positive integers less than or equal to n.

Definition: Let n be a non-negative integer. Then, n factorial, denoted by n!, is defined as $n! = n(n-1)(n-2)\dots 3 \cdot 2 \cdot 1$, where $0! = 1! = 1$

Permutations

Frequently, our interest may be to know the different ways in which a group of objects may be ordered or arranged. The fundamental principle of counting and its generalization are used when different choices are made from one set such that their order is important.

An arrangement of n objects in a specified order is called permutation of the objects.

Definition: The arrangement (ordering) of n distinct objects by taking r of them at a time is called permutation, $r \leq n$.

Permutation Rules:

1. The number of permutations of n distinct objects taken all together is n!

$$\text{Where } n! = n * (n-1) * (n-2) * \dots * 3 * 2 * 1$$

2. The arrangement of n objects in a specified order using r objects at a time is called the permutation of n objects taken r objects at a time. It is written as ${}_n P_r$ and the formula is

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

3. The number of permutations of n objects in which k_1 are alike k_2 are alike ---- etc is

$${}_n P_r = \frac{n!}{k_1! * k_2 * \dots * k_n}$$

Definition: The number of permutations of n objects of which n_1 are of one kind, n_2 a second kind, ..., n_k of a k^{th} kind, for $n = n_1 + n_2 + \dots + n_k$ is $n! / n_1! \cdot n_2! \cdot \dots \cdot n_k!$

Example 1: Suppose we have a letters A, B, C, D

- a) How many permutations are there taking all the four?
- b) How many permutations are there two letters at a time?

Example2: How many different permutations can be made from the letters in the word “CORRECTION”?

Example 3: In how many ways can the 10 members of the student council elect a president, a vice president, and a secretary?

Example 4: Determine the number of different permutation of any two of the five vowels a, e, i, o, u, and list all of them.

Solutions:

1.

a)

Here $n = 4$, there are four distinct objects
 $\Rightarrow$ *There are $4! = 24$ permutations.*

b)

Here $n = 4$, $r = 2$

$\Rightarrow$ *There are ${}_4P_2 = \frac{4!}{(4-2)!} = \frac{24}{2} = 12$ permutations.*

2.

Here $n = 10$

Of which 2 are C, 2 are O, 2 are R, 1E, 1T, 1I, 1N

$\Rightarrow K_1 = 2, k_2 = 2, k_3 = 2, k_4 = k_5 = k_6 = k_7 = 1$

Using the 3rd rule of permutation, there are

$$\frac{10!}{2! \cdot 2! \cdot 2! \cdot 1! \cdot 1! \cdot 1! \cdot 1!} = 453600 \text{ permutations.}$$

3. Since there are 10 students to take the first position, we say $n_1 = 10$; after this position is filled, there are 9 students to take the second position, so that $n_2 = 9$; finally, there are $n_3 = 8$ students to take the third position.

Then, it follows that there are altogether $n_1 \cdot n_2 \cdot n_3 = 10 \times 9 \times 8 = 720$ ways in which the students can elect president, vice president, and secretary. In general, if r objects are selected from a set of n distinct objects, any particular arrangement (order) of these objects is called **permutation**.

4. There are $n_1 \cdot n_2 = 5 \times 4 = 20$ different permutations, and they are

Ae, ai, ao, au, ea, ei, eo, eu, ia, ie, io, iu, oa, oe, oi, ou, ua, ue, ui, uo.

Note: The number of permutations of n distinct objects is $n!$.

Combinations

There are many problems in which we are interested in finding the number of ways in which r objects can be selected from n distinct objects **without regard to the order of selection**. Such selections are called **combinations**.

Example 1: Given the letters A, B, C, and D list the permutation and combination for selecting two letters.

Solutions:

Permutation

Combination

AB BA CA DA
AC BC CB DB
AD BD CD DC

AB BC
AC BD
AD DC

Note that in permutation AB is different from BA. But in combination AB is the same as BA.

Example 2: In how many ways can a committee of 2 students be formed out of 6?

Solution: We substitute $n=6$ and $r=2$ to get

$${}_6C_2 = 6! / 2!(6-2)! = 15$$

Combination Rule

The number of combinations of r objects selected from n objects is denoted by

${}_nC_r$ or $\binom{n}{r}$ and is given by the formula:

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

Example 1: In how many ways a committee of 5 people be chosen out of 9 people?

Solutions:

$$n = 9, r = 5$$

$$\binom{n}{r} = \frac{n!}{(n-r)!*r!} = \frac{9!}{4!*5!} = 126 \text{ ways}$$

Example 2: Among 15 clocks there are two defectives .In how many ways can an inspector chose three of the clocks for inspection so that:

- a) There is no restriction.
- b) None of the defective clock is included.
- c) Only one of the defective clocks is included.
- d) Two of the defective clock is included.

Solutions:

$$n = 15 \text{ of which } 2 \text{ are defective and } 13 \text{ are non – defective.}$$
$$r = 3$$

- a) If there is no restriction select three clocks from 15 clocks and this can be done in :

$$n = 15, r = 3$$

$$\binom{n}{r} = \frac{n!}{(n-r)!*r!} = \frac{15!}{12!*3!} = 455 \text{ ways}$$

- b) None of the defective clocks is included.

This is equivalent to zero defective and three non-defective, which can be done in:

$$\binom{2}{0} * \binom{13}{3} = 286 \text{ ways.}$$

- c) Only one of the defective clocks is included.

This is equivalent to one defective and two non-defective, which can be done in:

$$\binom{2}{1} * \binom{13}{2} = 156 \text{ ways.}$$

- d) Two of the defective clock is included.

This is equivalent to two defective and one non-defective, which can be done in:

$$\binom{2}{2} * \binom{13}{1} = 13 \text{ ways.}$$

PROBABILITIES

In section above, we have studied only what is possible in a given situation. In some cases all the possibilities have been listed and in others we simply determine how many different possibilities there are.

In this sub-unit, we shall go one step further and judge what is probable and what is improbable. There are three sub-sections. The first part, will give the meanings of some basic terms in probability. In the second part, different concepts of measuring probabilities will be presented. The last part will study further probabilities like conditional probability and probability of independent events.

In general: The definition and application of probability is defined as follows

- Probability theory is the foundation upon which the logic of inference is built.
- It helps us to cope up with uncertainty.
- In general, probability is the chance of an outcome of an experiment.

Definitions of some probability terms

1. **A Experiment:** Any process of observation or measurement or any process which generates well defined outcome. Like:
 - Flipping (tossing) a fair (balanced) coin and see the face that turns upward.
 - Rolling affair die
2. **Probability Experiment:** It is an experiment that can be repeated any number of times under similar conditions and it is possible to enumerate the total number of outcomes without predicting an individual outcome. It is also called random experiment.

Example: If a fair die is rolled once it is possible to list all the possible outcomes i.e.1, 2, 3, 4, 5, 6 but it is not possible to predict which outcome will occur.

3. **Outcome:** The result of a single trial of a random experiment.
4. **Sample Space:** Set of all possible outcomes of a probability experiment. It is usually denoted by the letter S. Each element of the sample space is called sample point.
5. **Event:** It is a subset of sample space. It is a statement about one or more outcomes of a random experiment .They are denoted by capital letters.

Example: Considering the above experiment let A be the event of odd numbers, B be the event of even numbers, and C be the event of number 8.

$$\Rightarrow A = \{1, 3, 5\}$$

$$B = \{2, 4, 6\}$$

$$C = \{ \} \text{ or empty space or impossible event}$$

Remark: If S (sample space) has n members then there are exactly 2^n subsets or events.

6. **Equally Likely Events:** Events which have the same chance of occurring.

7. **Complement of an Event:** the complement of an event A means non-occurrence of A and is denoted by A' , or A^c , or $\bar{A}$ contains those points of the sample space which don't belong to A .

8. **Elementary Event:** an event having only a single element or sample point.

9. **Mutually Exclusive Events:** Two events which cannot happen at the same time.

10. **Independent Events:** Two events are independent if the occurrence of one does not affect the probability of the other occurring.

11. **Dependent Events:** Two events are dependent if the first event affects the outcome or occurrence of the second event in a way the probability is changed.

12. **Compound event:** This is the sub set containing two or more points in the sample space. It is also called composite event.

13. **Certain event:** This is an event sure to occur.

14. **Simple event:** If an event E consists of a single outcome, then it is called a simple or elementary event.

15. **Impossible event:** This is an event which will never occur.

Example: .What is the sample space for the following experiment

- a) Toss a die one time.
- b) Toss a coin two times.
- c) A light bulb is manufactured. It is tested for its life length by time.

Solution

- a) $S = \{1, 2, 3, 4, 5, 6\}$
- b) $S = \{(HH), (HT), (TH), (TT)\}$
- c) $S = \{t / t \geq 0\}$
 - Sample space can be
 - Countable (finite or infinite)

Example: In an example of rolling a pair die, $S=\{1, 2, 3, 4, 5, 6\}$, each sample point is an equally likely outcome. It is possible to define many events on this sample space as follows:

$A= \{1, 4\}$ __ the event of getting a perfect square number.

$B= \{2, 4, 6\}$ __ the event of an even number.

$C= \{1, 3, 5\}$ _ The event of an odd number.

$D=$ the event of even or odd number.

$E=$ the event of getting number 8.

Then, $A^c= \{2,3, 5,6\}$; B and C are complementary ;

D is certain event; and E is an impossible event.

Example: Roll a fair die and flip balanced coin.

Let $A=$ the die shows an even number, $B=$ the coin shows head.

Then, A and B are independent events.

Different concepts of probability

There is always un certainty whether or not any event will occur. A number between 0 and 1 (inclusive) is assigned to the chance or probability of occurrence of an event. The probability of a certain event is 100% or 1, and the probability of an impossible event is 0. There are three different conceptual approaches to the study of probability theory. These are:

1. The classical approach

This approach is used when:

- All outcomes are equally likely.
- Total number of outcome is finite, say N.

Definition: If a random experiment with N equally likely outcomes is conducted and out of these N_A outcomes are favorable to the event A, then the probability that event A occur denoted $P(A)$ is defined as:

$$P(A) = \frac{N_A}{N} = \frac{\text{No. of outcomes favourable to A}}{\text{Total number of outcomes}} = \frac{n(A)}{n(S)}$$

Examples:

1. A fair die is tossed once. What is the probability of getting
 - a) Number 4?
 - b) An odd number?
 - c) An even number?
 - d) Number 8?

e) A 3 or 5 will turn up

Solutions:

First identify the sample space, say S

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$\Rightarrow N = n(S) = 6$$

a) Let A be the event of number 4

$$A = \{4\}$$

$$\Rightarrow N_A = n(A) = 1$$

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

b) Let A be the event of odd numbers

$$A = \{1, 3, 5\}$$

$$\Rightarrow N_A = n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = 3/6 = 0.5$$

c) Let A be the event of even numbers

$$A = \{2, 4, 6\}$$

$$\Rightarrow N_A = n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = 3/6 = 0.5$$

d) Let A be the event of number 8

$$A = \emptyset$$

$$\Rightarrow N_A = n(A) = 0$$

$$P(A) = \frac{n(A)}{n(S)} = 0/6 = 0$$

e) $S = \{1, 2, 3, 4, 5, 6\}$; let $E = \{3, 5\}$. For a fair die, $p(1)=p(2)=\dots=p(6)$; then $p(E) = m/n = 2/6 = 1/3$.

2. A box of 80 candles consists of 30 defective and 50 non defective candles. If 10 of this candles are selected at random, what is the probability
- a) All will be defective.
 - b) 6 will be non-defective
 - c) All will be non-defective

Solutions:

$$\text{Total selection} = \binom{80}{10} = N = n(S)$$

- a) Let A be the event that all will be defective.

$$\text{Total way in which A occur} = \binom{30}{10} * \binom{50}{0} = N_A = n(A)$$

$$\Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{\binom{30}{10} * \binom{50}{0}}{\binom{80}{10}} = 0.00001825$$

- b) Let A be the event that 6 will be non-defective.

$$\text{Total way in which A occur} = \binom{30}{4} * \binom{50}{6} = N_A = n(A)$$

$$\Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{\binom{30}{4} * \binom{50}{6}}{\binom{80}{10}} = 0.265$$

- c) Let A be the event that all will be non-defective.

$$\begin{aligned} \text{Total way in which } A \text{ occur} &= \binom{30}{0} * \binom{50}{10} = N_A = n(A) \\ \Rightarrow P(A) &= \frac{n(A)}{n(S)} = \frac{\binom{30}{0} * \binom{50}{10}}{\binom{80}{10}} = 0.00624 \end{aligned}$$

Short coming of the classical approach:

This approach is not applicable when:

- The total number of outcomes is infinite.
- Outcomes are not equally likely.

2. The Frequentist Approach

This is based on the relative frequencies of outcomes belonging to an event.

Definition: The probability of an event A is the proportion of outcomes favorable to A in the long run when the experiment is repeated under same condition.

Definition: The frequency concept of probability

If there are n repetitions of an experiment, where n is very large, and event A is observed to occur in m of these, again the probability of A is $p(A) = m/n$.

Note that the proportion of times an event occurs is nothing but its relative frequency, since, in this relative frequency concept, the probability is obtained by repetitive empirical observation, it also known as “Empirical probability” or a statistical approach.

Example: If the records of Ethiopian airlines show that 468 of 600 of its flights from B/DAR to Addis arrived on time, what is the probability that any one of similar flights will arrive on time?

Solution: Let E= the event that the plane will arrive on time.

$$P(E) = m/n = 468/600 = 0.78.$$

Note: the probability of not arriving on time is

$$P(E^c) = n-m/n = 600-468/600 = 1-468/600 = 1-0.78 = 0.22$$

That is, the plane did not arrive on time for $600-468 = 132$ times,

$$\text{And } p(E^c) = 132/600 = 0.22.$$

In general, $p(E^c) = 1-p(E)$ or $p(E) + p(E^c) = 1$.

Example: If records show that 60 out of 100,000 bulbs produced are defective. What is the probability of a newly produced bulb to be defective?

Solution: Let A be the event that the newly produced bulb is defective.

$$P(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N} = \frac{60}{100,000} = 0.0006$$

3. Axiomatic Approach:

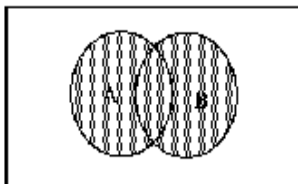
Let E be a random experiment and S be a sample space associated with E. With each event A a real number called the probability of A satisfies the following properties called axioms of probability or postulates of probability.

1. $P(A) \geq 0$
2. $P(S) = 1$, *S is the sure event.*
3. If A and B are mutually exclusive events, the probability that one or the other occur equals the sum of the two probabilities. i. e.

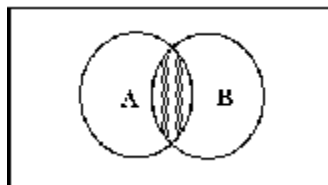
$$P(A \cup B) = P(A) + P(B)$$

4. $P(A') = 1 - P(A)$
5. $0 \leq P(A) \leq 1$
6. $P(\emptyset) = 0$, $\emptyset$ is the impossible event.

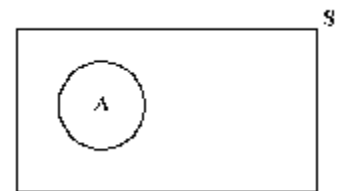
Remark: Venn-diagrams can be used to solve probability problems.



$A \cup B$



$A \cap B$



A

In general $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

5. Conditional probability and Independency

Conditional Events: If the occurrence of one event has an effect on the next occurrence of the other event then the two events are conditional or dependent events.

Example: Suppose we have two red and three white balls in a bag

1. Draw a ball with replacement

Let A= the event that the first draw is red $\rightarrow p(A) = \frac{2}{5}$

B= the event that the second draw is red $\rightarrow p(B) = \frac{2}{5}$, then A and B are independent.

2. Draw a ball without replacement

Let A= the event that the first draw is red $\rightarrow p(A) = \frac{2}{5}$

B= the event that the second draw is red $\rightarrow p(B) = ?$, This is conditional.

Let B= the event that the second draw is red given that the first draw is red $\rightarrow p(B) = 1/4$

Conditional probability of an event

The conditional probability of event A given that B has already occurred, denoted $p(A/B)$ is

$$p(A/B) = \frac{p(A \cap B)}{p(B)}, \quad p(B) \neq 0$$

Remark: (1) $p(A'/B) = 1 - p(A/B)$

$$(2) p(B'/A) = 1 - p(B/A)$$

Examples

1. For a student enrolling at freshman at certain university the probability is 0.25 that he/she will get scholarship and 0.75 that he/she will graduate. If the probability is 0.2 that he/she will get scholarship and will also graduate. What is the probability that a student who get a scholarship graduate?

Solution: Let A= the event that a student will get a scholarship

B= the event that a student will graduate

given $p(A) = 0.25$, $p(B) = 0.75$, $p(A \cap B) = 0.20$

Required $p(B/A)$

$$p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{0.20}{0.25} = 0.80$$

2. If the probability that a research project will be well planned is 0.60 and the probability that it will be well planned and well executed is 0.54, what is the probability that it will be well executed given that it is well planned?

Solution; Let A= the event that a research project will be well

Planned

B= the event that a research project will be well planned and well

Executed

given $p(A) = 0.60$, $p(A \cap B) = 0.54$

Required $p(B/A)$

$$p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{0.54}{0.60} = 0.90$$

Note; for any two events A and B the following relation holds.

$$p(B) = p(B/A).p(A) + p(B/A').p(A')$$

Probability of Independent Events

Two events A and B are independent if and only if $p(A \cap B) = p(A).p(B)$

Here $p(A/B) = p(A)$, $p(B/A) = p(B)$

Example; A box contains four black and six white balls. What is the probability of getting two black balls in drawing one after the other under the following conditions?

- a. The first ball drawn is not replaced
- b. The first ball drawn is replaced

Solution; Let A= first drawn ball is black

B= second drawn is black

Required $p(A \cap B)$

- a. $p(A \cap B) = p(B/A).p(A) = (4/10)(3/9) = 2/15$

$$b. \quad p(A \cap B) = p(A) \cdot p(B) = (4/10)(4/10) = 4/25$$

Multiplication Rule

Multiplication rule (Multiplicative Theorem)

For any n events $A_1, A_2, \dots, A_n$ with $P(\bigcap_{j=1}^{n-1} A_j) > 0$ it holds

$$P\left(\bigcap_{j=1}^n A_j\right) = P(A_n | A_1 \cap \dots \cap A_{n-1})P(A_{n-1} | A_1 \cap \dots \cap A_{n-2}) \dots P(A_2 | A_1)P(A_1).$$

Its significance is that we can calculate the probability of the intersection of n events, step by step, by means of conditional probabilities. The calculation of these conditional probabilities is far easier. Here is a simple example which simply illustrates the point.

Let $B_1, B_2, \dots, B_n$ be a partition of the sample space S ,

- (i) $B_i \cap B_j = \emptyset$ for $i \neq j$ where $i, j = 1, 2, 3, \dots, n$
- (ii) $\bigcup_{i=1}^n B_i = S$
- (iii) $P(B_i) > 0$ for all i .

Then, for any event A in S ,

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + \dots + P(A \cap B_n) \\ &= P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + \dots + P(B_n) \cdot P(A|B_n) \end{aligned}$$

Example: The members of a consulting firm rent cars from three rental agencies: 60% from agency I, 30% from agency II, and 10% from agency III. If 9% of the cars from agency I need a tune up, 20% of the cars from agency II need a tune-up, and 6% of the cars from agency III need a tune-up, what is the probability that a rental car delivered to the firm will need a tune-up?

Solution: If A is the event that the car needs a tune-up, and B_1, B_2 , and B_3 are the events that the car comes from rental agencies I, II, or III, we have $P(B_1)=0.60$, $P(B_2)=0.30$, $P(B_3)=0.10$, $P(A|B_1)=0.09$, $P(A|B_2)=0.20$, and $P(A|B_3)=0.06$. Then, using the rule of total probability,

$$\begin{aligned} P(A) &= P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3) \\ &= (0.60)(0.09) + (0.30)(0.20) + (0.10)(0.06) = 0.12. \end{aligned}$$

Thus, 12% of all the rental cars delivered to this firm will need a tune-up.

Baye`s Theorem

Suppose that $B_1, B_2, \dots, B_n$ are mutually exclusive events whose union is the sample space S . Then if A is any event, we have the following important theorem:

$$P(B_i | A) = \frac{P(B_i) P(A | B_i)}{\sum_{j=1}^n P(B_j) P(A | B_j)}$$

Example 1: With reference to example 1, if a rental car delivered to the consulting firm needs a tune-up, the probability that it came from rental agency II is:

$$P(B_2/A) = \frac{P(B_2) * P(A/B_2)}{P(A)} = \frac{(0.30)(0.20)}{0.120} = 0.5.$$

5. RANDOM VARIABLES

Definition: A random variable is a numerical description of the outcomes of the experiment or a numerical valued function defined on sample space, usually denoted by capital letters.

Example: If X is a random variable, then it is a function from the elements of the sample space to the set of real numbers. i.e.

X is a function $X: S \rightarrow R$

→ A random variable takes a possible outcome and assigns a number to it.

Example: Flip a coin three times, let X be the number of heads in three tosses.

$$\Rightarrow S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$$

$$\Rightarrow X(HHH) = 3,$$

$$X(HHT) = X(HTH) = X(THH) = 2,$$

$$X(HTT) = X(THT) = X(TTH) = 1$$

$$X(TTT) = 0$$

$$X = \{0, 1, 2, 3\}$$

→ X assumes a specific number of values with some probabilities.

Random variables are of two types:

1. Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted

Examples:

- Toss coin n times and count the number of heads.
 - Number of children in a family.
 - Number of car accidents per week.
 - Number of defective items in a given company.
 - Number of bacteria per two cubic centimeter of water
2. Continuous random variable: are variables that can assume all values between any two values.

Examples:

- Height of students at certain college.
- Mark of a student.
- Life time of light bulbs.
- Length of time required to complete a given training.

Definition: a probability distribution consists of a value a random variable can assume and the corresponding probabilities of the values.

Example: Consider the experiment of tossing a coin three times. Let X be the number of heads. Construct the probability distribution of X .

Solution:

- First identify the possible value that X can assume.
- Calculate the probability of each possible distinct value of X and express X in the form of frequency distribution.

$X = x$	0	1	2	3
$P(X = x)$	$1/8$	$3/8$	$3/8$	$1/8$

Probability distribution is denoted by P for discrete and by f for continuous random variable.

Properties of Probability Distribution:

1. $P(x) \geq 0$, and $\sum_x P(X = x) = 1$ if X is discrete

2. $f(x) \geq 0$, and $\int_x f(x)dx = 1$, if X is continuous

3. If X is a continuous random variable then $P(a < X < b) = \int_a^b f(x)dx$

4. Probability of a fixed value of a continuous random variable is zero.

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

1. If X is discrete random variable the

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

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

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

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

2. Probability means area for continuous random variable.

Introduction to expectation

1. Let a discrete random variable X assume the values $X_1, X_2, \dots, X_n$ with the probabilities $P(X_1), P(X_2), \dots, P(X_n)$ respectively. Then the expected value of X , denoted as $E(X)$ is defined as:

$$E(X) = X_1P(X_1) + X_2P(X_2) + \dots + X_nP(X_n) = \sum_{i=1}^n X_iP(X_i)$$

2. Let X be a continuous random variable assuming the values in the interval (a, b) such

$$\text{that } \int_a^b f(x)dx = 1, \text{ then } E(X) = \int_a^b x f(x)dx$$

Examples: What is the expected value of a random variable X obtained by tossing a coin three times where is the number of heads

Solution:

First construct the probability distribution of X

$X = x$	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

$$\begin{aligned} \Rightarrow E(X) &= X_1P(X_1) + X_2P(X_2) + \dots + X_nP(X_n) \\ &= 0 \cdot 1/8 + 1 \cdot 3/8 + \dots + 2 \cdot 1/8 = 1.5 \end{aligned}$$

Example: Suppose a charity organization is mailing printed return-address stickers to over one million homes in the Ethiopia. Each recipient is asked to donate either \$1, \$2, \$5, \$10, \$15, or \$20. Based on past experience, the amount a person donates is believed to follow the following probability distribution:

$X = x$	\$1	\$2	\$5	\$10	\$15	\$20
$P(X = x)$	0.1	0.2	0.3	0.2	0.15	0.05

What is expected that an average donor to contribute?

Solution:

$X = x$	\$1	\$2	\$5	\$10	\$15	\$20	Total
$P(X = x)$	0.1	0.2	0.3	0.2	0.15	0.05	1
$xP(X = x)$	0.1	0.4	1.5	2	2.25	1	7.25

$$\Rightarrow E(X) = \sum_{i=1}^6 x_i P(X = x_i) = \$7.25$$

Mean and Variance of a random variable

Let X be given random variable.

1. The expected value of X is its mean

$$\Rightarrow \text{Mean of } X = E(X)$$

2. The variance of X is given by:

$$\text{Variance of } X = \text{var}(X) = E(X^2) - [E(X)]^2$$

Where:

$$E(X^2) = \sum_{i=1}^n x_i^2 P(X = x_i) \text{ , if } X \text{ is discrete}$$

$$= \int x^2 f(x) dx \text{ , if } X \text{ is continuous}$$

Example 1: Find the mean and the variance of a random variable X in example 2 above.

Solutions:

$X = x$	\$1	\$2	\$5	\$10	\$15	\$20	Total
$P(X = x)$	0.1	0.2	0.3	0.2	0.15	0.05	1
$xP(X = x)$	0.1	0.4	1.5	2	2.25	1	7.25
$x^2 P(X = x)$	0.1	0.8	7.5	20	33.75	20	82.15

$$\Rightarrow E(X) = 7.25$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 82.15 - 7.25^2 = 29.59$$

Exercise: Two dice are rolled. Let X be a random variable denoting the sum of the numbers on the two dice.

- i) Give the probability distribution of X
- ii) Compute the expected value of X and its variance

There are some general rules for mathematical expectation.

Let X and Y are random variables and k be a constant.

RULES

$$Var(k) = 0$$

$$E(k) = k$$

$$E(kX) = kE(X)$$

$$Var(kX) = k^2 Var(X)$$

$$E(X + Y) = E(X) + E(Y)$$

6. Common probability Distributions

Common Discrete Probability Distributions

1. Binomial Distribution

A binomial experiment is a probability experiment that satisfies the following four requirements called assumptions of a binomial distribution.

1. The experiment consists of n identical trials.
2. Each trial has only one of the two possible mutually exclusive outcomes, success or a failure.
3. The probability of each outcome does not change from trial to trial, and
4. The trials are independent, thus we must sample with replacement.

Examples of binomial experiments

- Tossing a coin 20 times to see how many tails occur.
- Asking 200 people if they watch BBC news.
- Registering a newly produced product as defective or non-defective.
- Asking 100 people if they favour the ruling party.
- Rolling a die to see if a 5 appears.

Definition: The outcomes of the binomial experiment and the corresponding probabilities of these outcomes are called Binomial Distribution.

Let $P =$ the probability of success

$q = 1 - p =$ the probability of failure on any given trial

Then the probability of getting X successes in n trials becomes:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n$$

And this is sometimes written as: $X \sim \text{Bin}(n, p)$

When using the binomial formula to solve problems, we have to identify three things:

- The number of trials (n)
- The probability of a success on any one trial (p) and
- The number of successes desired (X).

Example1: What is the probability of getting three heads by tossing a fair coin four times?

Solution: Let X be the number of heads in tossing a fair coin four times

$$X \sim \text{Bin}(n = 4, p = 0.50)$$

$$\begin{aligned} \Rightarrow P(X = x) &= \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, 3, 4 \\ &= \binom{4}{x} 0.5^x 0.5^{4-x} \\ &= \binom{4}{x} 0.5^4 \\ \Rightarrow P(X = 3) &= \binom{4}{3} 0.5^4 = 0.25 \end{aligned}$$

1. Suppose that an examination consists of six true and false questions, and assume that a student has no knowledge of the subject matter. The probability that the student will guess the correct answer to the first question is 30%. Likewise, the probability of guessing each of the remaining questions correctly is also 30%.
 - a. What is the probability of getting more than three correct answers?
 - b. What is the probability of getting at least two correct answers?
 - c. What is the probability of getting at most three correct answers?
 - d. What is the probability of getting less than five correct answers?

Solution

Let X = the number of correct answers that the student gets.

$$X \sim \text{Bin}(n = 6, p = 0.30)$$

a) $P(X > 3) = ?$

$$\begin{aligned} \Rightarrow P(X = x) &= \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, 6 \\ &= \binom{6}{x} 0.3^x 0.7^{6-x} \end{aligned}$$

$$\begin{aligned} \Rightarrow P(X > 3) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= 0.060 + 0.010 + 0.001 \\ &= 0.071 \end{aligned}$$

Thus, we may conclude that if 30% of the exam questions are answered by guessing, the probability is 0.071 (or 7.1%) that more than four of the questions are answered correctly by the student.

b) $P(X \geq 2) = ?$

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) \\ &= 0.324 + 0.185 + 0.060 + 0.010 + 0.001 = 0.58 \end{aligned}$$

c) $P(X \leq 3) = ?$

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.118 + 0.303 + 0.324 + 0.185 = 0.93 \end{aligned}$$

d) $P(X < 5) = ?$

$$\begin{aligned} P(X < 5) &= 1 - P(X \geq 5) \\ &= 1 - \{P(X = 5) + P(X = 6)\} \\ &= 1 - (0.010 + 0.001) = 0.989 \end{aligned}$$

Remark: If X is a binomial random variable with parameters n and p then

$$E(X) = np, \text{Var}(X) = npq$$

2. Poisson Distribution

- A random variable X is said to be a Poisson distribution if its probability distribution is given by:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots$$

Where λ = the average number.

- The Poisson distribution depends only on the average number of occurrences per unit time of space.
- The Poisson distribution is used as a distribution of rare events, such as:
 - Number of misprints.
 - Hereditary.
 - Natural disasters like earth quake.
 - Arrivals
 - Accidents.
- The process that gives rise to such events is called Poisson process.

Example: If 1.6 accidents can be expected at an intersection on any given day, what is the probability that there will be 3 accidents on any given day?

Solution; Let X = the number of accidents, $\lambda = 1.6$

$$X = \text{poisson}(1.6) \Rightarrow p(X = x) = \frac{1.6^x e^{-1.6}}{x!}$$

$$p(X = 3) = \frac{1.6^3 e^{-1.6}}{3!} = 0.1380$$

1. On the average, five smokers pass a certain street corner every ten minutes, what is the probability that during a given 10 minutes the number of smokers passing will be
 - a. 6 or fewer
 - b. 7 or more
 - c. Exactly 8.....(Exercise)

If X is a Poisson random variable with parameters λ then

$$E(X) = \lambda \text{ and } Var(X) = \lambda$$

Note: The Poisson probability distribution provides a close approximation to the binomial probability distribution when n is large and p is quite small or quite large with $\lambda = np$.

$$P(X = x) = \frac{(np)^x e^{-(np)}}{x!}, \quad x = 0, 1, 2, \dots$$

Where $\lambda = np = \text{the average number}$.

Usually we use this approximation if $np \leq 5$. In other words, if $n > 20$ and $np \leq 5$ [or $n(1-p) \leq 5$], then we may use Poisson distribution as an approximation to binomial distribution.

Example: Find the binomial probability $P(X=3)$ by using the Poisson distribution if $p = 0.01$ and $n = 200$

Solution:

Using Poisson, $\lambda = np = 0.01 * 200 = 2$

$$\Rightarrow P(X = 3) = \frac{2^3 e^{-2}}{3!} = 0.1804$$

Using Binomial, $n = 200, p = 0.01$

$$\Rightarrow P(X = 3) = \binom{200}{3} (0.01)^3 (0.99)^{197} = 0.1814$$

Common Continuous Probability Distributions

1. Normal Distribution

A random variable X is said to have a normal distribution if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

Where $\mu = E(X)$, $\sigma^2 = \text{Variance}(X)$

μ and σ^2 are the Parameters of the Normal Distribution.

Properties of Normal Distribution:

1. It is bell shaped and is symmetrical about its mean and it is mesokurtic. The maximum ordinate is at $x = \mu$ and is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}$$

2. It is asymptotic to the axis, i.e., it extends indefinitely in either direction from the mean.
3. It is a continuous distribution.
4. It is a family of curves, i.e., every unique pair of mean and standard deviation defines a different normal distribution. Thus, the normal distribution is completely described by two parameters: mean and standard deviation.
5. Total area under the curve sums to 1, i.e., the area of the distribution on each side of the

$$\text{mean is } 0.5. \Rightarrow \int_{-\infty}^{\infty} f(x)dx = 1$$

6. It is unimodal, i.e., values mound up only in the center of the curve.
7. $\text{Mean} = \text{Median} = \text{mode} = \mu$
8. The probability that a random variable will have a value between any two points is equal to the area under the curve between those points.

Note: To facilitate the use of normal distribution, the following distribution known as the standard normal distribution was derived by using the transformation

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

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

Properties of the Standard Normal Distribution:

Same as a normal distribution, but also...

- Mean is zero
 - Variance is one
 - Standard Deviation is one
- Areas under the standard normal distribution curve have been tabulated in various ways.

The most common ones are the areas between

$Z = 0$ and a positive value of Z .

- Given a normal distributed random variable X with

Mean μ and standard deviation σ

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

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

Note:

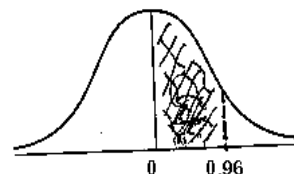
$$\begin{aligned} P(a < X < b) &= P(a \leq X < b) \\ &= P(a < X \leq b) \\ &= P(a \leq X \leq b) \end{aligned}$$

Examples:

1. Find the area under the standard normal distribution which lies

- a) Between $Z = 0$ and $Z = 0.96$

Solution:

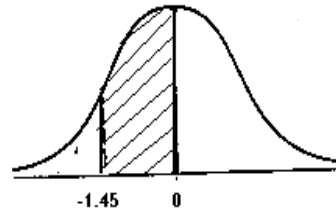


$$\text{Area} = P(0 < Z < 0.96) = 0.3315$$

b) Between $Z = -1.45$ and $Z = 0$

Solution:

$$\begin{aligned}\text{Area} &= P(-1.45 < Z < 0) \\ &= P(0 < Z < 1.45) \\ &= 0.4265\end{aligned}$$



c) To the right of $Z = -0.35$

Solution:

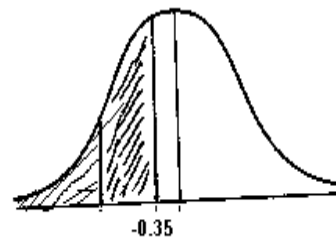
$$\begin{aligned}\text{Area} &= P(Z > -0.35) \\ &= P(-0.35 < Z < 0) + P(Z > 0) \\ &= P(0 < Z < 0.35) + P(Z > 0) \\ &= 0.1368 + 0.50 = 0.6368\end{aligned}$$



d) To the left of $Z = -0.35$

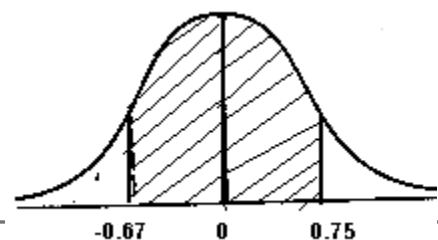
Solution:

$$\begin{aligned}\text{Area} &= P(Z < -0.35) \\ &= 1 - P(Z > -0.35) \\ &= 1 - 0.6368 = 0.3632\end{aligned}$$



e) Between $Z = -0.67$ and $Z = 0.75$

Solution:



$$\begin{aligned}
 \text{Area} &= P(-0.67 < Z < 0.75) \\
 &= P(-0.67 < Z < 0) + P(0 < Z < 0.75) \\
 &= P(0 < Z < 0.67) + P(0 < Z < 0.75) \\
 &= 0.2486 + 0.2734 = 0.5220
 \end{aligned}$$

f) Between $Z = 0.25$ and $Z = 1.25$

Solution:

$$\begin{aligned}
 \text{Area} &= P(0.25 < Z < 1.25) \\
 &= P(0 < Z < 1.25) - P(0 < Z < 0.25) \\
 &= 0.3934 - 0.0987 = 0.2957
 \end{aligned}$$



2. Find the value of Z if

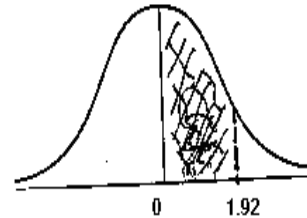
a) The normal curve area between 0 and z(positive) is 0.4726

Solution

$P(0 < Z < z) = 0.4726$ and from table

$P(0 < Z < 1.92) = 0.4726$

$\Leftrightarrow z = 1.92 \dots \text{uniqueness of Area.}$



b) The area to the left of z is 0.9868

Solution

$$\begin{aligned}
 P(Z < z) &= 0.9868 \\
 &= P(Z < 0) + P(0 < Z < z) \\
 &= 0.50 + P(0 < Z < z) \\
 \Rightarrow P(0 < Z < z) &= 0.9868 - 0.50 = 0.4868 \\
 &\text{and from table} \\
 P(0 < Z < 2.2) &= 0.4868 \\
 \Leftrightarrow z &= 2.2
 \end{aligned}$$

3. A random variable X has a normal distribution with mean 80 and standard deviation 4.8. What is the probability that it will take a value
- a) Less than 87.2
 - b) Greater than 76.4
 - c) Between 81.2 and 86.0

Solution

X is normal with mean, $\mu = 80$, standard deviation, $\sigma = 4.8$

$$\begin{aligned} P(X < 87.2) &= P\left(\frac{X - \mu}{\sigma} < \frac{87.2 - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{87.2 - 80}{4.8}\right) \\ \text{a) } &= P(Z < 1.5) \\ &= P(Z < 0) + P(0 < Z < 1.5) \\ &= 0.50 + 0.4332 = \underline{0.9332} \end{aligned}$$

$$\begin{aligned} P(X > 76.4) &= P\left(\frac{X - \mu}{\sigma} > \frac{76.4 - \mu}{\sigma}\right) \\ \text{b) } &= P\left(Z > \frac{76.4 - 80}{4.8}\right) \\ &= P(Z > -0.75) \\ &= P(Z > 0) + P(0 < Z < 0.75) \\ &= 0.50 + 0.2734 = \underline{0.7734} \end{aligned}$$

$$\begin{aligned} P(81.2 < X < 86.0) &= P\left(\frac{81.2 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{86.0 - \mu}{\sigma}\right) \\ \text{c) } &= P\left(\frac{81.2 - 80}{4.8} < Z < \frac{86.0 - 80}{4.8}\right) \\ &= P(0.25 < Z < 1.25) \\ &= P(0 < Z < 1.25) - P(0 < Z < 1.25) \\ &= 0.3934 - 0.0987 = \underline{0.2957} \end{aligned}$$

4. A normal distribution has mean 62.4. Find its standard deviation if 20.05% of the area under the normal curve lies to the right of 72.9

Solution

$$\begin{aligned}P(X > 72.9) &= 0.2005 \Rightarrow P\left(\frac{X - \mu}{\sigma} > \frac{72.9 - \mu}{\sigma}\right) = 0.2005 \\&\Rightarrow P\left(Z > \frac{72.9 - 62.4}{\sigma}\right) = 0.2005 \\&\Rightarrow P\left(Z > \frac{10.5}{\sigma}\right) = 0.2005 \\&\Rightarrow P\left(0 < Z < \frac{10.5}{\sigma}\right) = 0.50 - 0.2005 = 0.2995 \\&\text{And from table } P(0 < Z < 0.84) = 0.2995 \\&\Leftrightarrow \frac{10.5}{\sigma} = 0.84 \\&\Rightarrow \sigma = \underline{\underline{12.5}}\end{aligned}$$

5. A random variable has a normal distribution with $\sigma = 5$. Find its mean if the probability that the random variable will assume a value less than 52.5 is 0.6915.

Solution

$$\begin{aligned}P(Z < z) &= P\left(Z < \frac{52.5 - \mu}{5}\right) = 0.6915 \\&\Rightarrow P(0 < Z < z) = 0.6915 - 0.50 = 0.1915. \\&\text{But from the table} \\&\Rightarrow P(0 < Z < 0.5) = 0.1915 \\&\Leftrightarrow z = \frac{52.5 - \mu}{5} = 0.5 \\&\Rightarrow \mu = \underline{\underline{50}}\end{aligned}$$