BASIC STATISTICS

TQM
for better Future

Foundation for Problem Solving

BASIC STATISTICS

TQM for better Future

Contents:

1.
Basic Statistics and Scale of Measurement

Basic Statistics

Variable

Scale of Measurement

What is Meant by Statistics?

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

Who Uses Statistics?

Statistical techniques are used extensively by marketing, accounting, quality control, consumers, professional sports people, hospital administrators, educators, politicians, physicians, etc...

Why Statistics?

- We face numerical data all the time
  newspapers, sports magazines, business magazines
- Statistical techniques are used to make decisions that effect our daily lives
- Insurance companies use statistics to set rates for different insurance
- City corporation may want to examine the contamination level of different lake water
- Medical professional assess the performance of a new drug over the existing one
- **No matter what is your future line of work, you will make decisions that will involve data**

Types of Statistics

Descriptive Statistics:

Methods of organizing, summarizing, and presenting data in an informative way.

- EXAMPLE 1: A social survey found that 49% of the people knew the name of the first book of the Bible. The statistic 49 describes the number out of every 100 persons who knew the answer.

- EXAMPLE 2: According to Consumer Reports, Whirlpool washing machine owners reported 9 problems per 100 machines during 1995. The statistic 9 describes the number of problems out of every 100 machines.

Types of Statistics (Cont..)

Inferential statistics:

Taking a sample from a population and making estimates/ assumptions about a population, based on a sample.

- A population is a collection of all possible individuals, objects, or measurements of interest.
- A sample is a representative portion or part of the population of interest.

Types of Statistics (Cont.....) (examples of inferential statistics)

- EXAMPLE 1: TV networks constantly monitor the popularity of their programs by hiring research organizations to sample the preferences of TV viewers.
- EXAMPLE 2: The accounting department of a large firm will select a sample of the invoices to check for accuracy for all the invoices of the company.
- EXAMPLE 3: Wine tasters sip a few drops of wine to make a decision with respect to all the wine waiting to be released for sale.

Types of Variables

Variable is a characteristics that can assume any set of prescribed values

Age, Height, Weight, Eye color, Total population of a country

- **Qualitative Variable (Attribute)**
- **Quantitative Variable**

Types of Variables (Cont....)

Qualitative variable or Attribute: the characteristic or variable being studied is non-numeric.

EXAMPLES: Gender, Religious affiliation, Type of automobile owned, Eye color.

Types of Variables (Cont.....)

Quantitative variable: the variable which can be measured and reported numerically.

EXAMPLE: Balance in your savings account, Time remaining in class, Number of children in a family.

Types of Variables (Cont....)

Quantitative variables can be classified as either discrete or continuous.

Discrete variables: can only assume certain values and there are usually "gaps" between values.

EXAMPLE: the number of bedrooms in a house. (1,2,3,..., etc...).

Types of Variables (Cont...)

Quantitative Variables can be classified as either discrete or continuous.

Continuous variables: can assume any value within a specific range.

EXAMPLE: The time it takes to fly from Florida to New York.

Scales of Measurement

Measurement means assigning numbers or other symbols to characteristics of objects according to certain prescribed rules.

Scaling is the generation of a continuum upon which measured objects are located

Arithmetic and statistical operations for summarizing and presenting data depend on the levels of measurement

Ex. Average height of Bangladeshi male, Proportion of smokers in a community

Four scales of measurement

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

Scales of Measurement –Nominal

Nominal level (scaled):

Data that can only be classified into categories and cannot be arranged in an ordering scheme.

EXAMPLES: eye color, gender, religious affiliation.

Scales of Measurement-Nominal (Cont...)

- Categories must be mutually exclusive and exhaustive
- **Mutually exclusive:** Categories are so defined that each member of the population is correctly allocated to one and only one category
- **Exhaustive:** each person, object, or item must be classified in at least one category i.e. A classification is exhaustive when it provides sufficient categories to accommodated all members of the population

Scales of Measurement- Ordinal

Ordinal level:

involves data that may be arranged in some order, but differences between data values cannot be determined or are meaningless.

EXAMPLE: During a taste test of 4 colas, cola C was ranked number 1, cola B was ranked number 2, cola A was ranked number 3, and cola D was ranked number 4.

Scales of Measurement- Interval

Interval Level:

similar to the ordinal level, with the additional property that meaningful amounts of differences between data values can be determined.

There is no natural zero point.

EXAMPLE: Temperature on the Fahrenheit scale.

Scales of Measurement- Interval (Cont.....)

- Includes all the characteristics of ordinal level
- Differences between the values is a constant size
- Examples : temperature in Centigrade, calendar year
  - Difference between 96^0C and 98^0C is equal to the difference between 100^0C and 102^0C
- 0^0 C does not mean the 'no temperature!'
- Cannot say place A (with temperature 100^0C) is 10 times more warmer than place B (with temperature 10^0C)
- No absolute zero, zero is arbitrary

Scales of Measurement- Ratio

Ratio level:

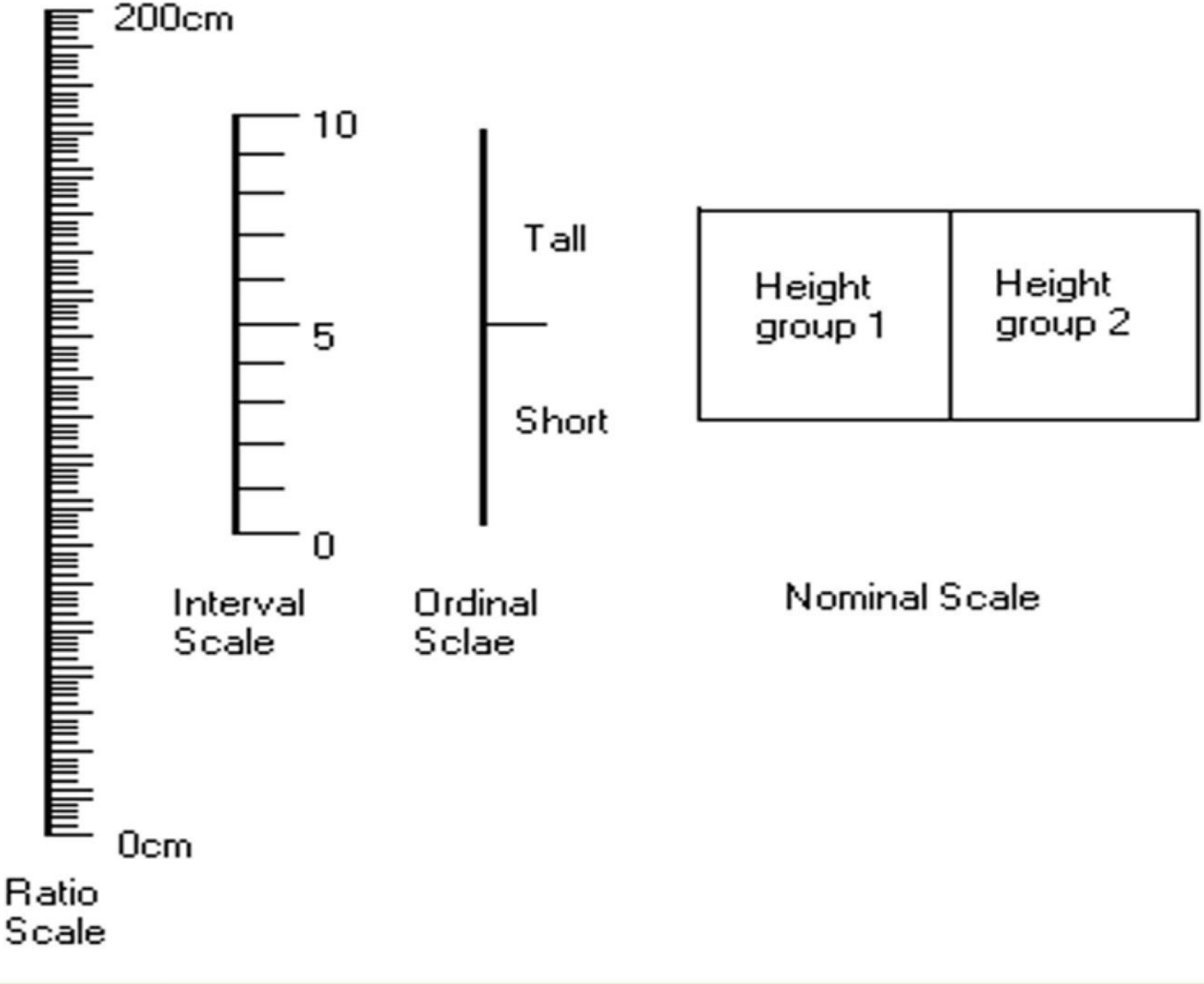
the interval level with an inherent zero starting point. Differences and ratios are meaningful for this level of measurement.

EXAMPLES: money, heights of NBA players.

Scales of Measurement - Ratio (Cont.....)

- The highest level of measurement that includes all the characteristics of the interval level
- The point 0 indicates the absence of the characteristics. Ratio between two numbers is meaningful
- Can say company A (with net wealth 100 b) is 10 times more rich than company B (with net wealth 10 b)

Levels	Arithmetic	Features	Examples
Nominal	Counting	Categories	Religion
Ordinal	Counting Ranking	Categories Ranks	Economic stratus
Interval	Counting Ranking Addition Subtraction	Categories Ranks Has equal units	IQ score
Ratio	Counting Ranking Addition Subtraction Multiplication division	Categories Ranks Has equal units Has absolute zero	Family size


200cm
10
5
0
Tall
Short
Height group 1
Height group 2
Interval Scale
Ordinal Sclae
Nominal Scale
0cm
Ratio Scale

DATA COLLECTION

FREQUENCY DISTRIBUTION

Collection, Processing and Presentation of Data

- Sources of data
- Data collection methods
- Processing of data
- Presentation of data: Graphs and Tables

Sources of data

Researching problems usually requires data.

- Data on these research problems may be found in published articles, journals, and magazines (Secondary data).
- If published data is not available on a given subject. In such cases, information have to be collected (Primary data).
- There are different methods of collecting data.

Sources of data (Cont....)

- **Primary data** are originated by a researcher for the specific purpose of addressing the research problem at hand. It is expensive and time consuming.

- **Secondary data** are data that have already been collected for purposes other than the research problem at hand. It is inexpensive and fast.
- Example: Price of rice in Dhaka city market.

Researcher may collect data from the retailers directly the data obtained this way is primary data.

He/She may use data published in different daily news papers, then it will be secondary data.

Data collection methods (Primary data)

- Data collection instruments (Questionnaire/ Schedule): Questionnaire or Schedule is formalized set of questions for obtaining information from respondents:

- Both Questionnaire and Schedule serve the same purpose, the only difference is that Questionnaires are filled by the respondent and the Schedules are filled by the interviewer.

- Observation
- Personal interview
- Telephone Conversation
- Mailing
- Combination

Processing of data

- Checking
- Editing
- Coding
- Data entry
- Cleaning
- Ready for analysis

Data Presentation (Frequency Distribution)

- Frequency distribution: A grouping of data into categories showing the number of observations in each mutually exclusive category.

- Goal is to establish a table that will quickly reveal the underlying shape of the data

Data Presentation (Frequency Distribution)

Example: Selling prices (in Lac TK.) of Apartments sold last year by Sheltech

38.70, 42.93, 41.07, 45.66, 48.16, 15.27, 43.15, 54.88,

19.98, 33.64, 49.94, 55.45, 39.58, 47.58, 41.94, 34.39,

30.34, 36.81, 20.56, 47.33, 11.92, 14.95, 37.36, 17.67,

18.29, 12.24, 45.74, 19.39, 30.85, 23.45, 48.51, 12.45,

54.29, 57.61, 15.40, 44.10, 33.55, 32.56, 21.42, 22.12,

20.32, 45.38, 34.21, 44.84, 19.71, 15.04, 38.91, 21.59, 48.51, 28.09 (n=50)

Present the data in a frequency distribution.

Example : Frequency Distribution

Table: Frequency Distribution of Selling Prices of Apartments at Dhaka City

Class Interval	Tallies	Frequency
10 to 20	𝍸 ||||	9
20 to 30	𝍸 𝍸 ||	10
30 to 40	𝍸 𝍸 |	11
40 to 50	𝍸 𝍸 𝍸 ||	15
50 to 60	𝍸 ||	5
Total		**50**

- Selling prices ranged from about Tk.10 lacs to TK. 60 lac
- About 30% of houses are sold in the range 40 to 50 lac
- Only five houses are sold by the price more than 50 lac

Frequency Distribution

Disadvantages

- We cannot pinpoint the exact selling price
- Price of the least expensive apartment was TK. 11.92 lac
- Price of the most expensive apartment was TK. 57.61 lac

Advantage

- Consider the data into a more understandable form

Frequency Distribution

Class mark (midpoint): A point that divides a class into two equal parts. This is the average between the upper and lower class limits.

Class Midpoint = (Lower Limit + Upper Limit)/2

Class interval: For a frequency distribution having classes of the same size, the class interval is the difference between upper and lower limits of a class.

Class Interval = Upper Limit − Lower Limit

Construction of Frequency Distributions

- The class interval used in the frequency distribution should be equal
- Sometime unequal class intervals are necessary to avoid a large number of empty or almost empty classes
- Too many classes or too few classes might not reveal the basic shape of the data
- Use your professional judgment to select the number of classes

Construction of Frequency Distributions

Expected number of classes can be calculated using the formula: $k=1 + 3.322 \times \log_{10} n$

Calculation of Class Interval if number of classes is known

The suggested class interval is:

(highest value-lowest value)/number of classes.

Hence the class interval is:

(highest value-lowest value)/k.

Consider a suitable rounded value as number of class and class interval

Stem-and-Leaf Displays

A statistical technique for displaying data

Each numerical value is divided into two parts: stem and leaf

Stem is the leading digit of the value

Leaf is the trailing digit of the value

Number	Stem	Leaf
9	0	9
16	1	6
108	10	8

Note: An advantage of the stem-and-leaf display over a frequency distribution is: we do not lose the identity of each observation.

Stem-and-Leaf Displays

EXAMPLE: Colin achieved the following scores on his twelve accounting quizzes this semester: 86, 79, 92, 84, 69, 88, 91, 83, 96, 78, 82, 85. Construct a stem-and-leaf chart for the data.

stem	leaf
6	9
7	8 9
8	2 3 4 5 6 8
9	1 2 6

Stem-and-Leaf Displays

Example: Represent the data in a stem and leaf display

96, 93, 88, 117, 127, 95, 108, 94, 148, 156, 139, 142, 94, 107, 125, 155, 155, 103, 112, 127, 117, 120, 112, 135, 132, 111, 125, 104, 106, 139, 134, 119, 97, and 89

Stem	Leaf
8	89
9	344567
10	34678
11	122779
12	05577
13	24599
14	28
15	556

Graphic Presentation

The commonly used graphic forms are Bar Diagram, Histograms, Frequency polygon, and a cumulative frequency curve (ogive).

Histogram: A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars (equal class interval) and the bars are drawn adjacent to each other.

Histogram for Hours Spent Studying

Graphical presentation

Bar Chart: A graph in which the categories are marked on X-axis the frequencies on the Y-axis. The class frequencies are represented by the heights of the bars, usually there are gaps between two bars.

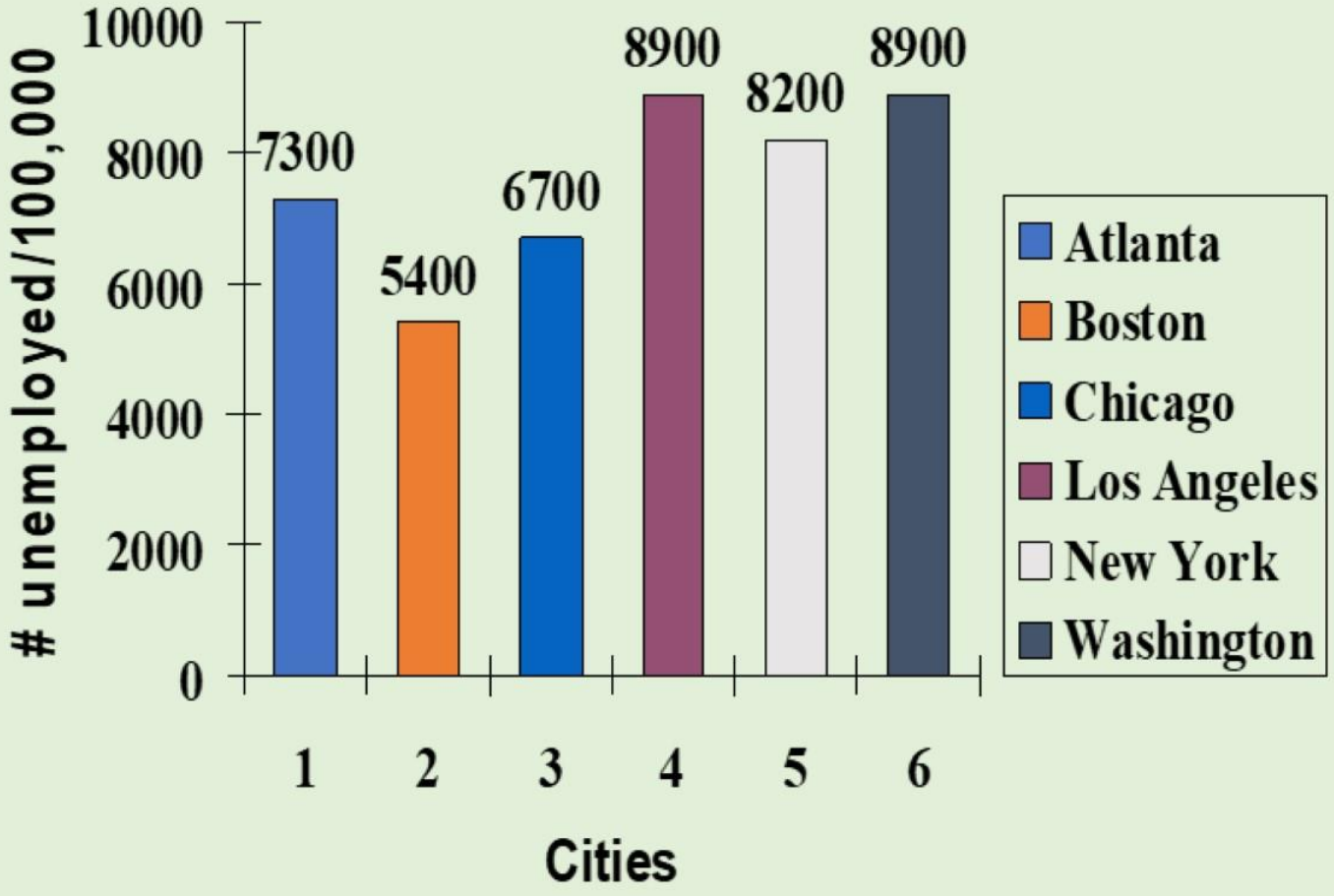
A bar chart is usually preferable for representing nominal or ordinal i.e. to represent qualitative data.

EXAMPLE: Construct a bar chart for the number of unemployed people per 100,000 population for selected six cities of USA in 1995.

Graphical Presentation

City	Number of unemployed per 100,000 population
Atlanta, GA	7300
Boston, MA	5400
Chicago, IL	6700
Los Angeles, CA	8900
New York, NY	8200
Washington, D.C.	8900

Graphical Presentation (Bar Chart)

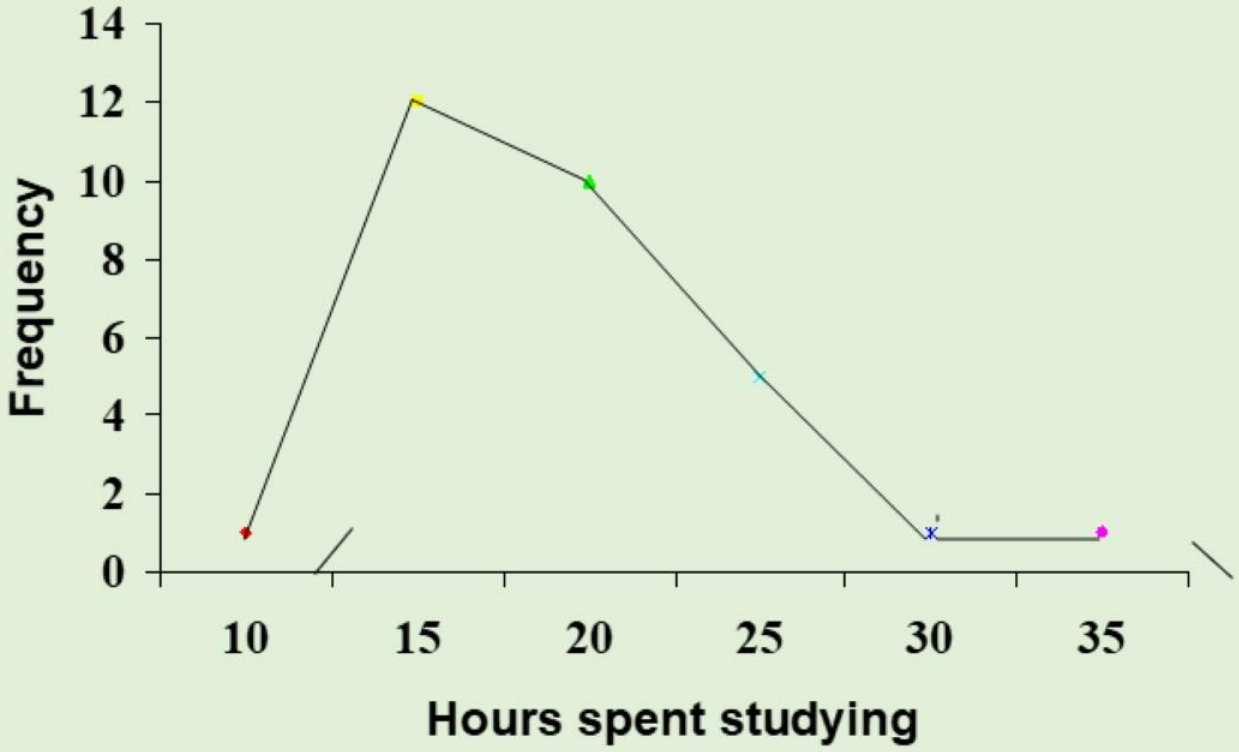


Graphic Presentation

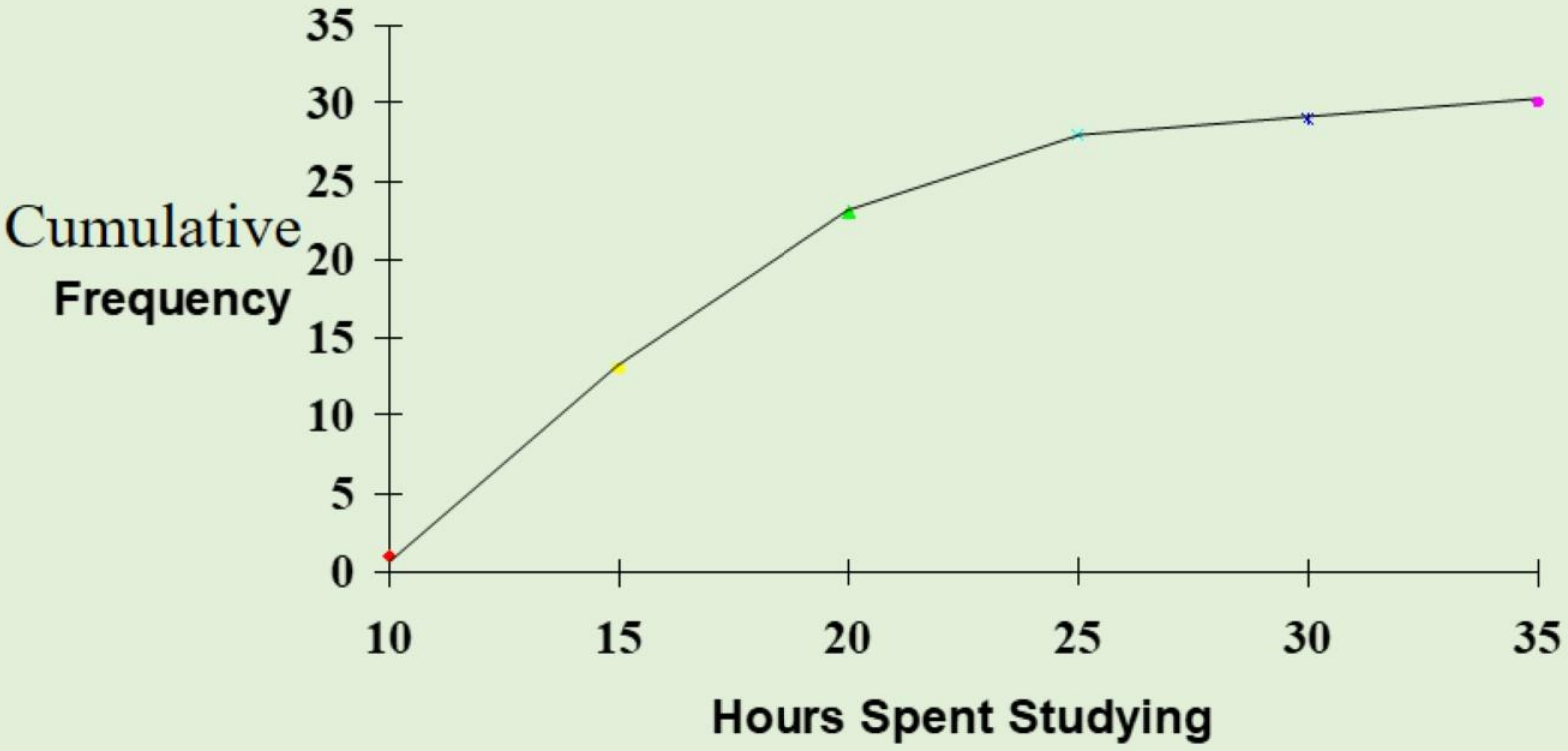
A frequency polygon consists of line segments connecting the points formed by plotting the midpoint and the class frequency for each class and than joined with X-axis at lower limit of first class and upper limit of last class.

A cumulative frequency curve (ogive) is a smooth curve obtained by joining the points formed by plotting upper limit (less than type) or lower limit (more than type) of and the cumulative frequency of each class. It is used to determine how many or what proportion of the data values are below or above a certain value.

Frequency Polygon



Cumulative Frequency Curve (less than)



Graphical Presentation (Pie Chart)

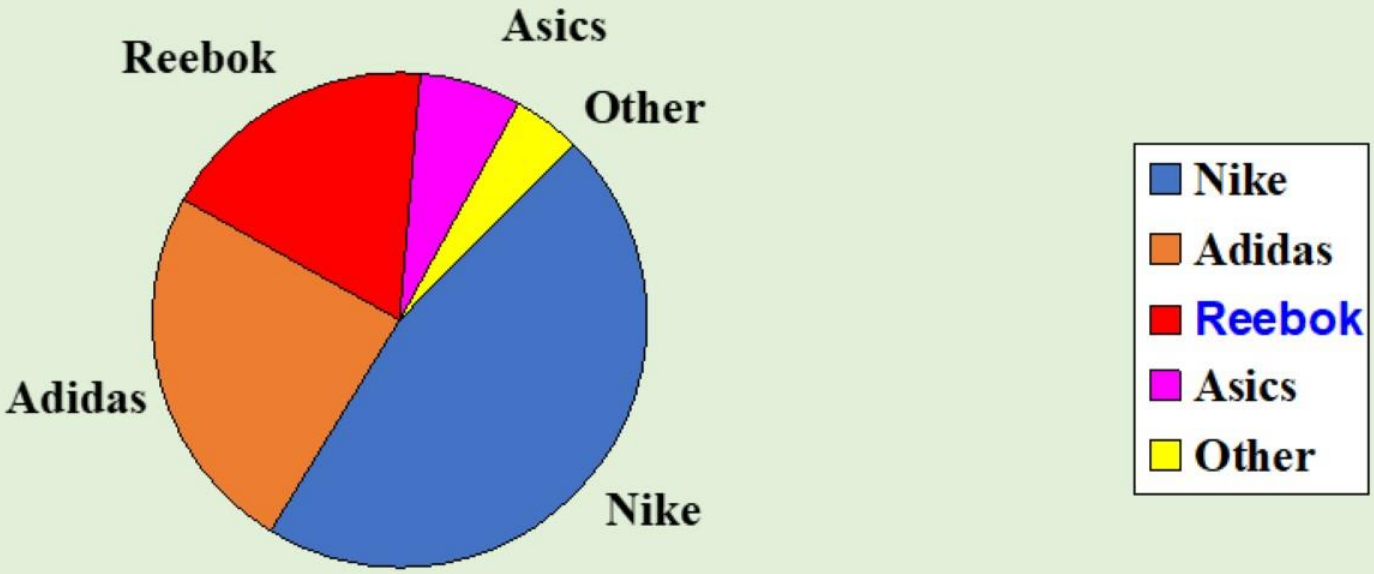
A pie chart is especially useful when there are many classes and class frequency is highly fluctuating. It displays a relative frequency distribution. A circle is divided proportionally to the relative frequency and portions of the circle are allocated for the different groups.

Graphical Presentation (Pie Chart)

EXAMPLE: A sample of 200 runners were asked to indicate their favorite type of running shoe. Draw a pie chart based on the information obtained.

Type of shoe	# of runners
Nike	92
Adidas	49
Reebok	37
Asics	13
Other	9

Graphical Presentation (Pie Chart)



Central Tendency

Measure of Central Tendency

Measure of Central Tendency is a single value that summarizes a set of data

It locates the center of the values

Also known as measure of location or **average**

Different Measures of Central Tendency

- Arithmetic mean
- Median
- Mode

Measures of Central Tendency (Arithmetic Mean)

- Arithmetic mean is the most widely used measures of central tendency
- The population mean is the sum of all values in the population divided by the total number of values:

$$(\mu) = \Sigma X \; / \; N$$

μ : Greek letter 'mu' represents the population mean

N : Number of the items in the population

X : Any particular value of the data

Σ : Greek letter 'Sigma' indicates the operation adding.

Example 1 (Arithmetic Mean)

Data on apartment prices

38.70, 42.93, 41.07, 45.66, 48.16, 15.27, 43.15, 54.88,
19.98, 33.64, 49.94, 55.45, 39.58, 47.58, 41.94, 34.39,
30.34, 36.81, 20.56, 47.33, 11.92, 14.95, 37.36, 17.67,
18.29, 12.24, 45.74, 19.39, 30.85, 23.45, 48.51, 12.45,
54.29, 57.61, 15.40, 44.10, 33.55, 32.56, 21.42, 22.12,
20.32, 45.38, 34.21, 44.84, 19.71, 15.04, 38.91, 21.59,
48.51, 28.09

Population mean $(\mu) = \Sigma X / N$

= (38.70+42.93+⋯+28.09)/50

= 1677.83/50

= 33.56

Here,

N = 50 and Σ X = 1677.83

Example 2 (Arithmetic Mean)

The Kiers family owns four cars. The following is the mileage attained by each car: 56,000, 23,000, 42,000, and 73,000. Find the average miles covered by each car.

The population mean is (56,000 + 23,000 + 42,000 + 73,000)/4 = 48,500

Parameter: a measurable characteristic of a population.

Measures of Central Tendency (Arithmetic Mean)

The Sample Mean

The sample mean is the sum of all the values in the sample divided by the number of values in the sample

Symbolically,

$$\bar{X} = \Sigma X / n$$

where,

$\bar{X}$: Sample mean

$\sum X$: Sum of the sample values

n : Sample size (No. of values in the sample)

Examples (Arithmetic Mean)

Example 3: A sample of five executives received the following amounts of bonus last year: $14,000, $15,000, $17,000, $16,000, and $15,000. Find the average bonus for these five executives.

Ans: The sample mean ($\overline{X}$) = (14,000 + 15,000 +17,000 + 16,000 +15,000)/5 = $15,400.

Example 4: Grades of five students chosen randomly from MBA students are as follows: 93, 90, 88, 90, 94

Ans: The sample mean ($\overline{X}$) = (93 + 90 + 88 + 90 + 94)/5

= 465 / 5 = 93

Statistic: a measurable characteristic of a sample.

Properties of the Arithmetic Mean

- Every set of interval-level and ratio-level data has a mean.
- All the values are included in computing the mean.
- A set of data has a unique mean.
- The mean is affected by unusually large or small data values.
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.

Properties of the Arithmetic Mean

- Sum of deviations of each value from the mean is zero, i.e.,

$$\Sigma(X - \bar{X}) = 0$$

- Consider the set of values: 3, 8, and 4. The mean is 5. Illustrating the fifth property, (3-5) + (8-5) + (4-5) = -2 +3 -1 = 0.

Exercise (Arithmetic Mean)

EX 1: The monthly income of a sample of several mid-level management employee of Grameen phone are : Tk 40,300, Tk 40,500, Tk 40,200, Tk 41,000, and Tk 44,000

(a) Give the formula for the sample mean.

(b) Find the sample mean.

(c) Is the mean you computed at (b) is a statistic or a parameter? Why?

EX 2 Compute the mean of the sample values: 5, 9, 4, 10
Hence show that $\sum(X - \bar{X}) = 0$

Weighted Mean

- The weighted mean of a set of numbers X1, X2, ..., Xn, with corresponding weights w1, w2, ...,wn, is computed from the following formula:

$$X_w = (w_1X_1 + w_2X_2 + \ldots + w_nX_n)/(w_1 + w_2 + \ldots w_n)$$

$$X_w = \Sigma(w * X)/\Sigma w$$

Weighted Mean

Example: Calico pizza sells colas in three sizes: small, medium, and large. The small size costs \$0.50, the medium \$0.75, and the large \$1.00. Yesterday 20 small, 50 medium, and 30 large colas were sold. What is the average price per cola?

The weighted mean

$$\bar{X} = \sum(wX)/\sum w$$

$$= \{20(0.50) + 50(0.75) + 30(1.0)\}/ (20+50+30)$$

$$= (10 + 37.5 + 30)/100 = 0.775$$

Measures of Central Tendency (Median)

Median:

The midpoint of the values after they have been ordered from the smallest to the largest, or the largest to the smallest. There are as many values above the median as below it in the data array.

Note: For a set of even number of values, the median is the arithmetic mean of the two middle values.

Measures of Central Tendency (Median)

Arithmetic mean of five values : 20, 30, 25, 290, 33 is 79.6

Problem?

- Does 79.6 represent the center point of the data?
- It seems a value between 20 to 33 is more representative average
- Arithmetic mean cannot give the representative value in this case

Measures of Central Tendency (Median)

To find the median, we first order the given values 20, 30, 25, 290, 33

The ordered values are 20, 25, 30, 33, 290

Since n is odd, $(n+1)/2^{th}$ observation is median

The middle of the five observations is the third observation

The third observation is 30, So Median = 30

30 is more representative center value than the mean 79.6

Properties of the Median

- There is a unique median for each data set.
- It is not affected by extremely large or small values and is therefore a valuable measure of central tendency when such values occur.
- It can be computed for ratio-level, interval-level, and ordinal-level data.
- It can be computed for an open-ended frequency distribution if the median does not lie in an open-ended class.

Measures of Central Tendency (Mode)

- The mode is the value of the observation that appears most frequently.
- It is useful measure of central tendency for nominal and ordinal level of measurements
- It can be computed for all levels of data
- Data set may have no mode because no value is observed more than once

- EXAMPLE : The exam scores for ten students are: 81, 93, 84, 75, 68, 87, 81, 75, 81, 87. Since the score of 81 occurs the most, the modal score is 81.

Measures of Central Tendency (Mode)

Exercise: The number of work stoppages in an automobile company for selected months are 6, 0, 10, 14, 8, 0

- What is the median number of stoppages?
- How many observations are below and above the median?
- What is the modal number of stoppages?

Group Data (The Arithmetic Mean)

The mean of a sample of data organized in a frequency distribution is computed by the following formula:

$$\overline{X} = \frac{\Sigma Xf}{\Sigma f} = \frac{\Sigma Xf}{n}$$

Where,

X : Class marks or Mid points of the class

f : Frequency in the class

Group Data (The Median)

The median of a sample of data organized in a frequency distribution is computed by the following formula:

$$Median = L + \left(\frac{\frac{N}{2} - f_c}{f_m} \right) \times c$$

L : lower limit of the median class

cf : cumulative frequency pre-median class

f : frequency of the median class

ci : class interval.

n : total number of observation

Finding the Median Class

To determine the median class for grouped data:

- Construct a cumulative frequency distribution.
- Divide the total number of data values by 2.
- Determine which class will contain this value. For example, if n=50, 50/2 = 25, then determine which class will contain the 25th value - the median class.

Group Data (The Mode)

The mode for grouped data is approximated by the midpoint of the class with the largest class frequency.

The formula used to calculate the median from grouped data is:

$$Mode = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times c$$

L : lower limit of the modal class

Δ1 : difference between frequencies of modal and pre-modal class

Δ2 : difference between frequencies of modal and post-modal class

c : class interval.

Exercise:

A sample of 50 antique dealers in Southeast USA revealed the following sales last year:

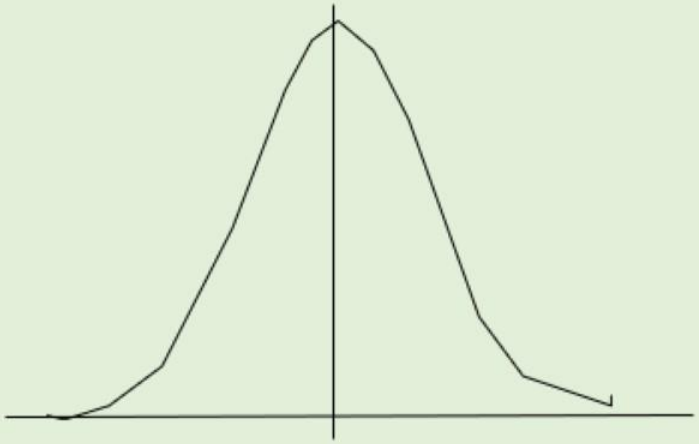
Sales (in 000 $)	No. of dealers
100 – 120	5
120- 140	7
140 - 160	9
160 - 180	16
180 - 200	10
200 - 220	3

Calculate the mean, median, and mode

Symmetric Distribution

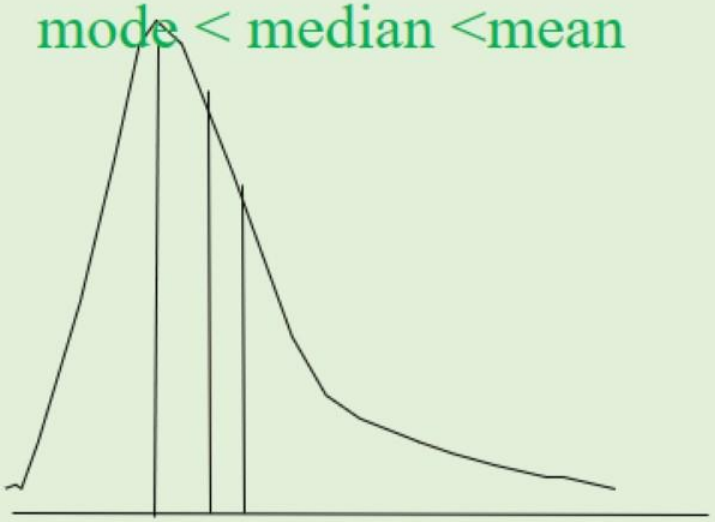
Zero skewness:

mode = median = mean

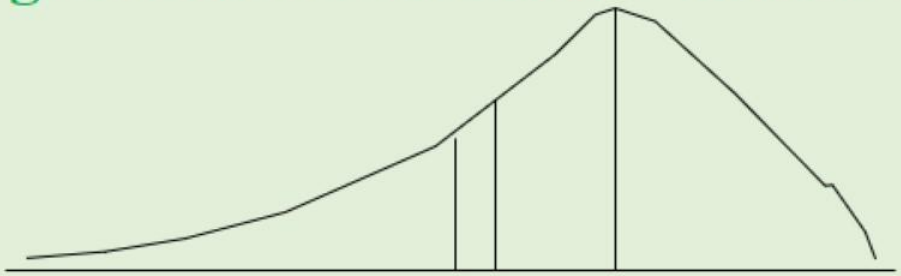


Positive skewness:

mode < median <mean



Negative skewness: mode > median >mean



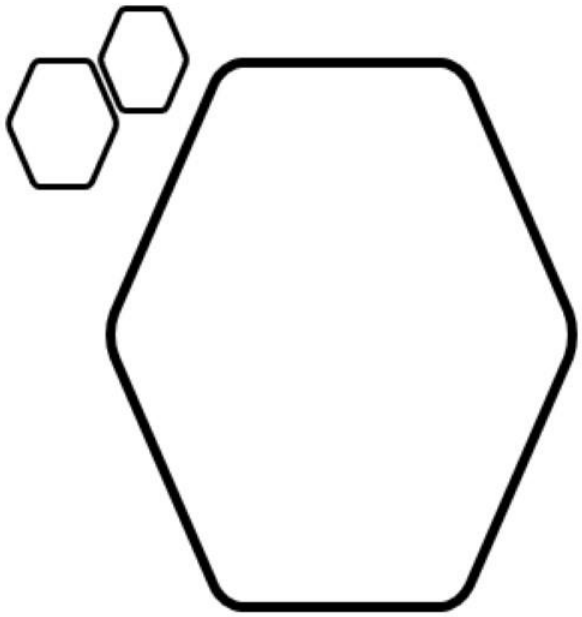
NOTE

If two averages of a moderately skewed frequency distribution are known, the third can be *approximated*.

- Mode = mean - 3(mean - median)
- Mean = [3(median) - mode]/2
- Median = [2(mean) + mode]/3

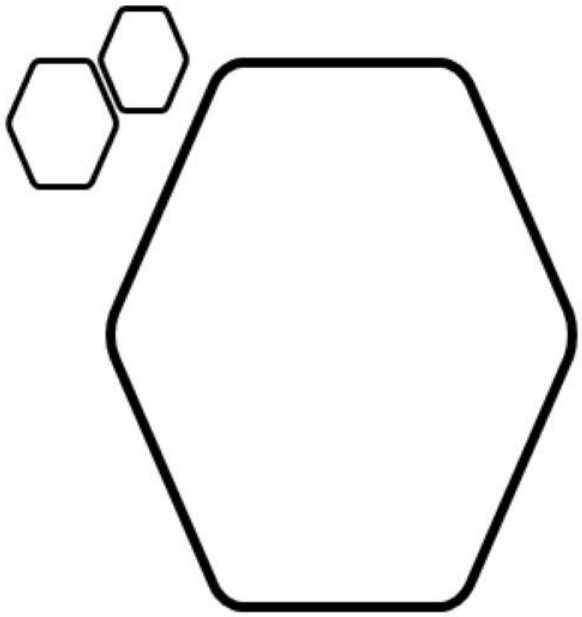
Measure of Dispersion

Measures of Dispersion



- It deals with spread of the data
- A small value of the measure of dispersion indicates that data are clustered closely
- A large value of dispersion indicates the estimate of central tendency is not reliable

Measures of Dispersion



Absolute measures:

Range

Mean Deviation

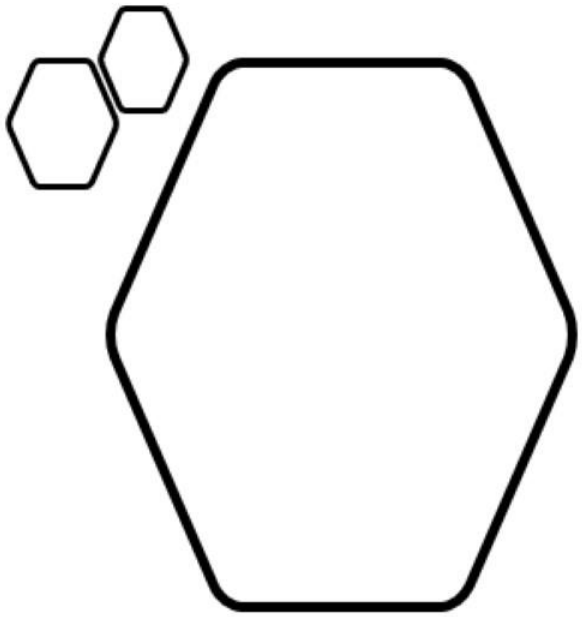
Variance

Standard Deviation

Relative measures:

Coefficient of Variation

Measures of Dispersion (Range)



Range indicates the difference between the highest and lowest value of the data set

Range = Highest value – Lowest value

Example : 93, 103, 105, 110, 104, 112, 105, 90

Range = 112 – 90 = 22

Properties

It depends only on two values

It can be influenced by extreme values

Measures of Dispersion (Mean Deviation)

Mean Deviation: The arithmetic mean of the absolute values of the deviations of the observations from the arithmetic mean.

$$MD = \frac{\Sigma |X - X|}{n}$$

Properties

MD uses all the values in the sample

Absolute values are difficult to work with

Measures of Dispersion (Mean Deviation)

Example : 93, 103, 105, 110, 104, 112, 105, 90

$\bar{X} = \sum X/n = 822/8 = 102.75$

X	$X - \bar{X}$	$\lvert X - \bar{X} \rvert$
93	-9.75	9.75
103	0.25	0.25
105	2.25	2.25
110	7.25	7.25
104	1.25	1.25
112	9.25	9.25
105	2.25	2.25
90	-12.75	12.75
		$\sum \lvert X - \bar{X} \rvert = 45$

$MD = \sum \lvert X - \bar{X} \rvert / n = 45/8 = 5.62$

Measures of Dispersion (Variance)

Variance is the arithmetic mean of the squared deviations of observations from the mean

Population variance: $$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Sample Variance: $$S^2 = \frac{\Sigma(X-\overline{X})^2}{n-1}$$

σ^2 is a Parameter and S^2 is a Statistic

Measures of Dispersion (Variance)

Example : 93, 103, 105, 110, 104, 112, 105, 90

$\overline{X} = \sum X/n = 822/8 = 102.75$

X	$X - \overline{X}$	$(X - \overline{X})^2$
93	-9.75	95.06
103	0.25	0.06
105	2.25	5.06
110	7.25	52.56
104	1.25	1.56
112	9.25	85.56
105	2.25	5.06
90	-12.75	162.56

$\sum(X - \overline{X})^2 = 407.48$

Population Variance = $\sum(X - \mu)^2/N = 407.48/8 = 50.935$

Sample Variance = $\sum(X - \overline{X})^2/(n-1) = 407.48/7 = 58.21$

Measures of Dispersion (Variance)

Working formula for population variance is:

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$$

Working formula for sample variance is:

$$S^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}$$

Measures of Dispersion (Variance)

Example : 93, 103, 105, 110, 104, 112, 105, 90

X	X^2
93	8649
103	10609
105	11025
110	12100
104	10816
112	12544
105	11025
90	8100
$\sum X=822$	$\sum X^2=84868$

Population Variance

$$= \frac{\Sigma X^2}{N} - (\frac{\Sigma X}{N})^2 = \frac{84868}{8} - (\frac{822}{8})^2$$

$$= 10608.5 - 10557.6 = 50.9$$

Sample Variance

$$= \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1} = \frac{84868 - \frac{(822)^2}{8}}{8-1}$$

$$= \frac{84868 - 84460.5}{7} = \frac{407.5}{7} = 58.2$$

Measures of Dispersion (Standard Deviation)

Standard Deviation (SD) is the positive square root of the variance

Population SD: $\sigma = \sqrt{\sigma^2} = \sqrt{\left(\frac{\Sigma(X-\mu)^2}{N} \right)}$

Sample SD: $S = \sqrt{S^2} = \sqrt{\left(\frac{\Sigma(X-\bar{X})^2}{n-1} \right)}$

σ is a Parameter and S is a Statistic

Measures of Dispersion (Variance)

Example : 93, 103, 105, 110, 104, 112, 105, 90

Population SD =

$$\sigma = \sqrt{\sigma^2} = \sqrt{50.935} = 7.14$$

Sample SD

$$S = \sqrt{S^2} = \sqrt{58.21} = 7.63$$

Measures of Dispersion (Group Data)

Range: The difference between the upper limit of the highest class and the lower limit of the lowest class

Mean Deviation:

$$MD = \frac{\Sigma f |X - \bar{X}|}{\Sigma f}$$

Measures of Dispersion (Group Data)

Variance: The formula for the calculation of variance from a grouped data is:

$$S^2 = \frac{\sum f\left(X - \overline{X}\right)^2}{\Sigma f} = \frac{\Sigma fX^2 - \dfrac{(\Sigma fX)^2}{\Sigma f}}{\Sigma f}$$

Where, f is class frequency,

X is class midpoint and

$\overline{X}$ is arithmetic mean.

Standard Deviation: S = √Variance

Exercise (Measures of Dispersion)

A sample of 50 antique dealers in Southeast USA revealed the following sales last year:

Sales (in 000 $)	Number of dealers
100 – 120	5
120 – 140	7
140 – 160	9
160 – 180	16
180 – 200	10
200 – 220	3

Find range, mean deviation, variance, and standard deviation.

Relative Dispersion

The usual measures of dispersion cannot be used to compare the dispersion if the units are different, even if the units are same but means are different

It reports variation relative to the mean

It is useful for comparing distributions with different units

Relative Dispersion

Coefficient of Variation (CV):

$$CV = \frac{s}{\overline{X}} \times 100$$

The CV is the ratio of the standard deviation to the arithmetic mean, expressed as a percentage:

Relative Dispersion (CV)

Example: The variation in the monthly income of executives is to be compared with the variation in incomes of unskilled employees. For a sample of executives, X = TK 50, 000 and s = TK 5, 000. For a sample of unskilled employees, X = TK 2, 200 and s = TK 220.

Solution:

For executives,

CV =(s/X) × 100 = (Tk. 5000/TK 50, 000) × 100

= 10 percent

For unskilled employees,

CV = (s/X) × 100 = (Tk. 220 /TK 2, 200) × 100

= 10 percent

Interpretation and Uses of the Standard Deviation: Empirical Rule

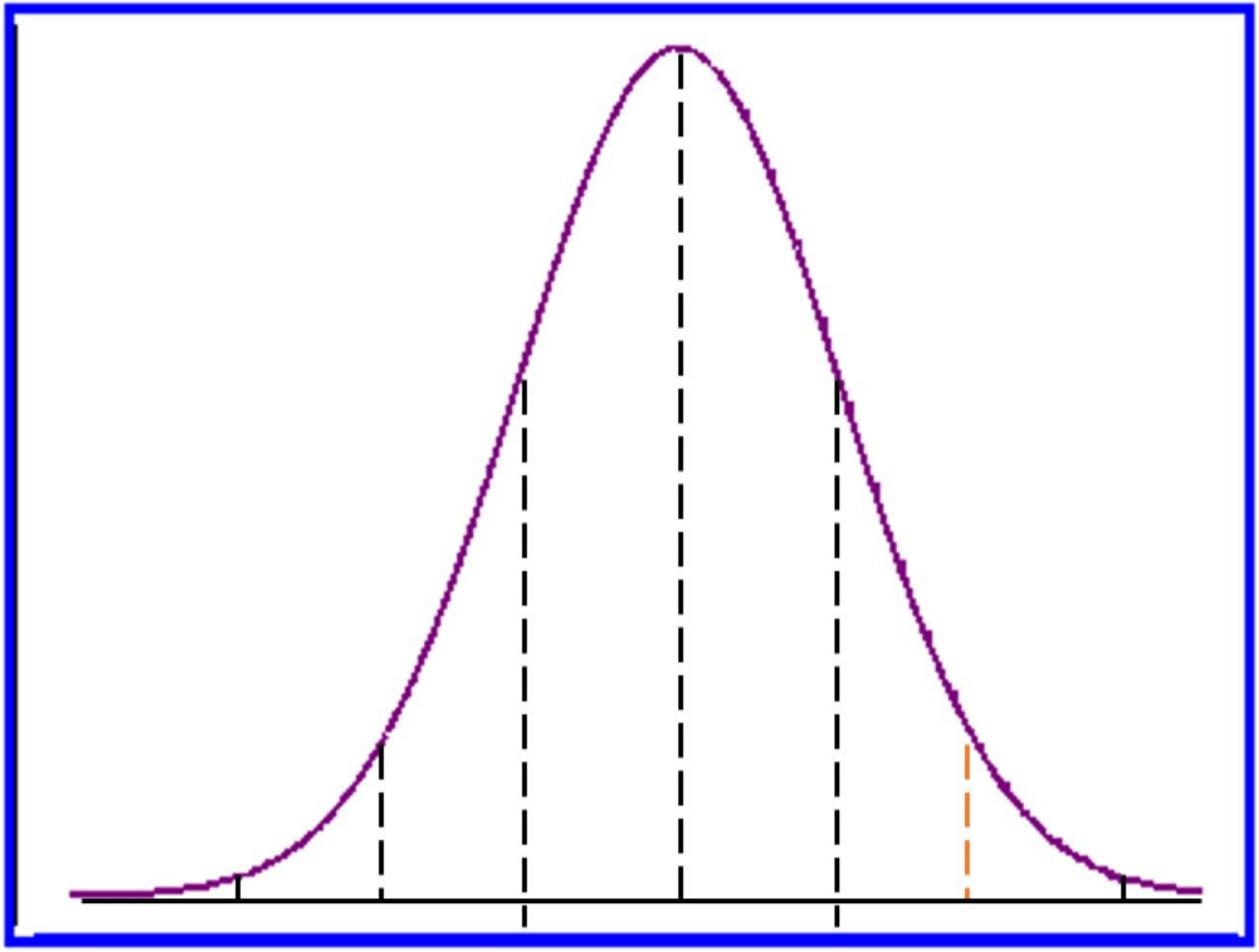
For any symmetrical, bell-shaped distribution,

- approximately 68% of the observations will lie within $\pm \ 1\sigma$ of the mean (μ)

- approximately 95% of the observations will lie within $\pm 2\sigma$ of the mean (μ)

- approximately 99.7% within $\pm 3\sigma$ of the mean (μ).

Bell - Shaped Curve showing the relationship between σ and μ.



Correlation and Regression

Correlation and Regression

- The most commonly used forms of bi-variate statistical analysis
- Useful in making business and economic decisions
- Helpful in identifying the nature of relationship among many business and economic variables
- Recognize that there is a quantifiable relationship between two or more variables
- One variable depends on another and can be determined by it

Correlation and Regression

The variables :

Students GPAs and amount of time they spend on studying

A firm's sale and expenditure on advertisement

Dependent variable and Independent variable

Determination of dependent and independent variable is crucial

Usually

X : Independent variable

Y : Dependent variable

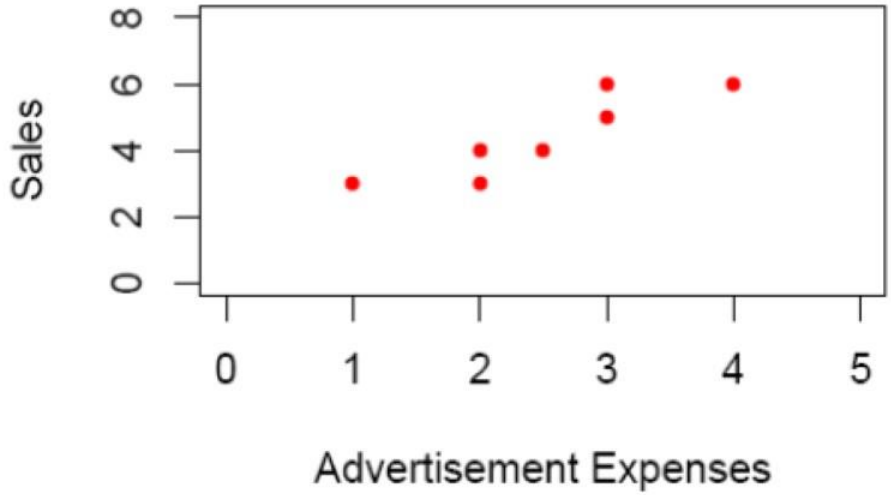
Scatter Diagram

- A plot of the paired observations of X and Y on a graph
- Graphically shows the relationship between two variables
- Common practice is to place the dependent variable on Y–axis and independent variable on X–axis

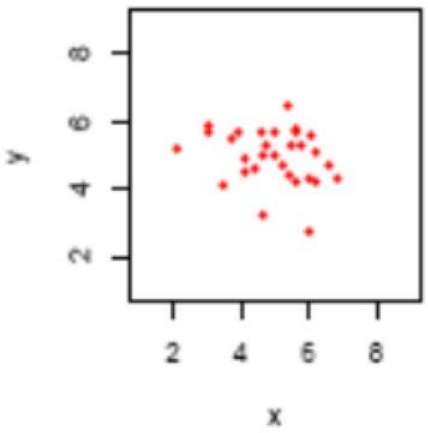
Ex. Sales and advertisement expenditures (in million Taka) of a firm on different months are

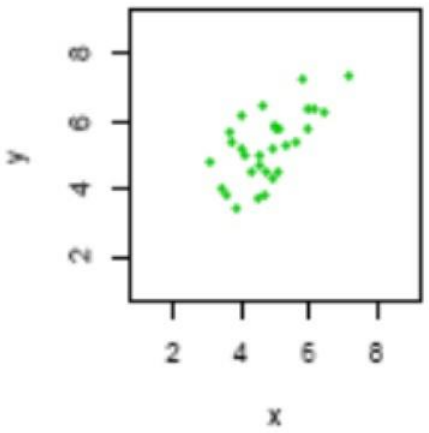
Sales	3	6	4	6	3	5	4
Advertisement	2	4	2	3	1	3	2.5

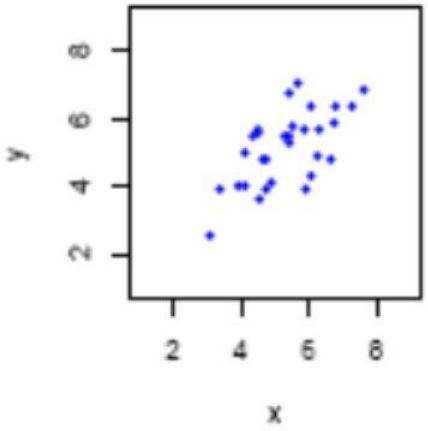
Scatter Diagram

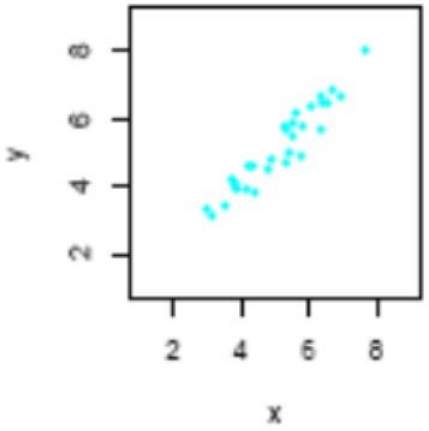


Scatter Diagram









Correlation Analysis

- **Correlation Analysis:** A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- **Scatter Diagram:** A chart that portrays the relationship between the two variables of interest.
- **Dependent Variable:** The variable that is being predicted or estimated.
- **Independent Variable:** The variable that provides the basis for estimation. It is the predictor variable.

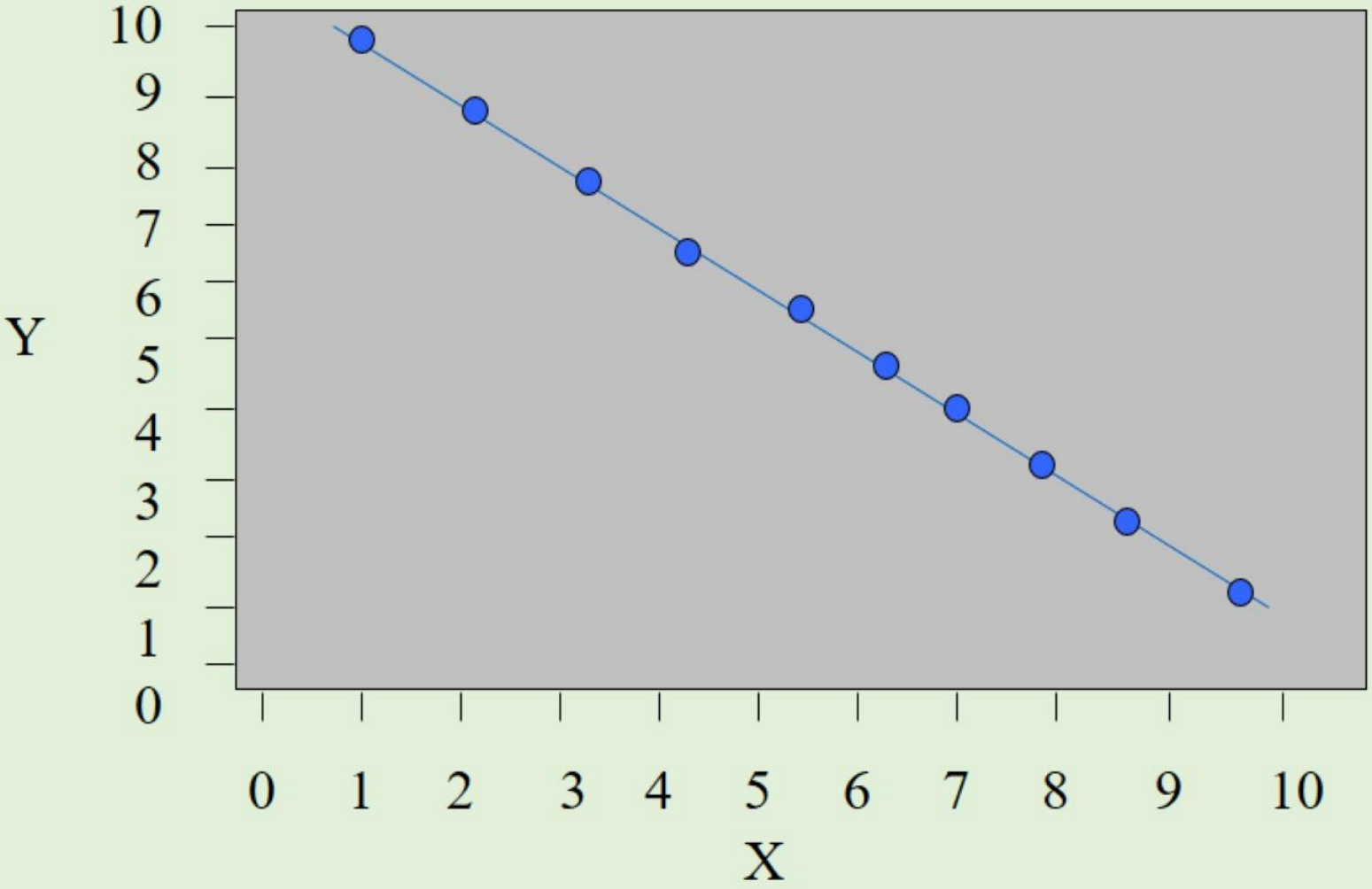
The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the strength of the relationship between two variables.
- It requires interval or ratio-scaled data (variables).
- It can range from -1.00 to 1.00.
- Values of -1.00 or 1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate no linear correlation.
- Negative values indicate an inverse relationship and positive values indicate a direct relationship.

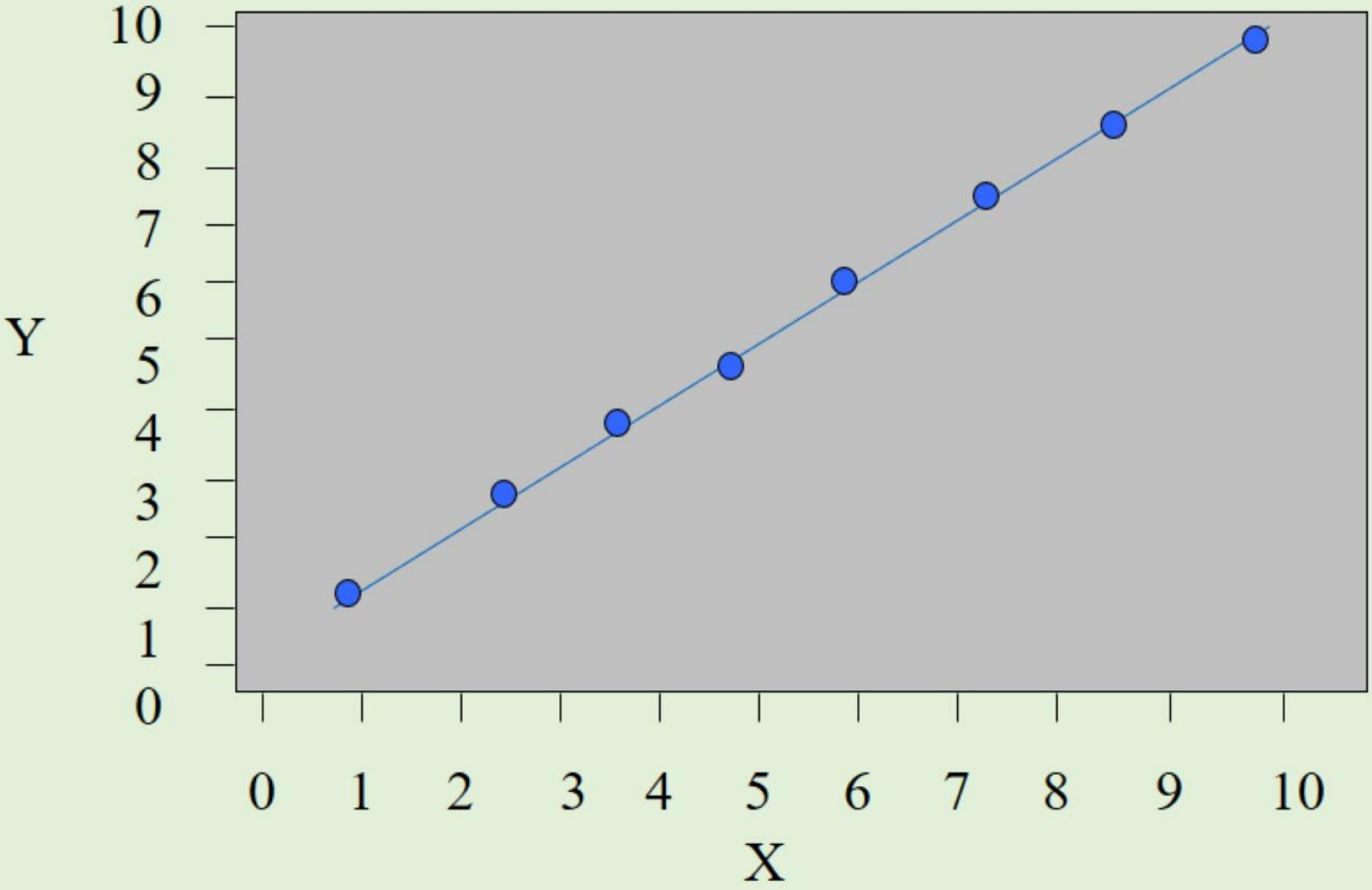
The Coefficient of Correlation, r

$r = -1$	Perfect negative correlation
$-1 < r < -.5$	Strong negative correlation
$r = -.5$	Moderate negative correlation
$-.5 < r < 0$	Weak negative correlation
$r = 0$	No correlation

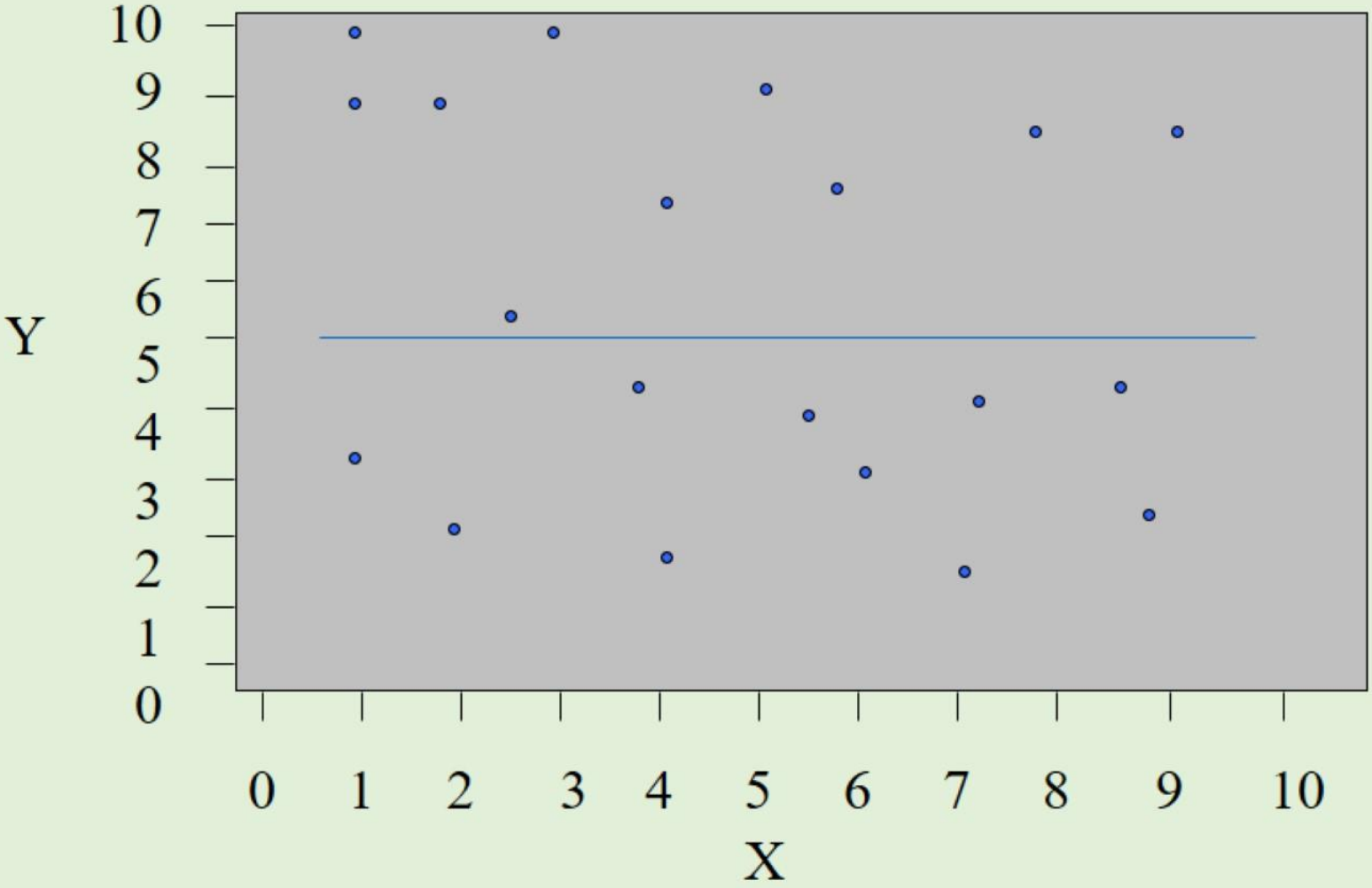
Perfect Negative Correlation



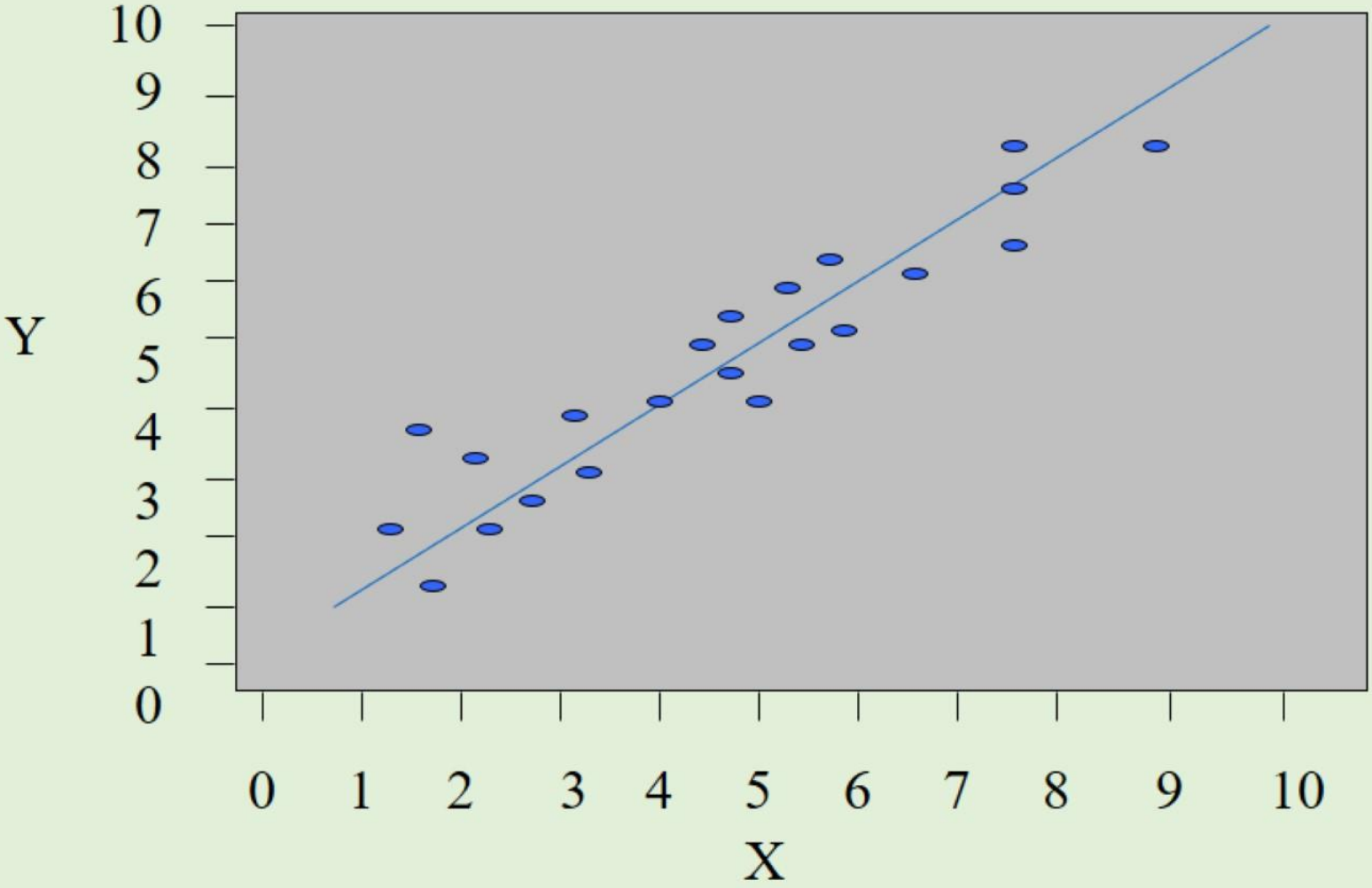
Perfect Positive Correlation



Zero Correlation



Strong Positive Correlation



Formula for r

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

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n(\Sigma X^2) - (\Sigma X)^2\right]\left[n(\Sigma Y^2) - (\Sigma Y)^2\right]}}$$

Coefficient of Determination, r^2

The Coefficient of Determination, r^2 - the proportion of the total variation in the dependent variable Y that is explained or accounted for by the variation in the independent variable X.

The coefficient of determination is the square of the coefficient of correlation, and ranges from 0 to 1.

Example: Sales and advertisement expense data,

$r = 0.759$ and $r^2 = (0.759)^2 = 0.576$

57.6% variation of sales can be explained by the variation in advertisement expenses

Regression Analysis

- In regression analysis an equation is developed to express the relationship between dependent and independent variables
- The equation must be linear
- Purpose: to determine the regression equation; it is used to predict the value of the dependent variable (Y) based on the independent variable (X).
- Procedure: select a sample from the population and list the paired data for each observation; draw a scatter diagram to give a visual portrayal of the relationship; determine the regression equation.

Regression Analysis

General form of linear regression model

$$Y = a + bX + e$$

Where,

- Y : dependent variable
- a : intercept term
- b : slope of the line
- X : independent variable
- e : error term

Want to estimate a and b such that $\sum e^2$ is minimum

Regression Analysis

The error sum of squares $\sum e^2$ will be minimum if

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}, \quad \hat{b} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

These estimates are known as least squares estimates

Sign of $\hat{b}$ is similar to that of correlation coefficient r

Estimated value of dependent variable: $\hat{Y} = \hat{a} + \hat{b}X$

Estimated error is: $e = Y - \hat{Y}$

Regression Analysis

$\hat{Y}$ is the average predicted value of Y for any X.

$\hat{a}$ is the Y-intercept, or the estimated Y value when $X=0$

$\hat{b}$ is the slope of the line, or the average change in Y' for each change of one unit in X

Regression Analysis (Coefficient of determination)

r^2 = Percentage of total variation in the dependent variable explained by the independent variable.

From a linear regression model one can write

r^2 = (Explained variation/total variation)

$$= \frac{\text{(Total variation} - \text{Unexplained variation)}}{\text{Total variation}}$$

Regression Analysis (Coefficient of determination)

Total Variation (TSS) = $\sum(Y - \bar{Y})^2$

Unexplained variation (ESS) = $\sum(Y - \hat{Y})^2$

Explained variation (RSS) = $\sum(\hat{Y} - \bar{Y})^2$

Coefficient of variation (r^2) = $\frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$

Regression Analysis (Coefficient of determination)

$$r^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

$$S_{Y \cdot X} = \sqrt{\frac{ESS}{n-2}}$$

The Standard Error of Estimate

- How precise the regression line is?
- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression
- The formulas that are used to compute the standard error:
- Standard error = √{ESS/(n-2)}

$$S_{Y \cdot X} = \sqrt{\frac{\Sigma (Y - Y')^2}{n - 2}}$$

$$= \sqrt{\frac{\Sigma Y^2 - a(\Sigma Y) - b(\Sigma XY)}{n - 2}}$$

Regression Analysis (Exercise)

(a) If we are interested in estimating selling price based on the age of the car, which variable is the dependent and which is independent variable?

(b) Draw a scatter diagram.

(c) Determine coefficient of correlation and coefficient of determination.

(d) Develop a regression equation.

(e) Determine the selling price of a car which is 15 years old.

(f) Estimate the standard error of the estimates.

Car	Age (in years)	Selling Price (Tk 0000)
1	9	81
2	7	60
3	11	36
4	12	40
5	8	50
6	7	100
7	8	76
8	11	80
9	10	80
10	12	60
11	6	86
12	6	80

Sampling

Sampling
Terminology

Population: Collection of all objects of interest
Finite and Infinite

Sample: Representative part of population

Sampling unit: Smallest unit of population from which information is to be collected

Sampling Frame: A complete list of the sampling units

Sampled unit: Sampling units which are included in the sample

Census: Collecting information from population

Sample survey: collecting information from sample

Respondent: From whom data/information is collected

Response: a measured value (answer to an inquiry from respondent)

Statistic: sample characteristics

Parameter: Population characteristics

Sampling in Research

Why Sample the Population?

- Three major issues: Time, manpower and cost
- Contacting the population would often be time-consuming.
- More manpower is needed to investigate population
- The cost of studying all the items in a population is high.
- Some other issues:
  - The physical impossibility of checking all items in the population.
  - The destructive nature of certain tests.
  - The sample results are usually adequate.

Sampling Method

Probability Sampling:

A form that uses random selection so different units in the population has a positive (or known) chance of being selected in the sample.

Non-Probability Sampling

Do not use probability theory

Convenience

Quota sampling

Mixed Sampling

Use a mixture of probability and non-probability sampling

Probability Sampling

- A probability sample is a sample selected in such a way that each item or person in the population being studied has a known likelihood of being included in the sample.
  - **Simple Random Sampling**
  - **Systematic Random Sampling**
  - **Stratified Random Sampling**
  - **Cluster Sampling**

Methods of Probability Sampling

Simple Random Sample:

A sample formulated so that each item or person in the population has the same chance of being included.

- Method of selecting a sample:
  - Lottery Method
  - Random Number method
- Selection Process
- Start with a sampling frame, assign an ID number to each sampling unit, decide how many to select,
- Make a lottery to decide who will be included in the sample (lottery method).
- Select a random number from random number table between 1 to n, the corresponding sampling unit will be selected in the sample. Continue the process until you get a sample of n units (Random number method).

Methods of Probability Sampling

Systematic Random Sampling:

- The items or individuals of the population are arranged in some order.
- A random starting point is selected and then every r^{th} member of the population is selected for the sample.
- If we want to select a sample of size n, divide the population into n groups having r units (N = nr), First select a unit (k^{th}, say) by simple random sampling method, So, k^{th} unit from first group will be the first unit in the sample, $(k+r)^{th}$ unit from second group is the second unit in the sample, $(k+2r)^{th}$ unit from third group is the third unit and so on. Ultimate the sample will be constituted with k, k+r, k+2r, k+3r, ….., $k+(n-1)r^{th}$ unit of the population.

7	64 86	66 31	55 04	88 40	10 30	84 38	06 13	58 83	62 04	63 52
8	22 69	58 45	49 23	09 81	98 84	05 04	75 99	27 70	72 79	32 19
9	23 22	14 22	64 90	10 26	74 23	53 91	27 73	78 19	92 43	68 10
10	42 38	59 64	72 96	46 57	89 67	22 81	94 56	69 84	18 31	06 39
11	17 18	01 34	10 98	37 48	93 86	88 59	69 53	78 86	37 26	85 48
12	39 45	69 53	94 89	58 97	29 33	29 19	50 94	80 57	31 99	38 91
13	43 18	11 42	56 19	48 44	45 02	84 29	01 78	65 77	76 84	88 85
14	59 44	06 45	68 55	16 65	66 13	38 00	95 76	50 67	67 65	18 83
15	01 50	34 32	38 00	37 57	47 82	66 59	19 50	87 14	35 59	79 47
16	79 14	60 35	47 95	90 71	31 03	85 37	38 70	34 16	64 55	66 49
17	01 56	63 68	80 26	14 97	23 88	59 22	82 39	70 83	48 34	46 48
18	25 76	18 71	29 25	15 51	92 96	01 01	28 18	03 35	11 10	27 84
19	23 52	10 83	45 06	49 85	35 45	84 08	81 13	52 57	21 23	67 02
20	91 64	08 64	25 74	16 10	97 31	10 27	24 48	89 06	42 81	29 10
21	80 86	07 27	26 70	08 65	85 20	31 23	28 99	39 63	32 03	71 91
22	31 71	37 60	95 60	94 95	54 45	27 97	03 67	30 54	86 04	12 41
23	05 83	50 36	09 04	39 15	66 55	80 36	39 71	24 10	62 22	21 53
24	98 70	02 90	30 63	62 59	26 04	97 20	00 91	28 80	40 23	09 91
25	82 79	35 45	64 53	93 24	86 55	48 72	18 57	05 79	20 09	31 46

Eighth Thousand

	1–4	5–8	9–12	13–16	17–20	21–24	25–28	29–32	33–36	37–40
1	37 52	49 55	40 65	27 61	08 59	91 23	26 18	95 04	98 20	99 52
2	48 16	69 65	69 02	08 83	08 83	68 37	00 96	13 59	12 16	17 93
3	50 43	06 59	56 53	30 61	40 21	29 06	49 60	90 38	21 43	19 25
4	89 31	62 79	45 73	71 72	77 11	28 80	72 35	75 77	24 72	98 43
5	63 29	90 61	86 39	07 38	38 85	77 06	10 23	30 84	07 95	30 76
6	71 68	93 94	08 72	36 27	85 89	40 59	83 37	93 85	73 97	84 05
7	05 06	96 63	58 24	05 95	56 64	77 53	85 64	15 95	93 91	59 03
8	03 35	58 95	46 44	25 70	31 66	01 05	44 44	62 91	36 31	45 04
9	13 04	57 67	74 77	53 35	93 51	82 83	27 38	63 16	04 48	75 23
10	49 96	43 94	56 04	02 79	55 78	01 44	75 26	85 54	01 81	32 82
11	24 36	24 08	44 77	57 07	54 41	04 56	09 44	30 58	25 45	37 56
12	55 19	97 20	01 11	47 45	79 79	06 72	12 81	86 97	54 09	06 53
13	02 28	54 60	28 35	32 94	36 74	51 63	96 90	04 13	30 43	10 14

Methods of Probability Sampling

Stratified Random Sampling:

A population is first divided into subgroups, called strata, individuals having homogeneous criteria of the variable of interest belongs to same strata. Classification of the individuals into different strata based on the characteristic is known as Stratification. After stratification a sample is selected from each stratum.

Allocation of the sample size:

- Equal allocation ($n_h = N/k$)

  where, N = Population size, n_h = sample size from h^{th} stratum, k = number of stratum
- Proportional Allocation ($n_h \propto N_h$)
- Neyman Allocation ($n_h \propto N_h S_h$)
- Optimal Allocation ($n_h \propto N_h S_h/\sqrt{C_h}$)

Methods of Probability Sampling

Cluster Sampling:

- Population is first divided into subgroups (Cluster), Clusters are so formed that each individual cluster can represent whole population. A sample of the clusters is selected. Each individual of the selected clusters is included in the sample.
- Strata: Within strata individuals are homogenous and between strata heterogeneous.
- Cluster: Within cluster individuals are heterogeneous and between cluster homogeneous.

Errors involved in Sampling

- **Sampling error**
- **Non-Sampling error**
- Errors introduced only due to sampling is termed as Sampling error. A sampling error is the difference between a sample statistic and its corresponding parameter.
- All errors other than sampling error is termed as Non-Sampling error.
- Sampling error decreases but non-sampling error increases if the sample size increases, on the other hand Sampling error increases but non-sampling error decreases if the sample size decreases,
- Guess how we can reduce non-sampling error

Determination of sample size

There are 3 factors that determine the size of a sample, none of which has any direct relationship to the size of the population.

They are:

- **The degree of confidence selected.**
- **The maximum allowable error.**
- **The variation of the population.**

Determination of sample size

The formula for determining the sample size in the case of a proportion is:

$$n = p(1 - p)\left(\frac{Z}{E}\right)^2$$

Where, p= estimated proportion (assumed, past experience or a pilot survey); z = z value associated with the degree of confidence selected; E = allowable error

Example: The American PET society wanted to estimate the proportion of children that have a dog as a pet. If they wanted the estimate to be within 3% of the population proportion, how many children would they need to contact? Assume a 95% level of confidence and that the club estimated that 30% of the children have a dog as a pet.

$$n = (.30)(.70)(1.96/.03)^2 = 896.3733 \approx 897$$

Probability

Important Terms

Random Experiment – a process leading to an uncertain outcome

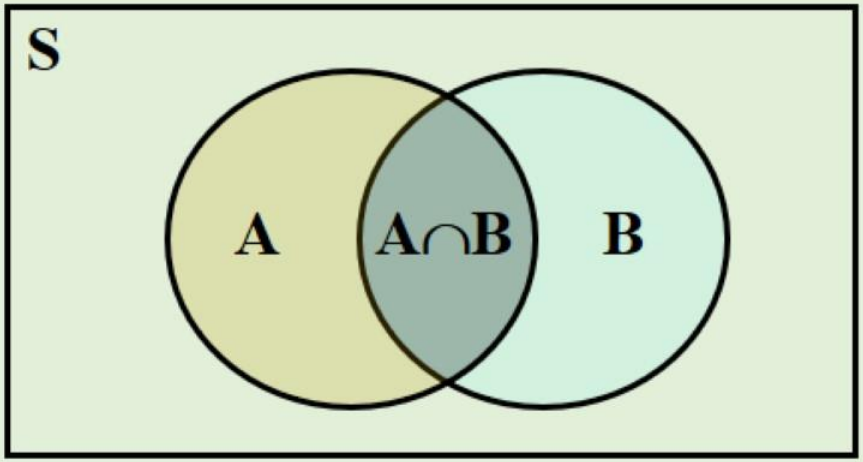
Basic Outcome – a possible outcome of a random experiment

Sample Space – the collection of all possible outcomes of a random experiment

Event – any subset of basic outcomes from the sample space

Important Terms: Intersection of Events

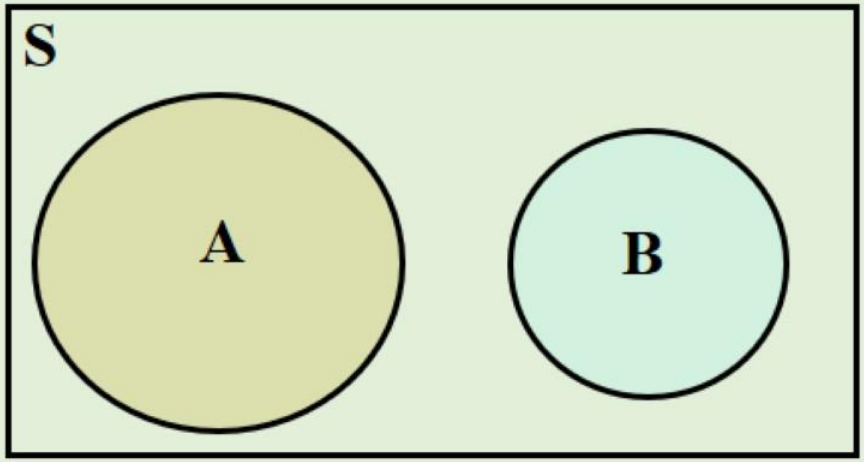
- **Intersection of Events** – If A and B are two events in a sample space S, then the intersection, $A \cap B$, is the set of all outcomes in S that belong to both A and B



Important Terms:
Mutually Exclusive Events

A and B are Mutually Exclusive Events if they have no basic outcomes in common

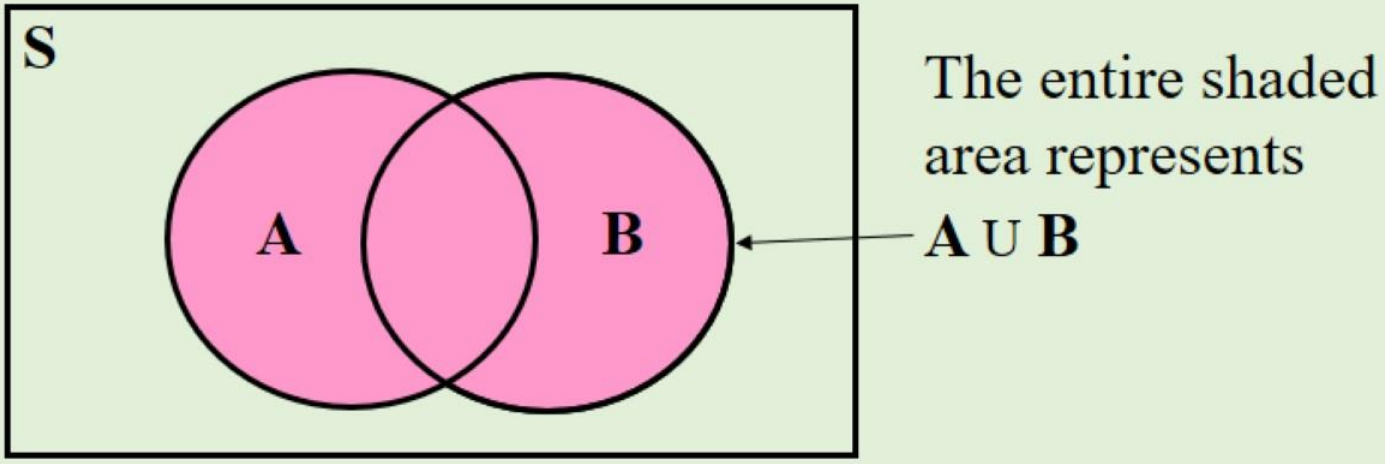
i.e., the set $A \cap B$ is empty



Important Terms:
Union of Events

- Union of Events – If A and B are two events in a sample space S, then the union, A U B, is the set of all outcomes in S that belong to either

  A or B

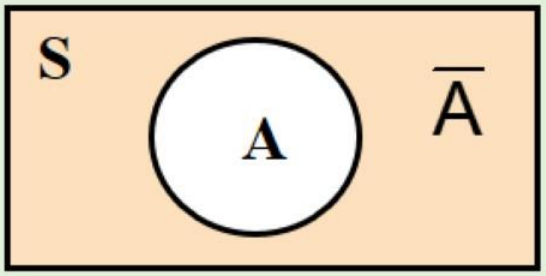


Important Terms: Complement of an Event

Events E_1, E_2, … E_k are Collectively Exhaustive events if $E_1 \cup E_2 \cup \ldots \cup E_k = S$

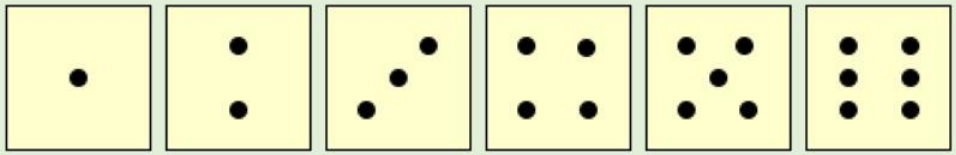
i.e., the events completely cover the sample space

The Complement of an event A is the set of all basic outcomes in the sample space that do not belong to A. The complement is denoted



Examples

Let the Sample Space be the collection of all possible outcomes of rolling one die:



$$S = [1, 2, 3, 4, 5, 6]$$

Let A be the event "Number rolled is even"

Let B be the event "Number rolled is at least 4"

$$A = [2, 4, 6] \quad \text{and} \quad B = [4, 5, 6]$$

Examples

$S = [1, 2, 3, 4, 5, 6]$ $A = [2, 4, 6]$ $B = [4, 5, 6]$

Complements:

$\overline{A} = [1, 3, 5]$ $\overline{B} = [1, 2, 3]$

Intersections:

$A \cap B = [4, 6]$ $\overline{A} \cap B = [5]$

Unions: $A \cup B = [2, 4, 5, 6]$

$A \cup \overline{A} = [1, 2, 3, 4, 5, 6] = S$

Examples

$S = [1, 2, 3, 4, 5, 6]$ $A = [2, 4, 6]$ $B = [4, 5, 6]$

Mutually exclusive:

A and B are not mutually exclusive

The outcomes 4 and 6 are common to both

Collectively exhaustive:

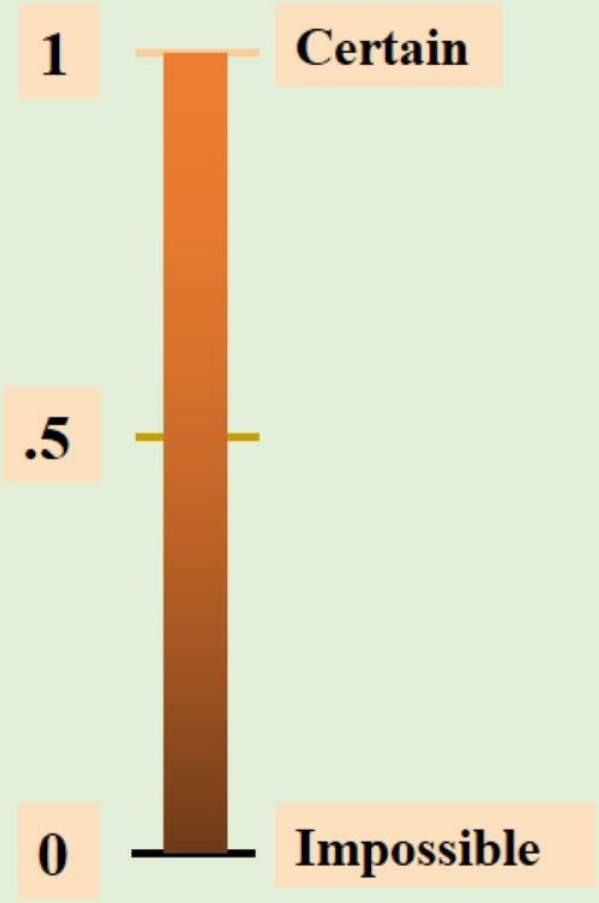
A and B are not collectively exhaustive

A U B does not contain 1 or 3

Probability

Probability – the chance that an uncertain event will occur (always between 0 and 1)

$$0 \leq P(A) \leq 1 \quad \text{For any event A}$$



Assessing Probability

There are three approaches to assessing the probability of an uncertain event:

Classical probability,

$$\text{probabilit y of event } A = \frac{N_A}{N} = \frac{\text{number of outcomes that satisfy the event}}{\text{total number of outcomes in the sample space}}$$

It is based on the assumption that the outcomes of an experiment are equally likely.

Using this classical viewpoint

Relative Frequency Concept

The probability of an event happening in the long run is determined by observing what fraction of the time like events happened in the past:

$$\text{Probability of event} = \frac{\text{Number of times event occured in the past}}{\text{Total number of observations}}$$

Subjective Probability

The likelihood (probability) of a particular event happening that is assigned by an individual based on whatever information is available.

An individual opinion or belief about the probability of occurrence

There will be a cold waive at the end of January.

Probability Postulates

1. If A is any event in the sample space S, then

$$0 \leq P(A) \leq 1$$

2. Let A be an event in S, and let O_i denote the basic outcome
(the notation means that the summation is over all the basic outcomes in A)

$$P(A) = \sum_{A} P(O_i)$$

3. $P(S) = 1$

EXAMPLE 1

Consider the experiment of tossing two coins once.

The sample space S = {HH, HT, TH, TT}

Consider the event of '*getting one head*'.

Event A = {HT, TH}

Probability of getting one head = 2/4 = 1/2.

EXAMPLE 2

- Throughout her career Professor Jones has awarded 186 A's out of the 1200 students she has taught. What is the probability that a student in her section this semester will receive an A?

- By applying the relative frequency concept, the probability of an A= 186/1200=.155

Mutually Exclusive and Exhaustive Events

Mutually Exclusive Events: The occurrence of any one event means that none of the others can occur at the same time.

Mutually Exhaustive Events: If the events are such that their union constitute the sample space. At least one of the events must occur when an experiment is conducted

Example: Mutually Exclusive and Exhaustive Events

Consider an experiment of tossing a fair coin

The sample space is : S = { 1, 2, 3, 4, 5, 6}

Consider three events: A, B, and C such that A: getting an ODD number, B: Getting an EVEN number, and C: getting a number above 3

So events, A = {1, 3, 5}, B = {2, 4, 6}, and C = {4, 5, 6}.

Here events A and B are mutually exclusive but A, C and B, C are not exclusive.

Here events A and B are mutually exhaustive but A, C and B, C are not exhaustive.

EXAMPLE 3

New England Commuter Airways recently supplied the following information on their commuter flights from Boston to New York:

Arrival	Frequency
Early	100
On Time	800
Late	75
Canceled	25
Total	1000

EXAMPLE 3 *continued*

If A is the event that a flight arrives early, then P(A) = 100/1000 = 0.1.

If B is the event that a flight arrives late, then P(B) = 75/1000 = 0.075.

If C is the event that a flight arrives on time, then P(C) = 800/1000 = .8.

If D is the event that a flight is canceled, then P(D) = 25/1000 = .025.

EXAMPLE 3 *continued*

The probability that a flight is either early or late is P(A or B) = P(A) + P(B) = 0.1 + 0.075 =0.175.

The probability that a flight either arrives on time or cancelled P(C or D) = P(C) + P(D) = 0.8 + 0.025 =0.825.

The Complement Rule

Complementary Event: If A is an event, then complement of A is denoted by A^c is the event that contain all the remaining outcome (Sample points) of the Sample space.

The complement rule of probability: If P(A) is the probability of event A and $P(A^c)$ is the probability of complement of A, then

$$P(A) + P(A^c) = 1$$

Hence, $P(A^c) = 1-P(A)$ OR $P(A) = 1-P(A^c)$.

A Probability Table

Marginal and joint probabilities for two events A and B are summarized in this table:

	B	$\overline{B}$	
A	$P(A \cap B)$	$P(A \cap \overline{B})$	$P(A)$
$\overline{A}$	$P(\overline{A} \cap B)$	$P(\overline{A} \cap \overline{B})$	$P(\overline{A})$
	$P(B)$	$P(\overline{B})$	$P(S) = 1.0$

Addition Rule Example

Consider a standard deck of 52 cards, with four suits:

♥ ♣ ♦ ♠

Let event A = card is an Ace

Let event B = card is from a red suit

Addition Rule Example

$$P(\text{Red} \cup \text{Ace}) = P(\text{Red}) + P(\text{Ace}) - P(\text{Red} \cap \text{Ace})$$

$$= 26/52 + 4/52 - 2/52 = 28/52$$

Type	Color		Total
	Red	Black	
Ace	2	2	4
Non-Ace	24	24	48
Total	26	26	52

Don't count the two red aces twice!

Conditional Probability

- A conditional probability is the probability of one event, given that another event has occurred:

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

→ The conditional probability of A given that B has occurred

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

→ The conditional probability of B given that A has occurred

Multiplication Rule

- Multiplication rule for two events A and B:

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

- also

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

Random Variables

A random variable is a numerical value determined by the outcome of an experiment.

EXAMPLE: Consider a random experiment in which a coin is tossed three times. Let X be the number of heads. Let H represent the outcome of a head and T the outcome of a tail.

EXAMPLE

- The *sample space* for such an experiment will be: TTT, TTH, THT, THH, HTT, HTH, HHT, HHH.
- Thus the possible values of X (number of heads) are x = 0,1,2,3.

EXAMPLE

- The outcome of zero heads occurred once.
- The outcome of one head occurred three times.
- The outcome of two heads occurred three times.
- The outcome of three heads occurred once.
- From the definition of a random variable, X as defined in this experiment, is a *random variable*.

Probability Distributions

A probability distribution is a listing of all the outcomes of an experiment and their associated probabilities. From previous example

Number of Heads	Probability of Outcomes
0	1/8 = .125
1	3/8 = .375
2	3/8 = .375
3	1/8 = .125
Total	8/8 = 1

Characteristics of a Probability Distribution

The probability of an outcome must always be between 0 and 1.

The sum of all mutually exclusive outcomes is always 1.

Thank you

for your kind attention and valuable time.

Follow us on LinkedIn:

https://www.linkedin.com/company/tqm-for-better-future

tqmforbetterfuture@gmail.com

Full Credit Goes To a Great Statistician Dr. M. Amir Hossain for the study materials.

Thank You!

#TQMforbetterFuture