Statistics Topic 8

- Definition: Methods of collection, organization and analysis of numerical information are collectively called **statistics.** Pieces of numerical and non-numerical information are called **data.** In order to collect data, you need to observe or measure some property; this property is called a **variable**.

- Definition: A variable is **qualitative** if it is not possible for it to take a numerical value. A variable is **quantitative** if it can take a numerical value. A quantitative variable which can take any value in a given range is called a **continuous** variable. A quantitative variable which has clear steps between its possible values is called a **discrete** variable.

- Types of Statistics: **Descriptive Statistics** comprises those methods concerned with collection and describing a set of data so as to yield meaningful information.

    - Example: Summarized large amounts of data collected from the pool games of 2011 cricket world cup to provide immediate meaningful information concerning the performance of each team.

Group A

Teams	Mat	Won	Lost	Tied	N/R	Pts	Net RR	For	Against
Pakistan	6	5	1	0	0	10	+0.758	1312/275.1	1155/288.0
Sri Lanka	6	4	1	0	1	9	+2.582	1336/218.4	882/250.0
Australia	6	4	1	0	1	9	+1.123	1181/218.5	1030/241.0
New Zealand	6	4	2	0	0	8	+1.135	1257/241.3	1156/284.0
Zimbabwe	6	2	4	0	0	4	+0.030	1288/288.0	1189/267.4
Canada	6	1	5	0	0	2	-1.987	1054/295.3	1582/284.5
Kenya	6	0	6	0	0	0	-3.042	932/300.0	1366/222.1

Group B

Teams	Mat	Won	Lost	Tied	N/R	Pts	Net RR	For	Against
South Africa	6	5	1	0	0	10	+2.026	1595/292.3	1028/300.0
India	6	4	1	1	0	9	+0.900	1673/282.3	1505/299.4
England	6	3	2	1	0	7	+0.072	1600/298.4	1576/298.1
West Indies	6	3	3	0	0	6	+1.066	1299/262.2	1138/292.5
Bangladesh	6	3	3	0	0	6	-1.361	1017/290.2	1276/262.2
Ireland	6	2	4	0	0	4	-0.696	1393/296.5	1595/296.0
Netherlands	6	0	6	0	0	0	-2.045	1182/300.0	1641/274.1

- **Inferential Statistics** comprises those methods concerned with analysis of a subset of data leading to predictions or inferences about the entire set of data.

    - Example: Suppose we collected data for 30 years regarding the average rainfall in the month of July in Lahore, and the amount came to be 3.3 centimeters. We make the inference that next year in the month of July we can expect between 3.2 and 3.4 centimeters of rain.

Where can Statistics Contribute?

- Business and Industry ○

Manufacturing

- Improve product quality
- Increase efficiency of processes ○ Marketing
- Conduct sample surveys
- Determine product viability
- Estimate advertisement effectiveness ○ Engineering
- Make consistent product
- Predict product life

○ Banking

- Estimate the risk of a company defaulting on loan
- Determine effective asset allocation for portfolios

- Health and Medicine ○ Epidemiology
  - Calculate cancer incidence rates
  - Monitor disease outbreaks
  - Study risk factors for various diseases ○ Public Health
  - Design community efforts
  - Education programs ○ Pharmacology
  - Drug discovery, development, and approval
  - Ensure validity of results in clinical trials
  - ○ Genetics
    - Identify potential indicators for specific diseases or traits ▪ Test gene modification for treatment of diseases

- Natural Resources ○ Agriculture
  - Evaluate differences in crop management
  - Evaluate the best combinations of fertilizers, pesticides and densities of planting.
  - ○ Ecology
    - Study changes in local and global climate
    - Develop strategies to improve the environment
    - Study the impact of new industrial plants on surrounding ecology ○ Geography
      - Evaluate the amount of rainfall one can expect for a given area based on longitude, latitude and distance from the sea.

- Social and Natural Sciences ○ Physics
  - Determine when an increase in the density of cosmic rays signals the presence of a supernova

        - Conducting tests to determine existence of new particles.
    - Chemistry
        - Predict shape of large molecules
        - Analysis of mass spectrometry data ○ Biology
        - Seek to better understand why insects cluster
        - Identify genes related to a particular disease ○ Sociology
        - Estimate the chances of a major war in the next five years
        - Study the increase in rates of marriage failure
        - Determine the characteristic of prisoners to study risk of them repeating criminal behavior. ○ Psychology
        - Study the effects of narcotics on schizophrenia
        - Determine if the existence of extrasensory perception can be demonstrated
        - Evaluate the relationship between shyness and loneliness ○ Anthropology
        - Determine the age of an archaeological site
        - Analyze the percentage difference in body fat between urban and rural dwellers in Pakistan ○ Zoology
        - Evaluate the differences in behavior of caged animals when they are outdoors and when they are indoors
        - Determine what techniques are more effective in counting a given species of bird.

- Education ○ Develop effective teaching strategies
    - Research appropriate and informative evaluation (testing) instruments ○ Identify risk factors for bullying, dropping out, failing
    - Identify factors contributing to a decrease/increase in student achievement
    - Study the proportion of graduates of various programs subsequently employed in their field of study

- Government Media and Law ○ Government Agencies
        - Design and implement effective sampling strategies
        - Estimate the unemployment rate
        - Track and report changes in the economy
        - Regulations on stock trading, drug approval, pollution ○ Journalism
        - Effective communication of statistical ideas to a broad audience
        - Participation in education efforts ○ Expert Witness
        - Testify in court cases involving DNA evidence, salary discrepancies, discrimination, and disease clusters

- Definitions: **A population** is defined as the set of all possible members of a stated group. A cross-section of the returns of all of the stocks traded on the New York Stock Exchange (NYSE) is an example of a population.

- Definition**: A sample** is defined as a subset of the population of interest. Once a population has been defined, a sample can be drawn from the population, and the sample's characteristics can be used to describe the population as a whole.

- Definition: A measure used to describe a characteristic of a population is referred to as a **parameter**.

- Definition: It is frequently too costly or time consuming to obtain measurements for every member of a population, if it is even possible. In this case, a sample may be used. In the same manner that a parameter may be used to describe a characteristic of a population, a **sample statistic** is used to measure a characteristic of a sample.

- Measurements Scales:

    - **Nominal scale:** Observations are classified or counted with *no particular order*. It consists of assigning items to groups or categories. *No quantitative information* is conveyed and *no ordering* of the items is implied. Nominal scales are therefore *qualitative* rather than quantitative.
        - Religious preference, race, and gender are all examples of nominal scales.

    - **Ordinal scale:** All observations are placed into separate categories and the *categories are placed in order with respect to some characteristic*. Differences between values makes no sense.
        - Political parties on left to right spectrum given labels 0, 1, 2; restaurant ratings, etc, are examples of ordinal scales.

    - **Interval scale:** This scale provides *ranking* and assurance that differences between scale values are equal. *Difference makes sense*, but ratio doesn't; and there is *no natural zero*.
        - temperature (C,F) and dates are examples of interval scale

    - **Ratio scale:** These represent the strongest level of measurement. In addition to providing ranking and equal differences between scale values, ratio scales have a *true zero point as the origin*.
        - Height, weight, age and length are all examples of ratio scale.

Representation of Data Topic 9

- Definition: A **frequency distribution** is a tabular presentation of statistical data that aids the analysis of large data sets. Frequency distributions summarize statistical data by assigning it to specified groups, or intervals.

  Three steps are required to construct a frequency distribution:
  - Define the intervals.
  - Tally the observations.
  - Count the observations.

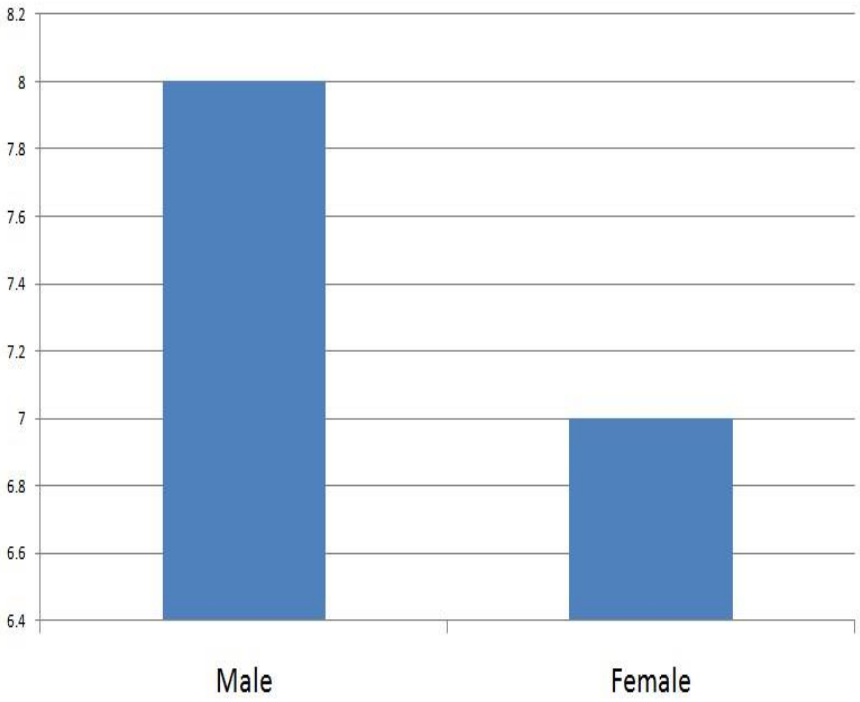
- Definition: **Relative frequency** is calculated by dividing the frequency of each interval by the total number of observations. Simply, relative frequency is the percentage of total observations falling within each interval.

- Definition**: Cumulative Frequency** is calculated by summing the frequencies starting at the lowest interval and progressing through the highest. Cumulative frequency for any given interval is the sum of the frequencies up to and including the given interval.

- Example:

Height Range	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
4.5 – 5.0 ft	25	25	0.25	0.25
5.0 – 5.5 ft	35	60	0.35	0.60
5.5 – 6.0 ft	29	89	0.29	0.89
6.0 – 6.5 ft	11	100	0.11	1.00

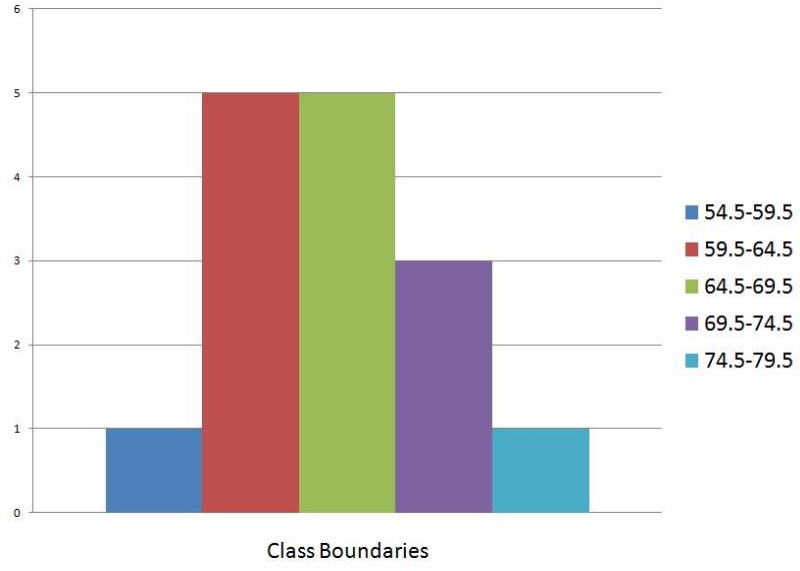
- Definition: A **Bar chart** graphically represents the data sets by representing the frequencies as heights of bars.
  - Example: Genderwise groupings of



students.

- Definition: A Bar chart which represents continuous data is called a **histogram** if
  - The bars have no spaces between them (though there may be bars with zero height, which could look like spaces).
    - The area of each bar is proportional to the frequency.
    - If all the bars have the same width, then the height is proportional to the frequency.

- Example: Height ranges of students in centimeters.

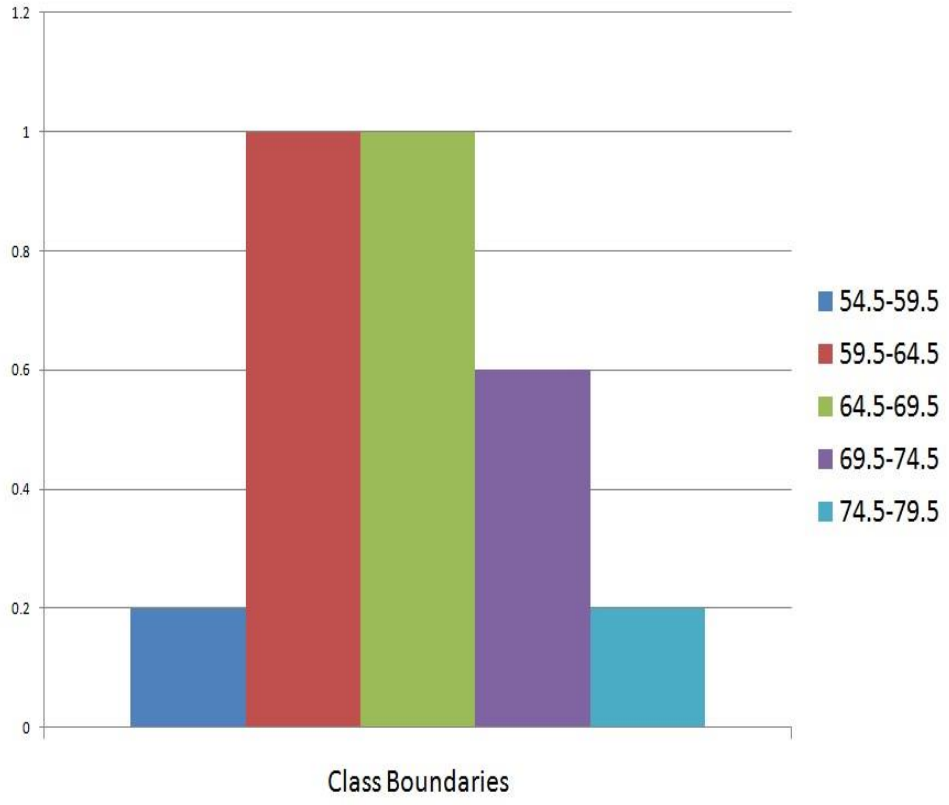


- Definition: **Frequency Density** is defined as the ratio between the frequency of a class and the class width. i.e. Frequency Density = Frequency/Class Width

- Example:

Height Range	Class Width	Frequency	Frequency Density
54.5 – 59.5	5	1	0.2
59.5 – 64.5	5	5	1.0
64.5 – 69.5	5	5	1.0
69.5 – 74.5	5	3	0.6
74.5 – 79.5	5	1	0.2

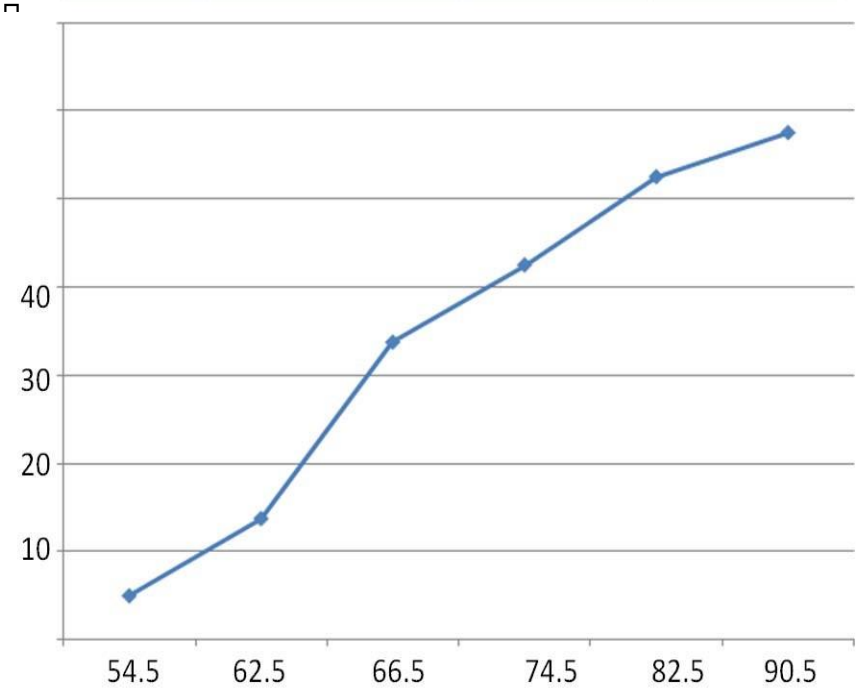
- Histogram: Frequency Density of height ranges of students



- Cumulative Frequency Graphs: Another way of representing continuous data is to draw a cumulative frequency graph. The cumulative frequencies are plotted against the upper class boundaries of the corresponding class

- Example: The grouped frequency distribution in the following table summarizes the masses in kilograms, measured to the nearest kilogram, of a sample of 38 students. Represent the data using a cumulative frequency graph

Mass (g)	Class Boundaries	Frequency	Cumulative Frequency
47 - 54	46.5 – 54.5	4	4
55 - 62	54.5 – 62.5	7	11
63 - 66	62.5 – 66.5	8	19
67 - 74	66.5 – 74.5	7	26
75 - 82	74.5 – 82.5	8	34
83 - 90	82.5 – 900.5	4	38



- We can use the cumulative frequency graph to get other information also. For example, what proportion of the students has mass less than 60 kg?

  From the graph the frequency is about 8.8.The proportion under 60 kg = 8.8/38 = 0.23 or 23%

- Pie Charts: Pie charts are useful for representing percentage allocation data.

  For example, the budget of a household can be represented effectively through a pie chart. Pie charts can be thought of as **circle graphs**. To calculate the quantities represented by each slice of the pie chart we need to take the angular fraction of the given total. i.e. If a particular frequency is *x*, and *n* is the total number, then the allocation angle for that class will be (*x*/*n*).*360*

- Example: In a survey, 90 people were asked to indicate which one of five musical instruments they played. The information is given in the following

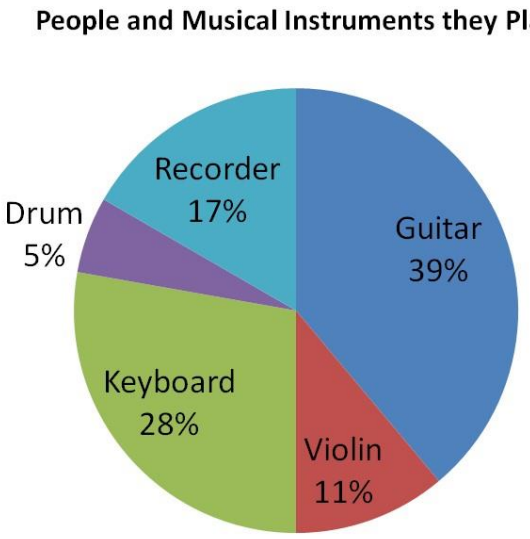
Instrument	Number of People
Guitar	35
Violin	10
Keyboard	25
Drum	5
Recorder	15

table.

Represent the data on a pie chart

Solution: We first calculate the allocation angle for each group:

Instrument	Number of People	Allocation Angle	Percentage
Guitar	35	140^o	39%
Violin	10	40^o	11%
Keyboard	25	100^o	28%
Drum	5	20^o	5%
Recorder	15	60^o	17%
TOTAL	90	360^o	100%



Measures of Central Tendency Topic 10

- Focus of Statistics:
  We are concerned in statistics with three things:
  - **Measures of Location**: where the data is clustered or centered. Also called measures of central tendency.
  - **Measures of Dispersion**: How the data is spread out from the center.
  - **Measures of Shape**: How the data sways and peaks.

- Measures of Central Tendency: The **central tendency** is measured by averages. These describe the point about which the various observed values cluster. In mathematics, an **average**, or **central tendency** of a data set refers to a measure of the "middle" or "expected" value of the data set.

- We will be studying the following measures of central tendency: Arithmetic Mean, Geometric Mean, Weighted Mean, Harmonic Mean, Median, and Mode.

- Definition: The **Arithmetic Mean** is the sum of the observation values divided by the number of observations. It is the most widely used measure of central tendency, and is the *only measure* where the *sum of the deviations* of each value from the mean is always *zero*. The formula for calculating the arithmetic mean of n values is: $\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$ or, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

- Arithmetic Mean of Grouped Data: If $z_1, z_2, z_3, \ldots, z_k$ are the mid-values and $f_1, f_2, f_3, \ldots, f_k$ are the corresponding frequencies, where the subscript 'k' stands for the number of classes, then the mean is $\bar{z} = \frac{\sum f_i z_i}{\sum f_i}$.

- Example: The math exam scores of 9 students are given below: 59, 66, 65, 74, 71, 67, 71, 62, 77.
  The mean score is: $\bar{x} = \frac{59 + 66 + 65 + 74 + 71 + 67 + 71 + 62 + 77}{9} = \frac{612}{9} = 68$

- Definition: The **Geometric Mean** is often used when calculating investment returns over multiple periods, or to find a compound growth rate. It is computed by taking the *nth* root of the product of n values. In general, when we are dealing with numbers that need to be multiplied, the geometric mean gives a more realistic picture than the arithmetic mean. The geometric mean for n values is:

$$GM = \sqrt[n]{a_1 a_2 \cdots a_n} = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}$$

- Geometric Mean of Grouped Data: If the "*n*" non-zero and positive variable values $x_1, x_2, \ldots\ldots, x_n$ occur $f_1, f_2, \ldots\ldots, f_n$ times, respectively, then the geometric mean of the set of observations is defined by: $GM = \left(x_1^{f_1} x_2^{f_2} \cdots\cdots\cdots x_n^{f_n}\right)^{\frac{1}{N}}$ where $N = \sum_{i=1}^{n} f_i$

$$\prod_{i=1}^{n} x_i^{f_i}$$

- Example (a): Company A has grown over the last 3 years by 10 million, 12 million and 14 million dollars. What is the average annual growth amount?

This is arithmetic mean = (10 + 12 + 14)/3 = 12 million dollars. • Example (b):

The profit of Company B has grown over last three years by 2.5%, 3%, and 3.5%.

What is the average growth rate?

This is geometric mean = $[(1.025)(1.030)(1.035)]^{1/3} - 1 = 2.9992\%$

- Definition**:** The **Weighted mean** is a special case of the mean that allows different weights on different observations. Formally, the weighted mean of a non-empty set of data, $x_1, x_2, \ldots, x_n$, with non-negative weights $w_1, w_2, \ldots, w_n$, is the quantity calculated by

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

Note: The arithmetic mean is weighted mean where all the weights equal 1

- Example: Given two school classes, one with 20 students, and one with 30 students, the grades in each class on a test were:

Morning class = 62, 67, 71, 74, 76, 77, 78, 79, 79, 80, 80, 81, 81, 82, 83, 84, 86, 89, 93, 98

Afternoon class = 81, 82, 83, 84, 85, 86, 87, 87, 88, 88, 89, 89, 89, 90, 90, 90, 90, 91, 91, 91, 92, 92, 93, 93, 94, 95, 96, 97, 98, 99

What is the average score of all the students if the average for the morning class is 80 and the average of the afternoon class is 90?

Solution: The straight average of 80 and 90 is 85, the mean of the two class means. However, this does not account for the difference in number of students in each class, and the value of 85 does not reflect the average student grade (independent of class). The average student grade can be

obtained by either averaging all the numbers without regard to classes as $x = \frac{4300}{50} = 86$, or weighting the class means by the number of students in each class: $x = \frac{20(80)+30(90)}{20+30} = 86$.

- Definition: The **Harmonic mean** is often used by investors to find the average cost of shares purchased over time. In certain situations, especially many situations involving *rates* and *ratios*, the harmonic mean provides the truest average. Formally, for a set of positive data values, $x_1, x_2, \ldots, x_n$, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals:

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

- Example (c)**:** You go on a 100 km trip. Suppose you travel half the **time** at 40 km/h and half the time at 60 km/h. What is the average speed?

  This is Arithmetic mean = (40 + 60)/2 = 50 km/hr

- Example (d): You go on a 100 km trip. Suppose you travel half the **distance** of your trip at 40 km/h, and the remaining half at 60 km/h**.** What is the average speed?

  This is Harmonic mean = 2/(1/40 + 1/60) = 48 km/hr

- Definition: The **median** is the middle value of the observations such that the number of observations above it is equal to the number of observations below it. If the number of values, n is odd, then the median is the middle value, i.e. $M_e = x_{\frac{1}{2}(n+1)}$; if n is even, then the median is the average of the middle two values, i.e. $M_e = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$

- Median of Grouped Data: For grouped data the median is given by: $M_e = L_o + h\left(\frac{\frac{n}{2} - F}{f_o}\right)$,

  where ○ L_0 = Lower class boundary of the median class ○

  h = Width of the median class ○ f_0 = Frequency of

  the median class ○ F = Cumulative frequency of the

  pre-median class

- Steps to Find the Median of Grouped Data:

- Compute the less than type cumulative frequencies. ○

  Determine n/2 , one-half of the total number of cases.

- Locate the median class for which the cumulative

  frequency is more than N/2 . ○ Determine the lower limit of

  the median class. This is L_0. ○ Sum the frequencies of all

  classes prior to the median class. This is *F*. ○ Determine the

  frequency of the median class. This is f_0.

- Determine the class width of the median class. This is *h*. ○

  Apply the Formula.

- Definition: **Mode** is the value of a distribution for which the frequency is maximum. In other words, mode is the value of a variable, which occurs with the highest frequency. •

  Example: The mode of the list (1, 2, 2, 3, 3, 3, 4) is 3. The mode is not necessarily well defined.

  The list (1, 2, 2, 3, 3, 5) has the two modes 2 and 3.

- Mode of Grouped Data: The formula for finding mode from grouped data is

$$M_0 = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} h$$, where

  - L_1 = Lower boundary of modal class ○ Δ_1 = difference of frequency between modal class and class before it ○ Δ_2 = difference of frequency between modal class and class after

  - *h* = class interval • Example: A commuter who travels to work by car has a choice of two different routes, V and W. He decides to compare his journey times for each route. So, he records the journey times, in minutes, for 10 consecutive working days, for each route. The results are:

Route V	53	52	48	51	49	47	42	48	57	53
Route W	43	41	39	108	52	42	38	45	39	51

- Calculate the means and medians for the two routes respectively.
- Which average do you think is more suitable for comparing the time taken on each route?

- Solution: For Route V: Mean = (53+52+48+51+49+47+42+48+57+53)/10 = 500/10 = 50 Arranging in ascending order, we get: 42, 47, 48, 48, 49, 51, 52, 53, 53, 57. So, the Median is (49 + 51)/2 = 50

  For Route W: Mean = (43+41+39+108+52+42+38+45+39+51)/10 = 498/10 = 49.8. Arranging in ascending order, we get: 38, 39, 39, 41, 42, 43, 45, 51, 52,108. So, the Median is (42 + 43)/2 = 42.5

  Route W is quicker.

- Example: Consider the two sets of data A and B

  A: 48, 52, 60, 60, 60, 68, 72
  B: 0, 10, 60, 60, 60, 110, 120

  For both sets of data the mean = median = mode = 60. Are the two data sets the same?

- If you were given nothing but the measures of location, you might be tempted to think that the two sets of data are similar. But B is much more spread out than A. It is necessary for us to devise some new measures to summarize the **spread of data**.

Measures of Dispersion Topic 11

- The Measures of Dispersion we will consider are:
  - Range
  - Inter Quartile Range ○ Mean Absolute Deviation ○ Variance and Standard Deviation.

- Definition: The **Range** is defined as the difference between the largest and the smallest data values. i.e. Range = Largest Value – Smallest Value
  - In the previous example, the range for A is: 72 – 48 = 24 and the range for B is: 120 – 0 = 120.

- Limitations of Range: Range is limited as it does not give the pattern of distribution in the middle.
    - Example:

        Dataset C: 2, 3, 4, 6, 8, 9 10

        Dataset D: 2, 5, 6, 6, 6, 7, 10

        Both the datasets have same mean, same median and same range. But the data patterns are different. In D, data is clustered around the center, but in C it is farther away from the center.

        We need more accurate measures of dispersion.

- Definition: The **quantiles** are values which divide the distribution such that there is a given proportion of observations below the quantile. It is mostly used to compare one's results with others.

- Commonly used quantiles are:
    - Quartiles: The distribution is divided into quarters. ○ Quintiles: The distribution is divided into fifths. ○ Deciles: The distribution is divided into tenths.
    - Percentile: The distribution is divided into hundredths (percents).

- At a given percentile, ***y***, with ***n*** data points **sorted in ascending order**, the position of the observation: $L_y = (n + 1)\ (y/100)$. Then from the dataset we can obtain P_y which is the $L_y th$ value
    - Example, For 11 observations, the third quartile (point below which 75% of observations lie) can be identified as the 9^{th} value since $L_y = (11 + 1)\ (75/100) = 9$

- Example: Consider the data set: 47 35 37 32 40 39 36 34 35 31 44 ○ Find the 75^{th} percentile point. ○ Find the 1^{st} quartile point.
    - Find the 5^{th} decile point.

    Solution: First arrange the data in ascending order: 31, 32, 34, 35, 35, 36, 37, 39, 40, 44, 47

    ○ Location of the 75^{th} percentile is the $L_{75} = (11 + 1)\ (75/100) = 9^{th}$ value. i.e. $P_{75}=40$ ○ Location of the 1^{st} quartile is the $L_{25} = (11 + 1)\ (25/100) = 3^{rd}$ value. i.e. $P_{25}=34$ ○ Location of the 5^{th} decile is the $L_{50} = (11 + 1)\ (50/100) = 6^{th}$ value. i.e $P_{50}=36$.

- Definition: Recall, the first quartile, denoted by Q_1, is the value below which 25% of the values lie, while the third quartile, denoted by Q_3, is the value below which 75% of the values lie. The **inter-quartile range, IQR**, is defined as $Q_3 – Q_1$.

IQR contains information about the spread of the middle 50% of the data.

Note: the second quartile, Q_2 = Median

- Another way to find Quartiles:
    - Case1: If there are even number of data values.
        - Arrange the data in ascending order
        - Split the data into upper half and lower half
        - The median of the upper half is Q_3 and the median of the lower half is Q_1
    - Case2: There are odd number of data values.
        - Arrange the data in ascending order
        - Find the median Q_2 and delete it from the list to get even number of values.
        - Use the method of case one to find Q_1 and Q_3

- How to Read the Quartiles from a cumulative graph: Results from two traffic surveys with 50 cars each taken at two observation points A and B are summarized below.

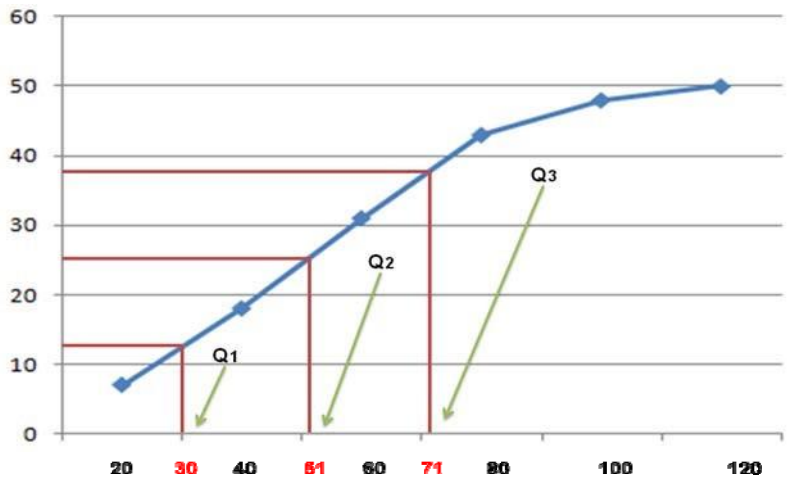
Speed	A Frequency	B Frequency	A C.Frequency	B C.Frequency
$0 \le x < 20$	7	1	7	1
$20 \le x < 40$	11	3	18	4
$40 \le x < 60$	13	5	31	9
$60 \le x < 80$	12	20	43	29
$80 \le x < 100$	5	18	48	47
$100 \le x < 120$	2	3	50	50
TOTALS	50	50		

Find the Median and the IQR

- Solution: Q_1 is the value below which 25% of the values lie. i.e. it is the value corresponding to CF = 0.25(50) = 12.5

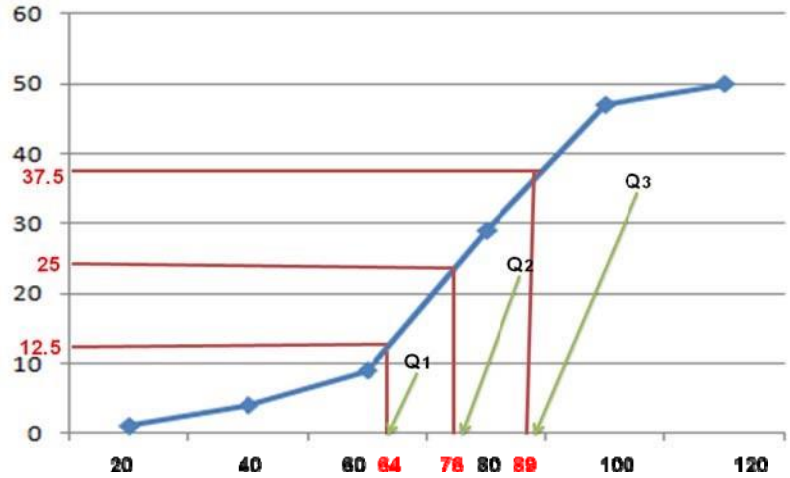
  Q_3 is the value below which 75% of the values lie. i.e. it is the value corresponding to CF = 0.75(50) = 37.5

  Q_2 is the value below which 50% of the values lie. i.e. it is the value corresponding to CF = 0.5(50) = 25 For A:



Median = 51, Q_1 = 30, Q_3 = 71, IQR = 41

For B:



Median = 76, Q_1 = 64, Q_3 = 89, IQR = 25
Median speed at A is lower than the median speed at B, and the IQR at A is greater than the IQR at B.
So, at A, the cars go more slowly and there is greater variation in their speeds

- Five Number Summary: One helpful way to summarize data is to give values which provide essential information about the data set. One such summary is called the **five-number**

summary. This summary gives the median, Q_2, the lower quartile, Q_1, the upper quartile, Q_3, the minimum value, *m*, and the maximum value, *M*.

- Box and Whiskers Plot: The five-number summary can be converted into a useful diagram, called the **box-and-whisker** plot, or a **boxplot**. Steps are:
  - Draw a scale (horizontal or vertical)
  - Above the scale draw a box (rectangle) in which the left side is above the point corresponding to Q_1 and the right side is above the point corresponding to Q_3. Then mark a third line inside the box above the point corresponding to Q_2.
  - Draw the two whiskers: the left whisker extends from Q_1 to *m*, and the right whisker extends from Q_3 to *M*.

- Example: The data below give the number of fish caught each day over a period of 11 days by a fisherman. Give a five number summary of the data, and plot a box and whiskers diagram:
  0, 2, 5, 2, 0, 4, 4, 8, 9, 8, 8

  Solution: Rearrange the data in ascending order: 0, 0, 2, 2, 4, 4, 5, 8, 8, 8, 9
  The median value is the 6^{th} value i.e. $Q_2 = 4$
  Lower quartile is the 3^{rd} value, i.e. $Q_1 = 2$
  Upper quartile is the 9^{th} value, i.e. $Q_3 = 8$
  The five-number summary is then: $m = 0$, Q1 = 2, Q2 = 4, Q3 = 8, $M = 9$
  Here is the box-and-whisker plot for the distribution of the number of fish caught over 11 days



- So far we have only looked at dispersion measuring methods that measure ranges of values. We need dispersion measures that measure variation for **each value** in the dataset.

- Definition: **Mean absolute deviation** (MAD) is the average of the absolute values of the deviations of individual observations from the arithmetic mean. $MAD = \frac{1}{n}\sum |X - \bar{X}|$

Example:

X	$X-\bar{X}$	$\lvert X-\bar{X}\rvert$
49	-9	9
56	-2	2
55	-3	3
68	10	10
61	3	3
57	-1	1
61	3	3
52	-6	6
63	5	5
Total	**0**	**42**

M.A.D. = 42/9 = 4.66

- Limitations of MAD ○ Absolute values are not very easy to manipulate algebraically.
  ○ Its graph is not “smooth” and that causes problems in dealing with it in calculus.

- Definition: The **Variance** of a dataset is the mean of the squared deviations from the mean. It is calculated using the formula: $Var\,X(\quad) = \frac{1}{n}\sum(X - X)^2$

  Standard deviation is the positive square root of the variance. i.e.

  $SD = \sqrt{\frac{1}{n}\sum (X - \bar{X})^2}$

- It can be shown through algebraic means that the variance formula can be written as

  $Var\,X(\quad) = \frac{1}{n}\sum(X - X)^2 = \frac{1}{n}\sum X^2 - (X)^2$

- Example:

X	$X-\bar{X}$	$(X-\bar{X})^2$	X^2
49	-9	81	2401
56	-2	4	3136
55	-3	9	3025
68	10	100	4624
61	3	9	3721
57	-1	1	3249
61	3	9	3721
52	-6	36	2704
63	5	25	3969
Total	**0**	**274**	**30550**

Using the definition: $\sigma^2 = \frac{1}{n}\sum (X - \bar{X})^2 = \frac{1}{9}(274) = 30.4$

Using the alternative formula:

$$\sigma^2 = \frac{1}{n}\sum X^2 - (\bar{X})^2 = \frac{1}{9}(30550) - 58^2 = 3394.4 - 3364 = 30.4$$

- Example: 12 boys and 13 girls, in a class of 25 students were given a test. The mean and SD of the 12 boys were 31 and 6.2 respectively; the mean and SD of the 13 girls were 36 and 4.3 respectively. Find the mean marks and SD of the whole class of 25 students.

Solution: Let the Boys' marks be represented by $x_1, x_2, \ldots x_{12}$. Let the Girls' marks be represented by $y_1, y_2, \ldots y_{13}$.

We are given $\bar{x} = \frac{\sum x_i}{12} = 31$. This means that $\sum x_i = (31)(12) = 372$. Given that SD of Boys is 6.2, the variance of Boys = 38.44. This means that $\frac{\sum x_{i2}}{12} - 31^2 = 38.44$. This gives

$$\sum x_i^2 = 12(38.44 - 31)^2 = 11993.28$$

We are also given $\bar{y} = \frac{\sum y_i}{13} = 36$. This means that $\sum y_i = (36)(13) = 468$

Given that SD of Girls is 4.3, the variance of Girls = 18.49. This means $\frac{\sum y_{i2}}{13} - 36^2 = 18.49$,

which gives $\sum y_i^2 = 13(18.49 + 36)^2 = 17088.37$

$$\sum x_i \quad \sum y_i \quad 372 \quad 468 \quad 840$$

The overall mean is $\frac{}{25} \quad \frac{}{25} \quad \frac{}{25} = 33.6$

The overall variance is $\frac{\sum x_i^2 \quad \sum y_i^2}{25} - (33.6)^2 = \frac{11993.28}{25} - 17088.37 - (33.6)^2 = 34.306$

So, the overall SD = $\sqrt{34.306} = 5.86$

- Variance for Grouped Data: The formula for variance of grouped data is:

$$\text{Variance} = \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2$$

- Definition: If the mean of a sample statistic is equal to the population parameter, the sample statistic is called an **unbiased estimator** of the population parameter. The sample variance becomes unbiased if we divide by *n – 1* instead of *n*.

- Estimate of Population Variance: When a sample is used to actually estimate the population variance there is a different formula for the sample variance and the sample standard deviation

  The unbiased estimator of the population variance is $s^2 = \frac{1}{n-1} \sum (X - \bar{X})^2$

  And the unbiased estimator of the population standard deviation is $s = \sqrt{\frac{1}{n-1} \sum (X - \bar{X})^2}$

- Chebyshev's Theorem: For any set of observations, the proportion of the observations within *k* standard deviations of the mean is at least: $1 - (1/k^2)$ for all $k > 1$.

- Chebyshev's inequality can be used to measure maximum amount of dispersion, regardless of the shape of the distribution.

- Example: A distribution has a mean of 3. If 75% of all of its observations lie between 1.25 and 4.75, what is the standard deviation of this distribution?

  Solution: We observe that the two values – 1.25 and 4.75 - are equidistant from the mean of

3. In other words,
$3 - k\sigma = 1.25$ and
$3 + k\sigma = 4.75$.
This gives us $2k\sigma = 3.5$ or $k\sigma = 1.75$
Now, by Chebyshev's theorem, we have $1 - (1/k^2) = 0.75$. This gives $k = 2$. Therefore $2\sigma = 1.75$, or $\sigma = 0.875$

- Definition: The **coefficient of variation** expresses how much dispersion exists relative to the mean of a distribution and allows for direct comparison of dispersion across different data sets. It is used in investment analysis to compare *relative risks*.

  $CV = \frac{s}{\bar{X}}$ = standard deviation of x / average value of x.

- Example: Investment A has a mean return of 7% and an std dev of 0.05. Investment B has a mean return of 12% and a std dev of 0.07. Which is riskier?

  Solution: A's CV is 0.05 / 0.07 = .714; B's CV is 0.07 / 0.12 = 0.583
  A has 0.714 units of risk for each unit of return while B has 0.583 units of risk for each unit of return. A has more risk per unit of return.

Skewness and Kurtosis Topic 12

- Definition: The first four **central moments about the mean** are defined as follows:
  - $\mu_1 = \frac{\sum (x - \bar{x})}{n}$
  - $\mu_2 = \frac{\sum (x - \bar{x})^2}{n}$
  - $\mu_3 = \frac{\sum (x - \bar{x})^3}{n}$
  - $\mu_4 = \frac{\sum (x - \bar{x})^4}{n}$

- Definition: **Skewness, or skew,** refers to the extent to which a distribution is **not symmetrical**. Non-symmetrical distributions may be either **positively or negatively skewed** and result from the occurrence of outliers in the data set.

  - **A positively skewed distribution** is characterized by many **outliers in the upper region, or right tail**. A positively skewed distribution is said to be skewed right because of its relatively **long upper (right) tail**.

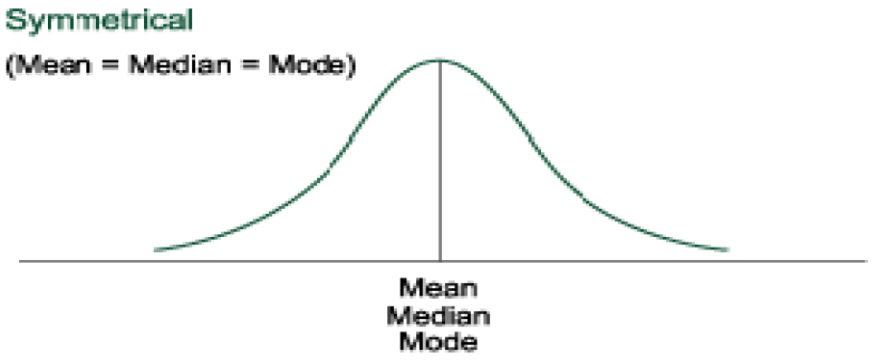
- **A negatively skewed distribution** has a disproportionately large amount of **outliers that fall within its lower (left) tail**. A negatively skewed distribution is said to be skewed left because of its **long lower tail**.

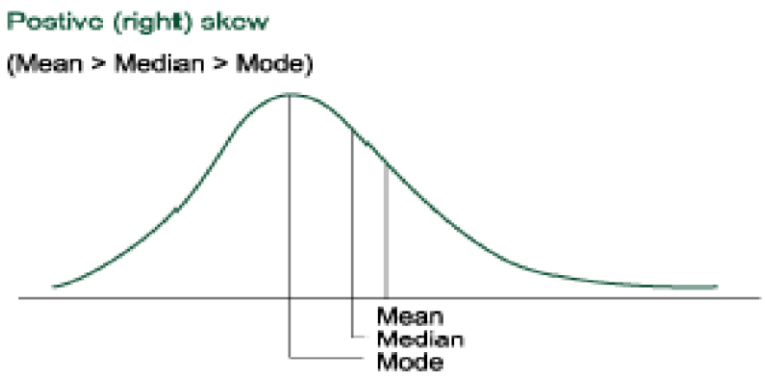
- Definition: The **moment coefficient of skewness, β_1**, is defined as μ_3^2 / μ_2^3, where the sign of skewness is determined by the sign of β_1 and the degree of skewness is determined by the absolute value of β_1.

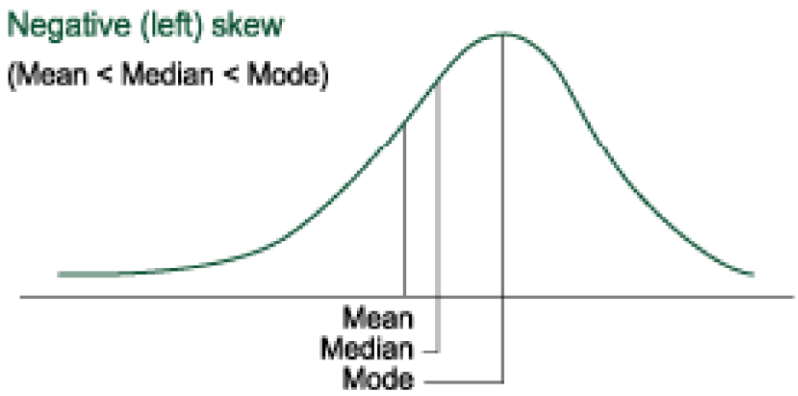
- Skewness and Measures of Central Tendency: Skewness affects the location of the **mean, median,** and **mode** of a distribution as summarized in the following bulleted list.

  - For a **positively skewed** distribution, the **mode is less than the median, which is less than the mean**. The mean is affected by outliers; in a positively skewed distribution, there are large, positive outliers which will tend to "pull" the mean upward, or more positive.

  - For a **negatively skewed** distribution, the **mean is less than the median, which is less than the mode**. In this case, there are large, negative outliers which tend to "pull" the mean downward (to the left).

- 





The diagrams illustrate an empirical relationship between mean, Median and the Mode. They suggest that the distance between the Mean and the Mode is three times the distance between the Mean and the Median, i.e. **Mean – Mode = 3(Mean – Median)**

Note: This is only an empirical and approximate relationship. There is no mathematical proof for it, and one can find counter examples where this is not true. However, in most smoothly skewed distributions this relationship holds. • Example: (a) Given Mean = 20, Mode = 15 find the value of Median. (b) Given Mode = 25, Median = 20, find the value of Mean.

Solution: (a) Mode = 3 Median — 2 Mean

15 = 3 Median — (2 x 20)

15 = 3 Median — 40

3 Median = 55 or Median = 18.3

(b) Mode = 3 Median — 2 Mean

25 = (3x20) — 2 Mean

25 = 60—2 Mean, or 2 Mean = 35

Mean = 17.5

- Another measure of skewness is the **Pearson's coefficient of skewness**:

$$\text{Pearson's coefficient of skewness} = \frac{\text{Mean-Mode}}{\text{Standard deviation}} = \frac{3(\text{Mean-Median})}{\text{Standard deviation}}$$

- Example: Find the pearson's Coefficient of Skewness from the following data:

Income of daily Labour	Frequency (f)	(x)
40-50	3	45
50-60	5	55
60-70	10	65
70-80	8	75
80-90	4	85
90-100	4	95
100-110	1	105
Sum	**35**	

Solution:

Income of daily Labour	Frequency (*f*)	(x)	*xf*	x^2f
f40-50	3	45	135	6075
50-60	5	55	275	15125
60-70	10	65	650	42250
70-80	8	75	600	45000
80-90	4	85	340	28900
90-100	4	95	380	36100
100-110	1	105	105	11025
Sum	**35**		**2485**	**184475**

Mean = 2485/35 = 71

$M_0 = L_1 + \frac{f_1}{f_1 + f_2} \times h = 60 + \frac{5}{5 + 2} \times (10) = 67.14$

$sd = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i} - (\bar{x})^2} = \sqrt{\frac{184475}{35} - (71)^2} = \sqrt{229.71} = 15.16$

$S_K = \frac{\text{(Mean-Mode)}}{\text{Standard deviation}} = \frac{(71 - 67.14)}{15.16} = 0.255$

The distribution is positively skewed.

- Measuring Peakedness: **Kurtosis** is a measure of the degree to which a distribution is more or less "peaked" than a normal distribution.

- The most important measure of kurtosis based on the second and fourth moments is

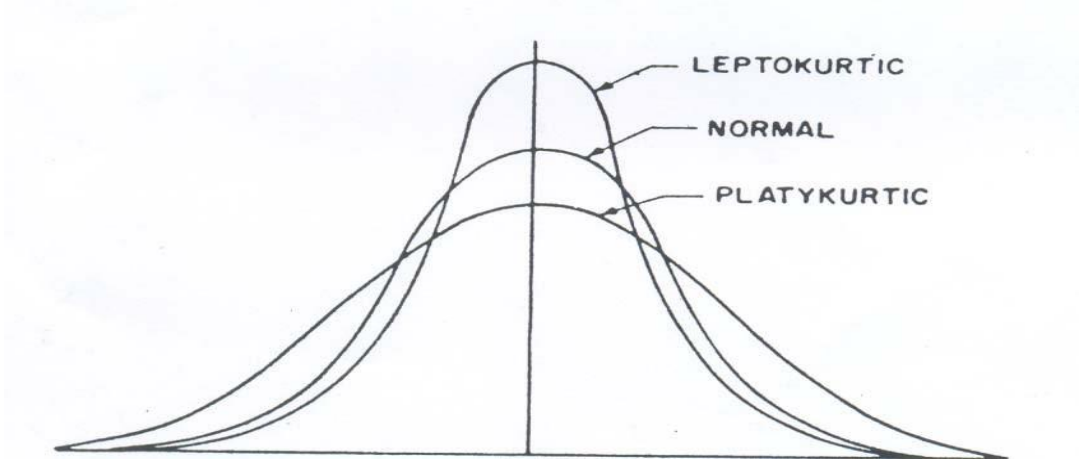
$$\frac{n\sum (x - \bar{x})^4}{\left[\sum (x - \bar{x})^2\right]^2}$$

If $\beta_2 - 3 = 0$, the distribution is **leptokurtic**. If $\beta_2 - 3 = 0$, the

distribution is **platykurtic**. ○ If $\beta_2 - 3 = 0$, the

distribution is **mesokurtic**.

These pictures are illustrated in the following diagram:



Leptokurtic distribution has a **higher peak** than the normal distribution and **fatter tails** than normal distribution.

Platykurtic distribution has a **flatter peak** than the normal distribution and **thinner tails** than the normal distribution.

Probability Concepts Topic 13

- Probability is the calculation of chance of a particular hypothesis being true or false.
- Example: Hypothesis: “Head will turn up when I toss a fair coin once” o The chances that this hypothesis will be true is 0.5 or 1 in 2.
    - We say that the probability that heads will show is 0.5.
- Definition: An **event** is a specified set of outcomes or a particular hypothesis.
    - We usually denote events by capital letters such as A, B, C, P, Q, R, etc.
- Notation: We denote probability by capital P, and the probability of an event A by P(A).

  Example: In the above example if we say that A is the event that head will turn up when we toss a fair coin once, then we write P(A) = 0.5
- An event can be a single outcome or a combination of outcomes. The following are all examples of events:
    - A = Head shows up when we toss a coin o B = Head shows up on the second and the third toss of a coin.
    - C = A number greater than 3 appears when we roll a die.
- Definition: Two events are **mutually exclusive** if they cannot happen at the same time. •

  Example: When we toss a coin, the events H {= Head shows up} and T {=Tail shows up} are mutually exclusive as they both cannot happen at the same time.
- Definition: A collection of events is **exhaustive** if all the possible elementary events under the experiment are covered by the event(s) considered together. In other words, the events are said to be exhaustive when they are such that at least one of the events compulsorily occurs.

- Example: When we toss a coin, the events H and T are exhaustive as they cover all the possibilities of the experiment.

- Definition: A **Probability Space** or **Sample Space** is the set of all the possible outcomes in an experiment.

- Example: When we toss a fair coin three times, we get the following Sample Space:

  $\{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

- The probability of an event A is defined by: $P(A) = \frac{\text{\# of ways in which A can occur}}{\text{\# of events in the sample space}}$

- Characteristics of a Probability Distribution:
  - The probability of any event A lies between 0 and 1, i.e. $0 \le P(A) \le 1$
  - The sum of probabilities of all the mutually exclusive and exhaustive events in a probability distribution equals 1.

- Definition: Two events A and B are **independent** if knowledge of one event happening does not affect the probability of occurrence of the other event.

- Example: Suppose an experiment consists of tossing a fair coin twice. Let A be the event that Head turns up on the first toss and let B be the event that Tail turns up on the second toss. Then A and B are independent as the occurrence of one does not affect the occurrence of the other.

- Definition: If A is an event, then A^c (or A') is called the **complement** of A, and $P(A) + P(A^c) = 1$

- Multiplication Rule: The probability that event A **and** event B occur together is given by the multiplication rule: $P(A \cap B) = P(A) \cdot P(B)$

- Addition Rule: The probability that event A **or** event B occur is given by the addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Example: The following table summarizes the availability of trucks with air bags and bucket seats at a dealership.

	Bucket Seats	No Bucket Seats	*Total*
Air Bags	75	50	***125***
No Air Bags	35	60	***95***
Total	***110***	***110***	***220***

What is the probability of selecting a truck at random that has either air bags or bucket seats?

Solution: The addition rule for probabilities is used to determine the probability of at least one event among two or more events occurring. The probability of each event is added and the joint probability (if the events are not mutually exclusive) is subtracted to arrive at the solution. P(air bags or bucket seats) = P(air bags) + P(bucket seats) - P(air bags and bucket seats) = (125/220) + (110/220) - (75/220) = 0.57 + 0.50 0.34 = 0.73 or 73 percent.

Alternative: 1 - P(no airbag and no bucket seats) = 1 - (60/220) = 73%

© Copyright Virtual University of Pakistan