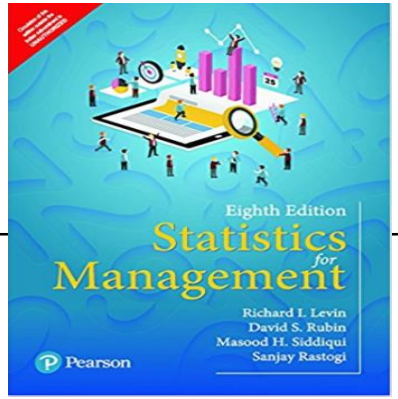
Chapter One

Introduction to Statistics

Course Summary

- Descriptive Statistics: Getting a better sense of data
  - Mean, Standard Deviation, Median, Quartiles, Distribution
- Inferential Statistics: Drawing conclusions about the population based on sample data
  - Properties of a single variable

Book for the Course



- Textbook is much more exhaustive than what we will cover in five weeks
- The best use of the book is as a reference, go to specific sections of chapter where you need more clarity
- First solve the exercises from the textbook before thinking of more practice problems

Definitions:

¬It is a science which helps us to **collect, analyze** and **present data** systematically.

¬It is the process of **collecting, processing, summarizing, presenting, analysing** and **interpreting of data in order to study** and **describe a given problem.**

¬Statistics is the **art of learning from data**.

¬Statistics may be regarded as (i)the study of populations, (ii) the study of variation, and (iii) the study of methods of the reduction of data.

Importance of Statistics:

- It simplifies mass of data (condensation);
- Helps to get concrete information about any problem;
- Helps for reliable and objective decision making;
- It presents facts in a precise & definite form;
- It facilitates comparison(Measures of central tendency and measures of dispersion);
- It facilitates Predictions (Time series and regression analysis are the most commonly used methods towards prediction.);
- It helps in formulation of suitable policies;

Limitation of statistics:

- Statistics does not deal with **individual items;**
- Statistics deals only with **quantitatively expressed items**, it does not study qualitative phenomena;
- Statistical results are **not universally true**;
- Statistics is **liable/responsible/ to be misused**.

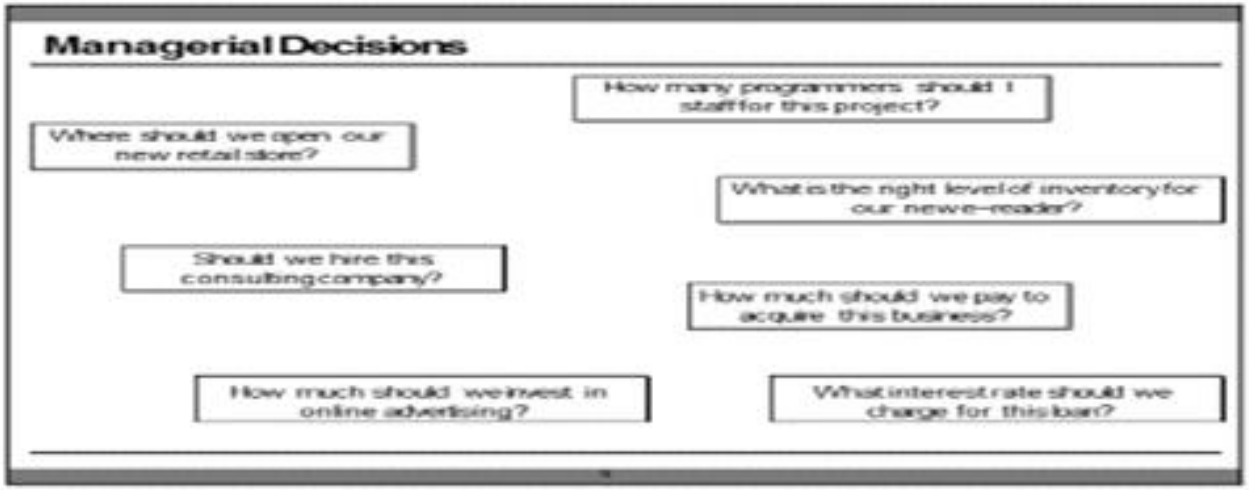
Application areas of statistics

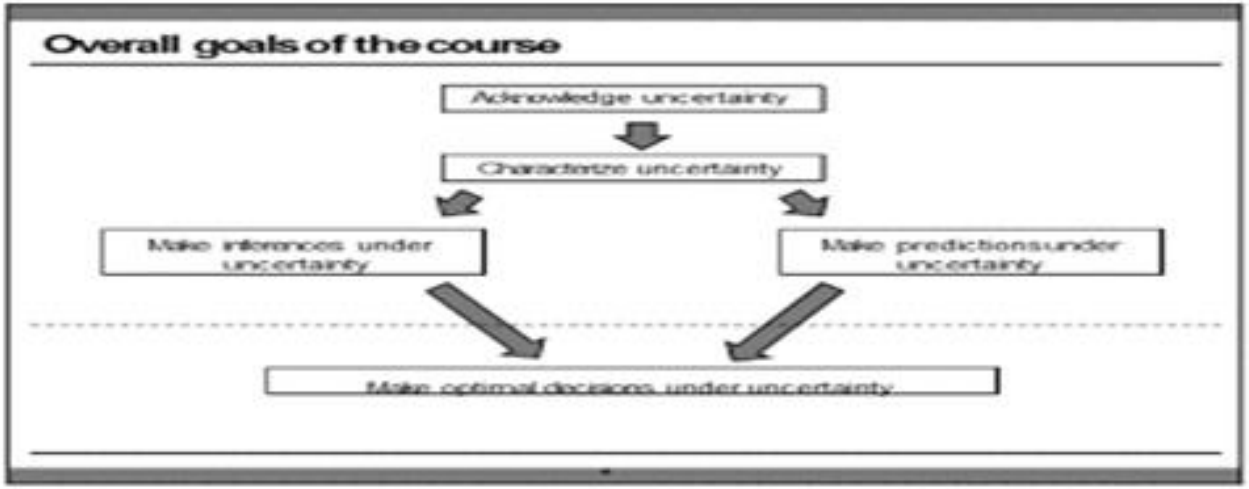
¬Engineering:

Improving product design, testing product performance, determining reliability and maintainability, working out safer systems of flight control for airports, etc.

¬Business:

Estimating the volume of retail sales, designing optimum inventory control system, producing auditing and accounting procedures, improving working conditions in industrial plants, assessing the market for new products.


Managerial Decisions
How many programmers should I staff for this project?
Where should we open our new retail store?
What is the right level of inventory for our new e-reader?
Should we hire this consulting company?
How much should we pay to acquire this business?
How much should we invest in online advertising?
What interest rate should we charge for this loan?


Overall goals of the course
Acknowledge uncertainty
Characterize uncertainty
Make inferences under uncertainty
Make predictions under uncertainty
Make optimal decisions under uncertainty

¬Quality Control:

Determining techniques for evaluation of quality through adequate sampling, in process control, consumer survey and experimental design in product development etc.

*Realizing its importance, large organizations are maintaining their own **Statistical Quality Control Department** *.

¬Economics:

Measuring indicators such as volume of trade, size of labor force, and standard of living, analyzing consumer behavior, computation of national income accounts, formulation of economic laws, etc.

Particularly, Regression analysis extensively uses in the field of Economics.

- **Health and Medicine:**

  Developing and testing new drugs, delivering improved medical care, preventing diagnosing, and treating disease, etc. Specifically, inferential Statistics has a tremendous application in the fields of health and medicine.

- **Biology:**

  Exploring the interactions of species with their environment, creating theoretical models of the nervous system, studying genetically evolution, etc.

- **Psychology:**

  Measuring learning ability, intelligence, and personality characteristics, creating psychological scales and abnormal behavior, etc.

- **Sociology:**

  Testing theories about social systems, designing and conducting sample surveys to study social attitudes, exploring cross-cultural differences, studying the growth of human population, etc.

There are two main branches of statistics:

1. Descriptive statistics
2. Inferential statistics

1. Descriptive statistics:

- It is the **first phase** of Statistics;
- involve any kind of data processing designed to the collection, organization, presentation, and analyzing the important features of the data **with out attempting to infer/conclude any thing that goes beyond the known data.**
- In short, descriptive Statistics describes the nature or characteristics of the observed data (usually a sample) **without making conclusion or generalization.**

The following are some examples of descriptive Statistics:

- The daily average temperature range of AA was 25 0c last week .
- The maximum amount of coffee export of Eth. (as observed from the last 20 years) was in the year 2004.
- The average age of athletes participated in London Marathon was 25 years.
- 75% of the instructors in AAU are male.
- The scores of 50 students in a Mathematics exam are found to range from 20 to 90.

2. Inferential statistics (Inductive Statistics):

- It is a **second phase** of Statistics which deals with techniques of making a **generalization** that lie outside the scope of **Descriptive Statistics;**
- It is concerned with the process of **drawing conclusions (inferences)** about specific characteristics of a population based on information obtained from samples;
- It is a process of performing hypothesis testing, determining relationships among variables, and making predictions.
- **The area of inferential statistics entirely needs the whole aims to give reasonable estimates of unknown population parameters.**

The following are some examples of inferential Statistics:

- The result obtained from the analysis of the income of 1000 randomly selected citizens in Ethiopia suggests that the average monthly income of a citizen is estimated to be 600 Birr.
- Here in the above example we are trying to represent the income of about entire population of Ethiopia by a sample of 1000 citizens, hence we are making inference or generalization.
- Based on the trend analysis on the past observations/data, the average exchange rate for a dollar is expected to be 18 birr in the coming month.
- The national statistical Bereau of Ethiopia declares the out come of its survey as "The population of Eth. in the year 2020 will likely to be 120,000,000."
- From the survey obtained on 15 randomly selected towns of Eth. it is estimated that 0.1% of the whole urban dwellers are victims of AIDS virus.

Exercise: Descriptive Statistics or Inferential Statistics

1. The manager of quality control declares the out come of its survey as "the average life span of the imported light bulbs is 3000 hrs.
2. Of all the patients taking the drug at a local health center 80% of them suffer from side effect developed.
3. The average score of all the students taking the exam is found to be 72.
4. The national statistical bureau of Eth. declares the out come of its survey for the last 30 years as "the average annual growth of the people of Ethiopia is 2.8%.
5. The national statistical bureau of Eth. declares the out come of its survey for the last 30 years as "the population of Ethiopia in the year 2015 will likely to be 100,000,000.
6. Based on the survey made for the last 10 years 30,000 tourists are expected to visit Ethiopia.
7. Based on the survey made for the last 20 years the maximum number of tourists visited Eth. were in the year 1993.
8. The Ethiopian tourism commission has announced that (as observed for the last 20 years) the average number of tourists arrived Ethiopia per year is 3000.
9. The maximum difference of the salaries of the workers of the company until the end of last year was birr 5000.

Main terms in statistics:

Data: Certainly known facts from which conclusions may be drawn.

Statistical data: Raw material for a statistical investigation which are obtained when ever **measurements or observations are made.**

i.Quantitative data: data of a certain group of individuals which is expressed numerically.

Example: Heights, Weights, Ages and, etc of a certain group of individuals.

ii.Qualitative data: data of a certain group of individuals that is not expressed numerically.

Example: Colors, Languages, Nationalities, Religions, health, poverty etc of a certain group of individuals.

Primary data and **Secondary data**

Variable: A variable is a factor or characteristic that can take on different possible values or outcomes.

A variable can be qualitative or quantitative (numeric).

Example: Income, height, weight, sex, age, etc of a certain group of individuals are examples of variables.

Population: A complete set of observation (data) of the entire group of individuals under consideration .

A population can be finite or infinite.

Example: The number of students in this class, the population in Addis Ababa etc.

Sample: A set of data drawn from population containing a part which can reasonably serve as a basis for valid generalization about the population.

A sample is a portion of a population selected for further analysis.

Sample size: The number of items under investigation in a sample.

Survey (experiment): it is a process of obtaining the desired data. Two types of survey:

1. **Census Survey:** A way of obtaining data referring the entire population including a total coverage of the population.
2. **Sample Survey:** A way of obtaining data referring a portion of the entire population consisting only a partial coverage of the population.

STEPS/STAGES IN STATISTICAL INVESTIGATION

1. Collection of Data:

Data collection is the process of gathering information or data about the variable of interest. Data are inputs for Statistical investigation. Data may be obtained either from primary source or secondary source.

2. Organization of Data

Organization of data includes three major steps.

1. ***Editing:*** checking and omitting inconsistencies, irrelevancies.
2. ***Classification :*** task of grouping the collected and edited data .
3. ***Tabulation:*** put the classified data in the form of table.

3. Presentation of Data

The purpose of presentation in the statistical analysis is to display what is contained in the data in the form of Charts, Pictures, Diagrams and Graphs for an easy and better understanding of the data.

4. Analyzing of Data

- In a statistical investigation, the process of analyzing data includes finding the various statistical constants from the collected mass of data such as measures of central tendencies (averages) , measures of dispersions and soon.
- It merely involves mathematical operations: different measures of central tendencies (averages), measures of variations, regression analysis etc. In its extreme case, analysis requires the knowledge of advanced mathematics.

5. Interpretation of Data

- involve interpreting the statistical constants computed in analyzing data for the formation of valid conclusions and inferences.
- It is the most difficult and skill requiring stage.
- It is at this stage that Statistics seems to be very much viable to be misused.
- Correct interpretation of results will lead to a valid conclusion of the study and hence can aid in taking correct decisions.
- Improper (incorrect) interpretation may lead to wrong conclusions and makes the whole objective of the study useless.

THE ENGINEERING METHOD AND STATISTICAL THINKING

- An engineer is someone who solves problems of interest to society by the efficient application of scientific principles.

- Engineers accomplish this by either refining an existing product or process or by designing a new product or process that meets customers' expectations and needs.

The steps in the engineering method are as follows:

1. Develop a clear and concise description of the problem.
2. Identify, the important factors that affect this problem or that may play a role in its solution.
3. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied. State any limitations or assumptions of the model.
4. Conduct appropriate experiments and collect data to test or validate the tentative model or conclusions made in steps 2 and 3.

5. Refine the model on the basis of the observed data.
6. Manipulate the model to assist in developing a solution to the problem.
7. Conduct an appropriate experiment to confirm that the proposed solution to the problem is both effective and efficient.
8. Draw conclusions or make recommendations based on the problem solution.

Cont'd

- The engineering method features a strong interplay between the problem, the factors that may influence its solution, a model of the phenomenon, and experimentation to verify the adequacy of the model and the proposed solution to the problem.
- Specifically, statistical techniques can be a powerful aid in designing new products and systems, improving existing designs, and designing, developing, and improving production processes.

Cont'd

- Therefore, Engineers must know how to efficiently plan experiments, collect data, analyze and interpret the data, and understand how the observed data are related to the model that have been proposed for the problem under study.

1.2. Data collection and graphical representation of data

Classification of Data based on source:

1. **Primary:** data collected for the purpose of specific study.

   It can be obtained by:

   - ¬Direct personal observation
   - ¬Direct or indirect oral interviews
   - ¬Administrating questionnaires

2.Secondary: refers to data collected earlier for some purpose other than the analysis currently being undertaken.

It can be obtained from:

- ¬ External Secondary data Sources(for eg. gov't and non gov't publications)
- ¬ Internal Secondary data Sources: the data generated within the organization in the process of routine business activities

Qualitative data are nonnumeric. **Quantitative** data are numeric.

Method of data presentation

¬The purpose of organizing data is to see quickly some of the characteristics of the data that have been collected.

¬**Raw data** is collected numerical data which has not been arranged in order of magnitude.

¬**An array** is an arranged numerical data in order of magnitude.

Method of data presentation

- Mechanism for reducing and summarizing data are:

1. Tabular method.
2. Graphical method
3. Diagrammatic method

1. Tabular presentation of data:

¬The collected raw data should be put into an ordered array in either ascending or descending order so that it can be organized in to a Frequency Distribution (FD)

¬Numerical data arranged in order of magnitude along with the corresponding frequency is called frequency distribution (FD).

¬FD is **of two kinds namely ungrouped /and grouped frequency distribution**.

A. Ungrouped (Discrete) Frequency Distribution

- It is a tabular arrangement of numerical data in order of magnitude showing the **distinct values** with the corresponding frequencies.

Example:

Suppose the following are test score of 16 students in a class, write un grouped frequency distribution.

"14, 17, 10, 19, 14, 10, 14, 8, 10, 17, 19, 8, 10, 14, 17, 14"

Sol: the ungrouped frequency distribution:

Array: 8,8,10,10,10,10,14,14,14,14,14, 17,17,17,19,19.

Then the ungrouped frequency distribution is then grouped:-

Test score	8	10	14	17	19
Frequency	2	4	5	3	2

- The difference between the highest and the lowest value in a given set of observation is called **the range (R)**
  R= L- S, R= 19-8 = 11

B. Grouped (continuous) Frequency Distribution (GFD)

- It is a tabular arrangement of data in order of magnitude by classes together with the corresponding class frequencies.
- In order to estimate the number of classes, the ff formula is used:

  Number of classes=1+3.322(log N) where N is the Number of observation.

$$\text{The Class size (class width)} = \frac{\text{Range}}{1+3.322(\log N)} \quad \text{(round up)}$$

Example:

Grouped/Continuous frequency distribution where several numbers are grouped into one class.

e.g.

Student age	Frequency
18-25	5
26-32	15
33-39	10

Components of grouped frequency distribution

1. **Lower class limit:**

   is the smallest number that can actually belong to the respective classes.

1. **Upper class limit:**

   is the largest number that can actually belong to the respective classes.

1. **Class boundaries:**

   are numbers used to separate adjoining classes which should not coincide with the actual observations.

1. **Class mark:**

   is the midpoint of the class.

5. Class width/ Class intervals

is the difference between two consecutive lower class limits or the two consecutive upper class limits. (OR)

can be obtained by taking the difference of two adjoining class marks or two adjoining lower class boundaries.

Class width = Range/Number of class desired.

Where: Number of classes=1+3.322(log N) where N is the Number of observation.

6. Unit of measure

is the smallest possible positive difference between any two measurements in the given data set that shows the degree of precision.

- **Class boundaries:**

  can be obtained by taking the averages of the upper class limit of one class and the lower class limit of the next class.

- **Lower class boundaries:**

  can be obtained by subtracting half a unit of measure from the lower class limits.

- **Upper class boundaries:**

  can be obtained by adding half the unit of measure to the upper class limits.

Example1 :

Suppose the table below is the frequency distribution of test score of 50 students.

Then the frequency table has 6 classes (class intervals).

Test score	Frequency
11-15	7
16-20	8
21-25	10
26-30	12
31-35	9
36-40	4

What is the Unit of Measure, LCLs, UCLs, LCBs, UCBs, CW, and CM

- The unit of measure is 1
- The lower class limits are:-11, 16, 21, 26, 31, 36
- The upper class limits are:- 15, 20, 25, 30, 35, 40
- The class marks are:- 13((11+15)/2), 18, 23, 28, 33, 38
- The lower class boundaries are:- 10.5(11-0.5), 15.5, …., 35,5
- The upper class boundaries are:- 15.5(15+0.5), ….35.5, 40.5
- Class width (size) is 5.

Rules to construct Grouped Frequency Distribution (GFD):

i. Find the **unit of measure** of the given data;
ii. Find the **range**;
iii. Determine the **number of classes** required;
iv. Find **class width (size)**;
v. Determine a **lowest class limit** and then find the **successive lower and upper class limits** forming non over lapping intervals such that each observation falls into exactly one of the class intervals;
vi. Find the **number of observations** falling into each class intervals that is taken as **the frequency of the class (class interval)** which is best done using a tally.

Exercise:

Construct a GFD of the following aptitude test scores of 40 applicants for accountancy positions in a company with

a. 6 classes b. 8 classes

96	89	58	61	46	59	75	54
41	56	77	49	58	60	63	82
66	64	69	67	62	55	67	70
78	65	52	76	69	86	44	76
57	68	64	52	53	74	68	39

Types of Grouped Frequency Distribution

1. Relative frequency distribution (RFD)
2. Cumulative Frequency Distribution (CFD)
3. Relative Cumulative Frequency Distribution (RCFD)

Types of Grouped Frequency Distribution

1. Relative frequency distribution (RFD):

- A table presenting the ratio of the **frequency of each class to the total frequency of all the classes**.
- Relative frequency generally expressed **as a percentage**, used to show the percent of the total number of observation in each class.

For example

Test score	F	RFD	PFD
37.5-47.5	4	4/40=0.1	10%
47.5-57.5	8	8/40=0.2	20%
57.5-67.5	13	13/40=0.325	32.5%
67.5-77.5	10	10/40=0.25	25%
77.5-87.5	3	3/40=0.075	7.5%
87.5-97.5	2	2/40=0.05	5%

2. Cumulative Frequency Distribution (CFD):

- It is applicable when we want to know how many observations lie **below or above a certain value/class boundary.**

- **CFD** is of two types, **LCFD and MCFD:**
  - **Less than Cumulative Frequency Distribution (LCFD):** shows the collection of cases lying **below the upper class boundaries of each class.**
  - **More than Cumulative Frequency Distribution (MCFD):** shows the collection of cases lying **above the lower class boundaries of each class.**

Test score	CF
Less than 37.5	0
Less than47.5	4
Less than 57.5	12
Less than 67.5	25
Less than 77.5	35
Less than 87.5	38
Less than 97.5	40

Test score	CF
more than 37.5	40
more than 47.5	36
more than 57.5	28
more than 67.5	15
more than 77.5	5
more than 87.5	2
more than 97.5	0

3. Relative Cumulative Frequency Distribution (RCFD)

It is used to determine the ratio or the percentage of observations that lie below or above a certain value/class boundary, to the total frequency of all the classes. These are of two types: The LRCFD and MRCFD.

- **Less than Relative Cumulative Frequency Distribution (LRCFD):** A table presenting the ratio of the cumulative frequency **less than upper class boundary of each class to the total frequency of all the classes**
- **More than Relative Cumulative Frequency Distribution (MRCFD):** A table presenting the ratio of the cumulative frequency **more than lower class boundary of each class to the total frequency of all the classes.**

LRCFD

Test score	LCF	LRCF	LPCF
Less than 37.5	0	0/40=0	0%
Less than 47.5	4	4/40=0.1	10%
Less than 57.5	12	12/40=0.3	30%
Less than 67.5	25	25/40=0.625	62.5%
Less than 77.5	35	35/40=0.875	87.5%
Less than 87.5	38	38/40=0.95	95%
Less than 97.5	40	40/40=1	100%

MRCFD

Test score	MCF	MRCF	MPCF
More than 37.5	40	40/40=1	100%
More than 47.5	36	36/40=0.9	90%
More than 57.5	28	28/40=0.7	70%
More than 67.5	15	15/40=0.375	37.5%
More than 77.5	5	5/40=0.125	12.5%
More than 87.5	2	2/40=0.05	5%
More than 97.5	0	0/40=0	0%

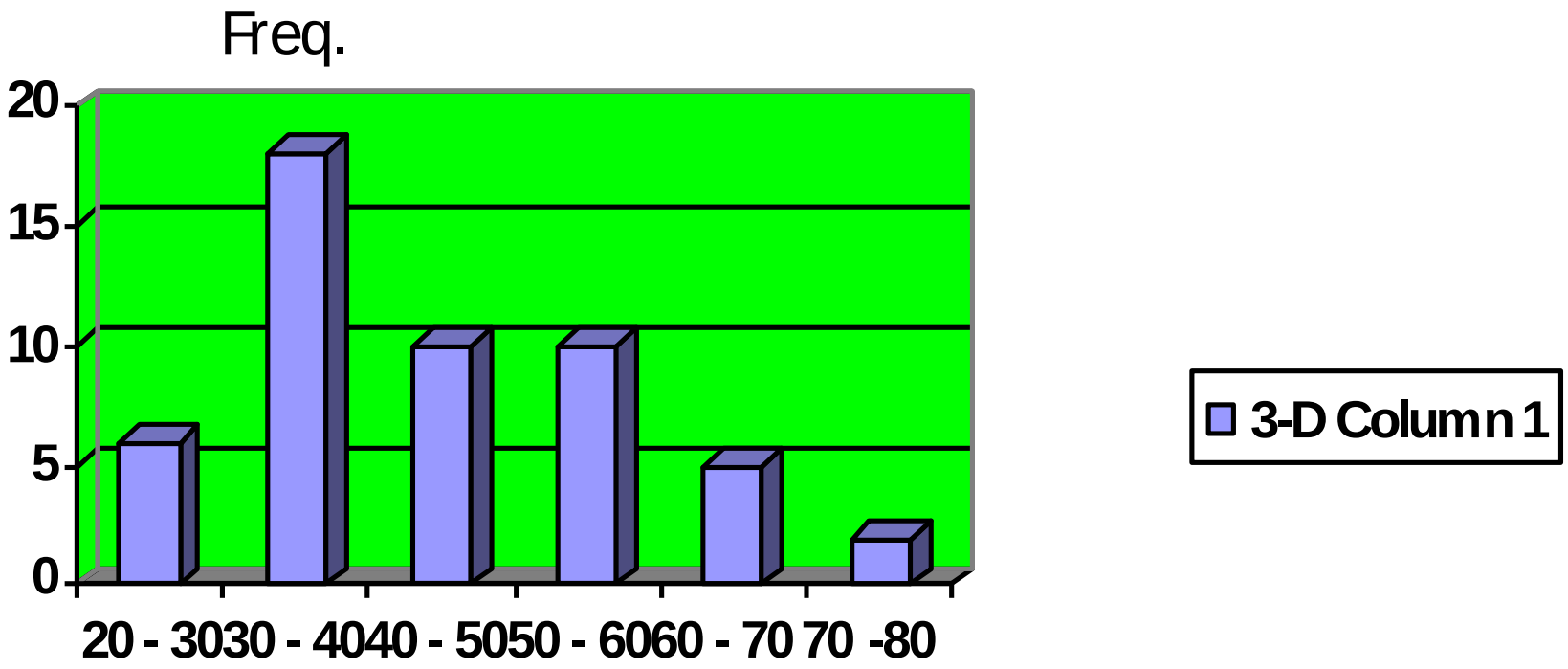
Graphic Methods of Data presentation

1. Histogram
2. Frequency Polygon (Line graph)
3. Cumulative frequency curve (o-give)

1. Histogram:

- A graphical presentation of grouped frequency distribution consisting of a series of adjacent rectangles whose bases are the class intervals specified in terms of class boundaries (equal to the class width of the corresponding classes) shown on the x-axis and whose heights are proportional to the corresponding class frequencies shown on the y-axis.

Histogram: E.g.



Steps to draw Histogram

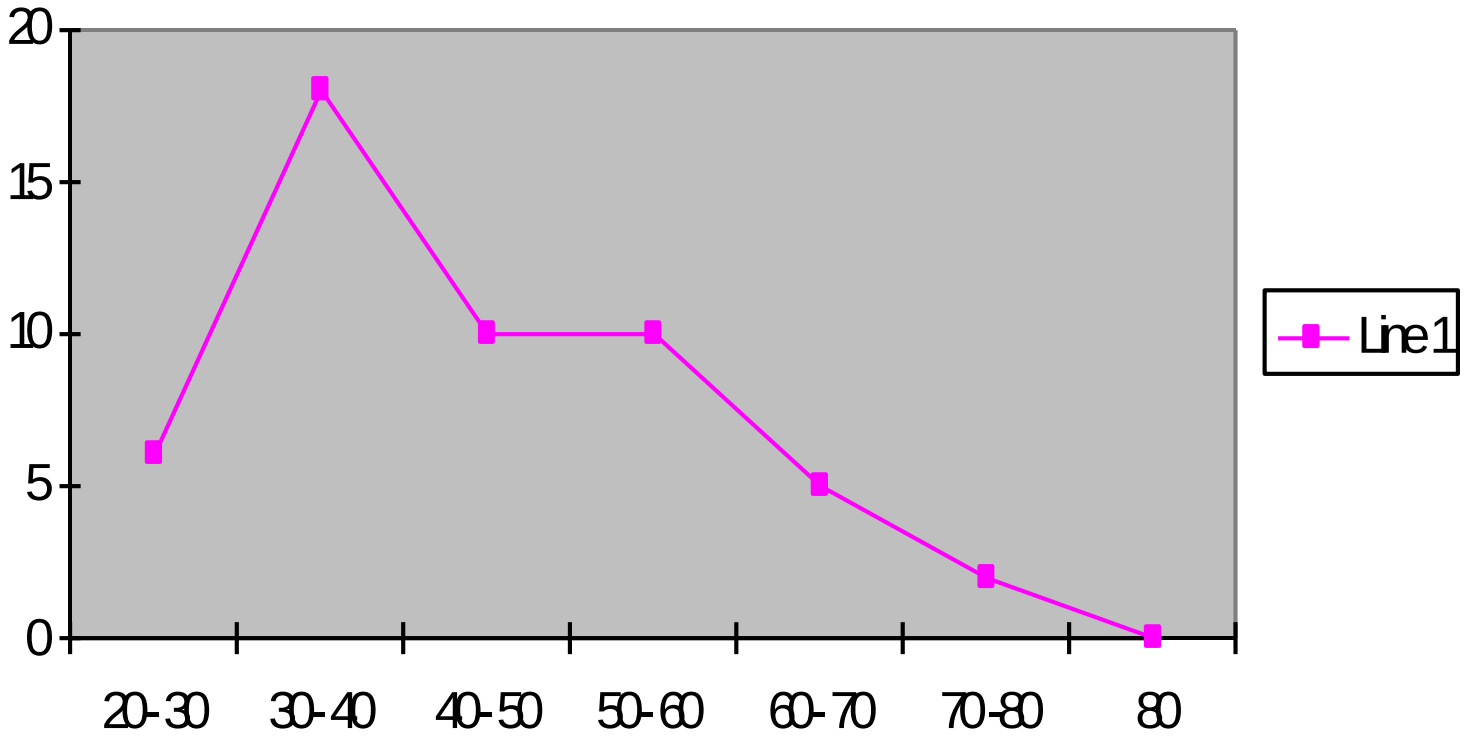
i. Mark the class boundaries on the horizontal axis (x- axis) and the class frequencies along the vertical axis (y- axis) according to a suitable scale.

ii. With each interval as a base draw a rectangle whose height equals the frequency of the corresponding class interval. It describes the shape of the data.

2. Frequency Polygon:

It is a line graph of grouped frequency distribution in which the class frequency is plotted against class mark that are subsequently connected by a series of line segments to form line graph including classes with zero frequencies at both ends of the distribution to form a polygon.

Frequency Polygon:



Steps to draw Frequency polygon

i. Mark the class mid points on the x-axis and the frequency on the y-axis.

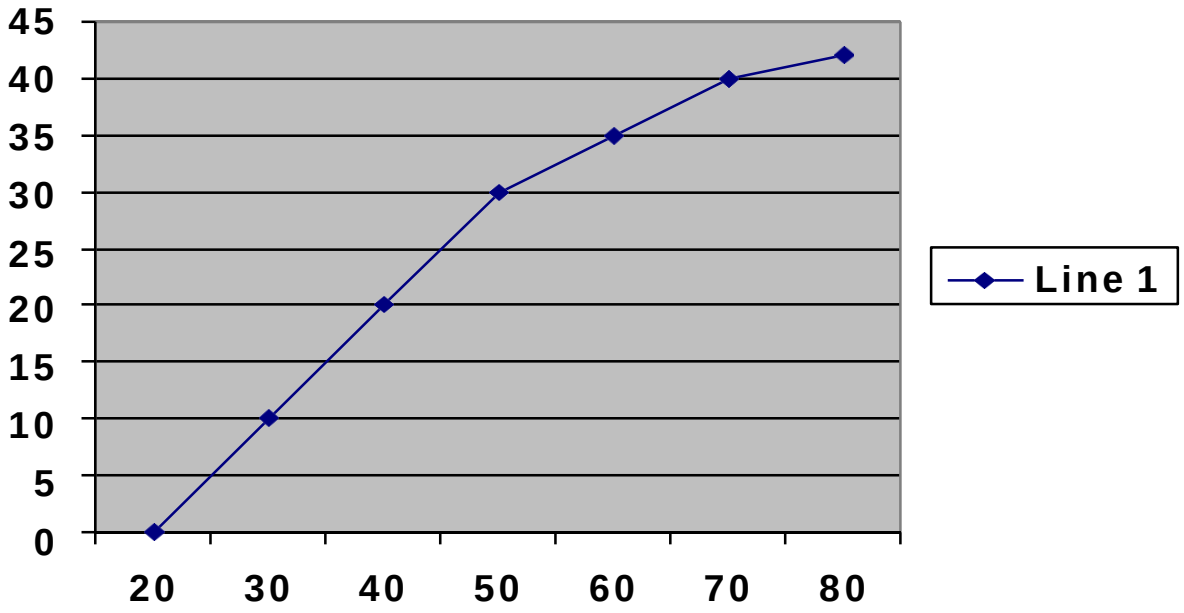
ii. Mark dots which correspond to the frequency of the marked class mid points.

iii. Join each successive dot by a series of line segments to form line graph, including classes with zero frequencies at both ends of the distribution to form a polygon.

3. O-GIVE curve (Cumulative Frequency Curve / percentage Cumulative Frequency Curve)

- It is a line graph presenting the cumulative frequency distribution.
- O-gives are of two types: The **Less than O-give** and The **More than O-give**.
- **The Less than O-give** shows the cumulative frequency less than the upper class boundary of each class; and
- **The More than O-give** shows the cumulative frequency more than the lower class boundary of each class.

Ogive: E.g.



Steps to draw O-gives

i. Mark class boundaries on the x-axis and mark non overlapping intervals of equal length on the y-axis to represent the cumulative frequencies.

ii. For each class boundaries marked on the x-axis, plot a point with height equal to the corresponding cumulative frequencies.

iii. Connect the marked points by a series of line segments where the less than O-give is done by plotting the less than cumulative frequency against the upper class boundaries

Diagrammatic Presentation Of Data

- Bar charts
- Pie chart
- Pictograph and
- Pareto diagram

Outline

3. **Measures of Central Tendency**
4. **Measures of Dispersion**
5. **Measure of skewness and kurtosis**

1.3 Measures of Central Tendency

- Central value refers to the location of the centre of the distribution of data.
- Measure of central value:
  - **The mean**
  - **The mode**
  - **The median**
  - **Percentile/Quantiles and**
  - **Midrange**

The Mean:

- It is the most commonly used measures of central value.
- Types of Mean:
  1. Arithmetic Mean
  2. Geometric Mean
  3. Harmonic Mean
  4. Quadratic Mean
  5. Trimmed Mean
  6. Weighted Mean
  7. Combination mean

1. Arithmetic Mean (simply Mean)

- The mean is defined as the arithmetic average of all the values.
- It is represented by **$\bar{x}$ read as x-bar for a sample and by μ for a population**

$$\bar{x} = \frac{\sum_{i=1}^{fc} m_i}{fc} \qquad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Advantages

- It is the most commonly used measure of location or central tendency for continuous variables.
- The arithmetic mean uses all observations in the data set.
- All observations are given equal weight.

Disadvantage

- The mean is affected by **extreme values** that may not be representative of the sample.

2. Geometric mean

- It is the ***n*th root** of the product of the data elements.
- It is used in business to find average rates of growth.

Geometric mean = $\sqrt[n]{\prod x_i}$ all n >=2.

- **Example:** suppose you have an IRA (Individual Retirement Account) which earned annual interest rates of 5%, 10%, and 25%. **Solution:** The proper average would be the geometric mean or the cube root of (1.05 • 1.10 • 1.25) or about 1.13 meaning 13%.
- Note that the data elements must be positive. Negative growth is represented by positive values less than 1. Thus, if one of the accounts lost 5%, the proper multiplier would be 0.95.

3. Harmonic mean

- It is used to calculate average rates.
- It is found by dividing the number of data elements by the sum of the reciprocals of each data element.

Harmonic mean = $\frac{n}{\sum x_i^{-1}}$

- **Example:** Suppose a boy rode a bicycle three miles. Due to the topography, for the first mile he rode 2 mph; for the second mile 3 mph; for the final mile the average speed was 4 mph. What was the average speed for the three miles?
  Solution: The arithmetic mean of (2+3+4)/3 = 3.0 is incorrect. This would imply it took 1 hour. Breaking it down into the separate components, it takes 30 minutes (1st) + 20 minutes (2nd) + 15 minutes (3rd) to walk (each mile) or 65 minutes total. His actual speed was thus 3/1.083 or 2.77 mph.
- Another way to show our work would be:

$$\frac{3 \text{ miles}}{1/2+1/3+1/4} = \frac{3}{13/12} = \frac{36}{13} = 2.77\text{mph}$$

4. Quadratic mean

- It is another name for **Root Mean Square or RMS.**
- **RMS** is typically used for data whose arithmetic mean is zero.

$$\mathrm{RMS} = \sqrt{\frac{\sum x_i^2}{n}}$$

Example

- Suppose measurements of 120, -150, and 75 volts were obtained.

Solution: The corresponding quadratic mean is $\sqrt{((120^2 + (-150)^2 + 75^2)/3)}$ or 119 volts RMS.

- The quadratic mean gives a physical measure of the average distance from zero.

5. Trimmed mean

- It is usually refers to **the arithmetic mean without the top 10% and bottom 10% of the ordered scores;**
- Removes extreme scores on both the high and low ends of the data.

- A **truncated mean** or **trimmed mean**

6. Weighted mean

- It is the average of differently weighted scores;
- It takes into account some measure of weight attached to different scores.

$$\text{Weighted mean} = \frac{\sum (w_i \cdot x_i)}{\sum w_i}$$

The Mode

- The mode is the most frequent or most typical value.
- The mode will not always be the central value; in fact it may sometimes be an extreme value. Also, a sample may have more than one mode.

Bimodal or **Multimodal**

Example

‘23, 22, 12, 14, 22, 18, 20, 22, 18, 18’

The mode is 18 and 22 - bimodal

Advantages

- Requires no calculations.
- Represents the value that occurs most often.

Disadvantage

- The mode for continuous measurements is dependent on the grouping of the intervals.
- We may not have mode

The Median

- The median is the middle value of a group of an odd number of observations when the data is arranged in increasing or decreasing order magnitude.
- If the number of values is even, the median is the average of the two middle values.

Example

‘23, 22, 12, 14, 22, 18, 20, 22, 18, 18’

Array: 12, 14, 18, 18, 18, 20, 22, 22,22,23

- The median (18+20)/2 = 19

Advantages

- The median always exists and is unique.
- The median is not affected by extreme values.

Disadvantages

- The values must be sorted in order of magnitude.
- The median uses only one (or two) observations.

Percentiles / Quartiles

- Percentiles are values that divide a distribution into two groups where the **Pth** percentile is larger than **P%** of the values.
- Some specific percentiles have special names:

First Quartile : Q_1 = the 25 percentile

Median : Q_2 = the 50 percentile

Interpretation

- Q_1= a. This means that 25% of the data values are smaller than a
- Q_2= b. This means that 50 %of the data has values smaller than b.
- Q_3= c. This means that 75 %of the data has values smaller than c.

Midrange

- The midrange is the average of largest and smallest observation.

Midrange = (Largest +Smallest)/2

- The percentile estimate (P25 + P75)/2 is sometimes used when there are a large number of observations.

1.4. Measure of Dispersion (of Variability)

- A measure of dispersion indicates how the observations are spread about the central value.
- Measures of dispersion are:
  - **The range**
  - **The variance**
  - **The standard deviation and**
  - **The coefficient of variation**

The range

- The range is the difference between the largest and the smallest value in the sample.
- The range is the easiest of all measures of dispersion to calculate.

R= Maximum Value - Minimum Value

Advantage

- The range is easily understood and gives a quick estimate of dispersion.
- The range is easy to calculate

Disadvantage

- The range is inefficient because it only uses the extreme value and ignores all other available data. The larger the sample size, the more inefficient the range becomes.

The Variance

- The variance is the mean square deviation of the observations from the mean. To calculate the variance, the following equation is used:

$$s^2 = \sigma 2 = \frac{\sum fi * (m_i - \vec{x})^2}{n - 1}$$

Advantages

- The variance is an efficient estimator
- Variances can be added and averaged

Disadvantage

- The calculation of the variance can be tedious without the aid of a calculator or computer

The Standard Deviation

- The square root of the variance is known as the standard deviation. The symbol for the standard deviation is s.

$$s = \sqrt{s^2}$$

Advantages

- The standard deviation is in the same dimension as the observed values.
- The standard deviation is an efficient estimator.

Disadvantages

- The calculations can be tedious without the aid of a good calculator

The coefficient of variation

- The coefficient of variation is a measure of relative dispersion, that is, a measure which expresses the magnitude of the variation to the size of the quantity that is being measured and is expressed as a percent.

$$CV = (s / \bar{x}) * 100$$

Advantages

- The coefficient of variation can be used for comparing the variation in different populations of data that are measured in two different units. (because the CV is unitless)

Disadvantages

- The coefficient of variation fails to be useful when x is close to zero.
- The coefficient of variation is often misunderstood and misused.

The Interquartile range

- It is the difference between the 25th and the 75th quartiles.

$$\text{Interquartile range} = Q3 - Q1$$

Measure of skewness and kurtosis

Skewness (mean deviation)

- **Skewness** is a measure of the tendency of the deviations to be larger in one direction than in the other.
- **Skewness** is the degree of asymmetry or departure from symmetry of a distribution.
- If the frequency curve of a distribution has a longer tail to the right of the central value than to the left, the distribution is said to be skewed to the right or to have positive skewness.
- If the reverse is true, it is said that the distribution is skewed to the left or has negative skewness.

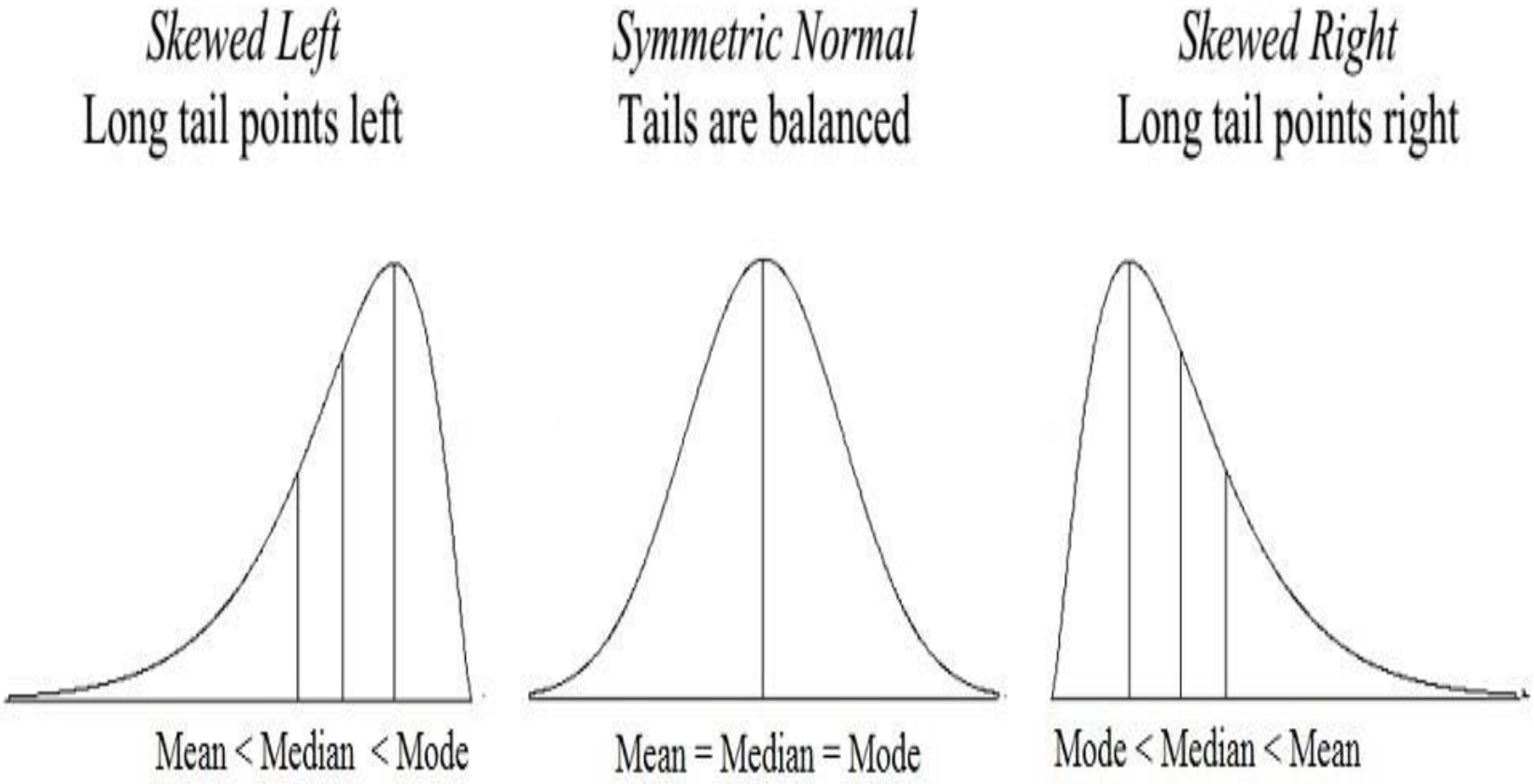


Figure 1. Sketches showing general position of mean, median, and mode in a population.

- Population skewness is defined as

$$\frac{E(x-\mu)^3}{\sigma^3}$$

where Estands for **expected value**

- A bell-shaped distribution which has no skewness, i.e., mean = median = mode is called **a normal distribution.**
- f mean > Median > mode, **the distribution is positively skewed distribution or it is said to be skewed to the right.**
- If mean < Median < mode, **the distribution is negatively skewed distribution or it is said to be skewed to the left.**

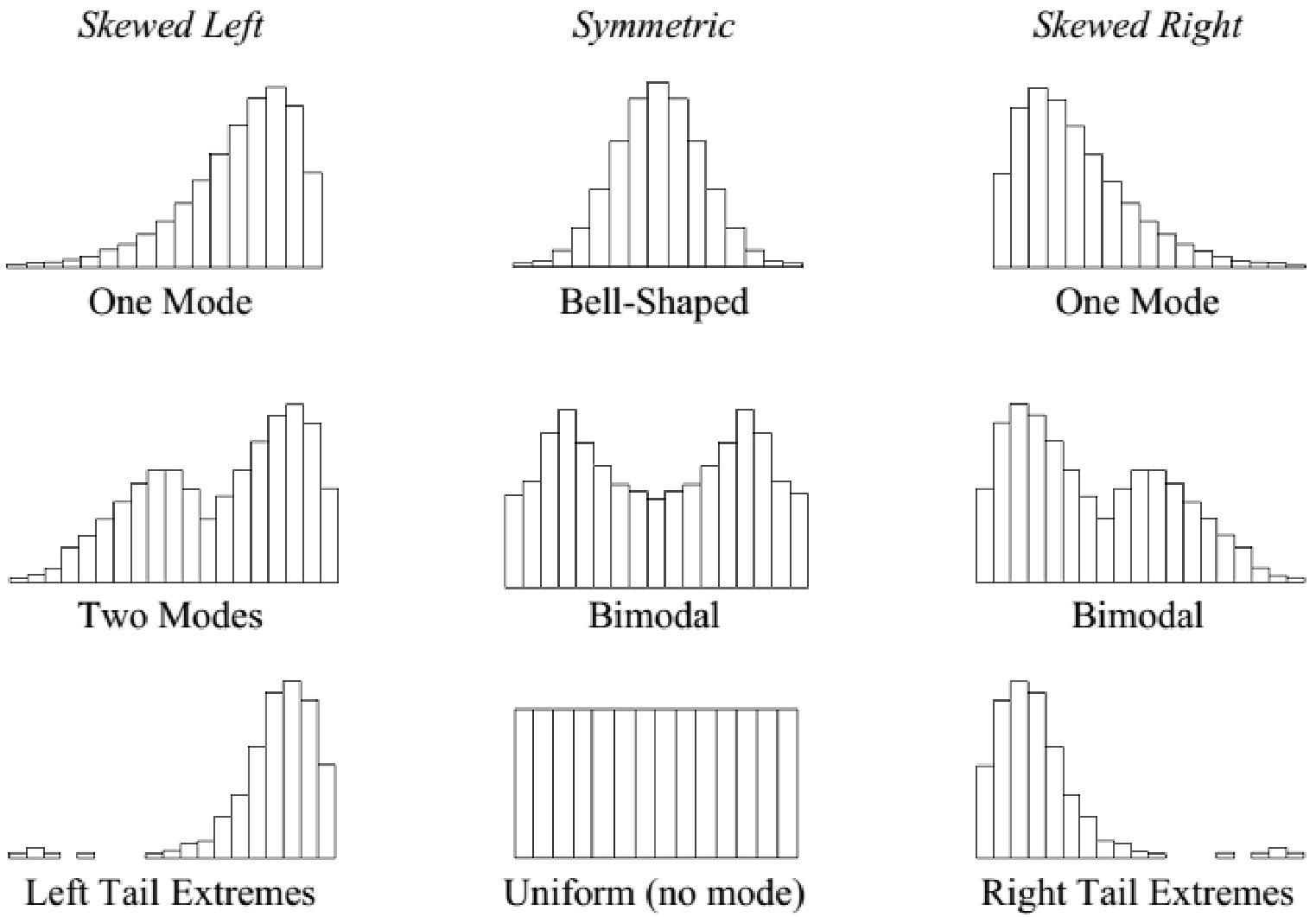


Figure 2. Illustrative prototype histograms.

Measure of kurtosis

- Kurtosis characterises the relative peakedness or flatness of a distribution compared with the normal distribution.
- **Positive kurtosis** indicates a **relatively peaked distribution.**
- Negative kurtosis indicates a **relatively flat distribution.**
- The population kurtosis is usually defined as:

$$\frac{E(x-\mu)^4}{N\sigma^4}-3$$

Data come in many flavors …

Type of data	Definition	Example
Nominal	Categories	Your previous degree
Ordinal	Can be ranked *f* ordered but not measured	Business school rankings
Interval scale	Intervals are meaningful but not ratios	Temperature in Fahrenheit or Celsius
Ratio scale	Ratios are meaningful	Sales of a new product

Source of data	Definition	Example
Observational	Analyst does not control data generating process	Stock returns on BSE
Experimental	Analyst has good control over data generation	Drug efficacy in clinical trials

Random Variable

- A random variable describes the probabilities for an uncertain future numerical outcome of a random process
- It is a variable because it can take one of several possible values
- It is random because there is some chance associated with each possible value
- Examples?

Probability Distribution: Discrete and Continuous

- Probability
  - Long run average of a random event occuring
  - Different from subjective beliefs
- A probability distribution is a rule that identifies possible outcomes of a random variable and assigns a probability to each
- A discrete distribution has a finite number of values
  - e.g. face value of a card, work experience of students rounded off to nearest month
- A continuous distribution has all possible values in some range
  - e.g. sales per month, height of students in this class
- Continuous distributions are nicer to deal with and are good approximations when there are a large number of possible values

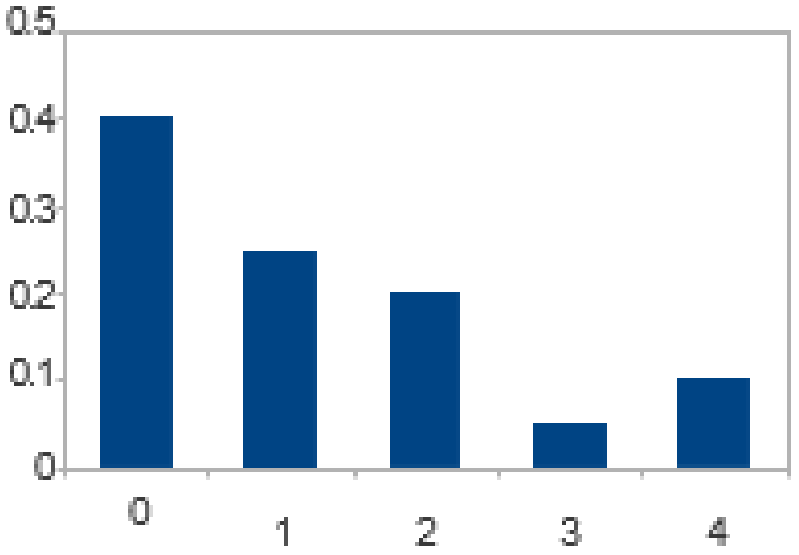
Discrete Probability Distribution: An example

- Suppose you randomly picked a card from the card deck. What is the probability that this card will be
  - Bigger than 10?
  - Equal to or bigger than 10?
  - Smaller than 3?
  - Greater than 4 and less than 8 (or, between 4 and 8)?

Another Example

- The daily sales of large flat panel TVs at a store (X)

x	P(X=x)
0	0.40
1	0.25
2	0.20
3	0.05
4	0.10



- What is the probability of a sale?
- What is the probability of selling at least three TVs?

Expected Value or Mean

- The expected value or mean (μ) of a random variable is the weighted average of its values
  - The probabilities serve as weights
  - $E(X) = \sum_i x_i P(X = x_i)$
- What is the mean number of TVs sold per day?
- What does this number imply?

Variance and Standard Deviation

- Both measures of variation or uncertainty in the random variable

- Variance (σ^2): The weighted average of the squared deviations from the mean
  - Probabilities serve as weights
  - $\sigma^2(X) = \sum_i [x_i - \mu]^2 P(X = x_i) = E[X - \mu]^2$
  - Units are square of the units of the variable

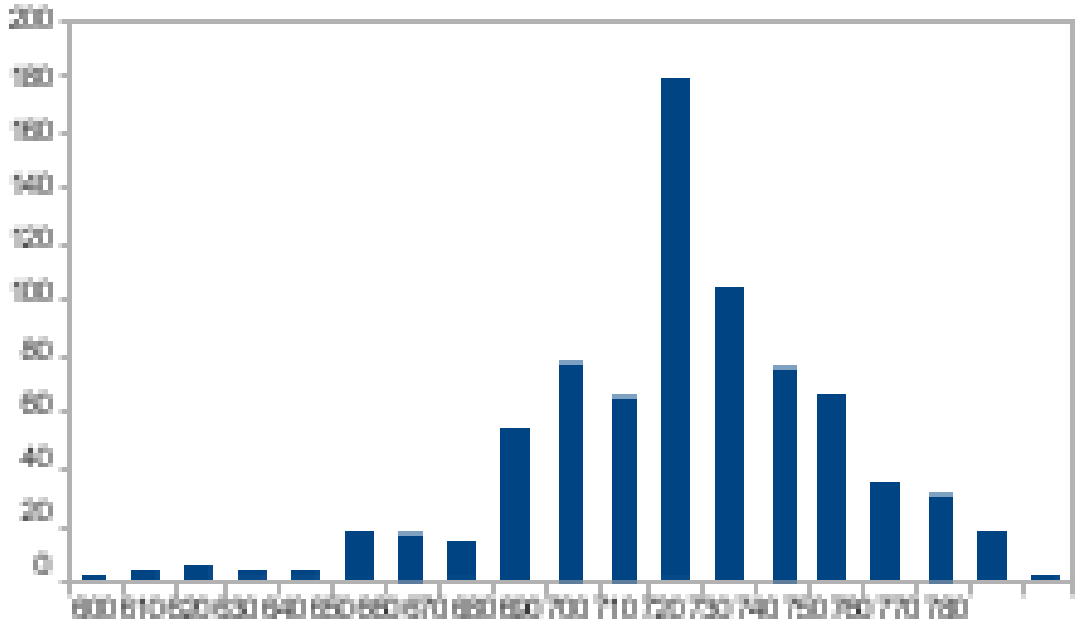
- Standard deviation (σ): Square root of variance
  - Has same units as the variable

GMAT Scores of an MBA Class

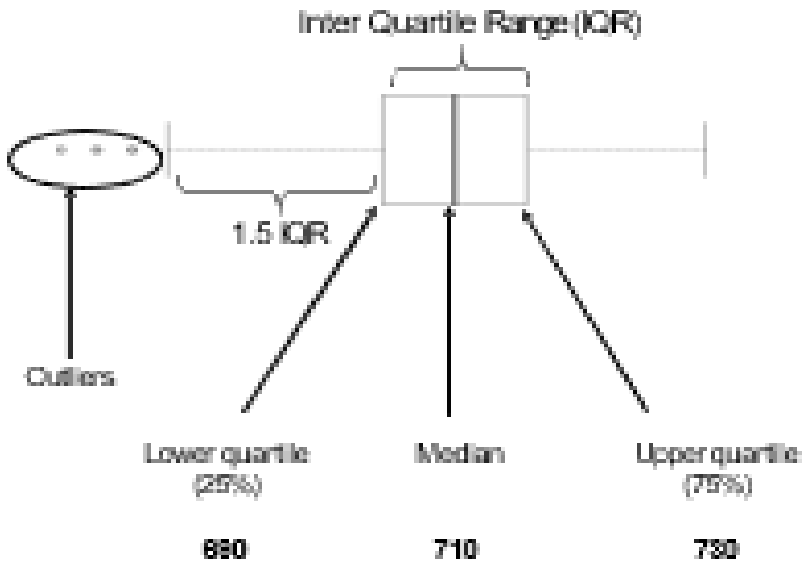
610	730	590	610	.	.	.	680	630
640	680	540	650	.	.	.	610	540
690	610	520	640	.	.	.	720	680
610	650	660	580	.	.	.	600	730
710	600	760	690	.	.	.	500	720
610	650	660	710	.	.	.	480	600
630	610	680	780	.	.	.	700	690
530	550	730	690	.	.	.	670	540
630	720	610	710	.	.	.	600	600
690	680	730	540	.	.	.	560	770

Data File: mba.csv

Pictorial summary of data: A bar chart



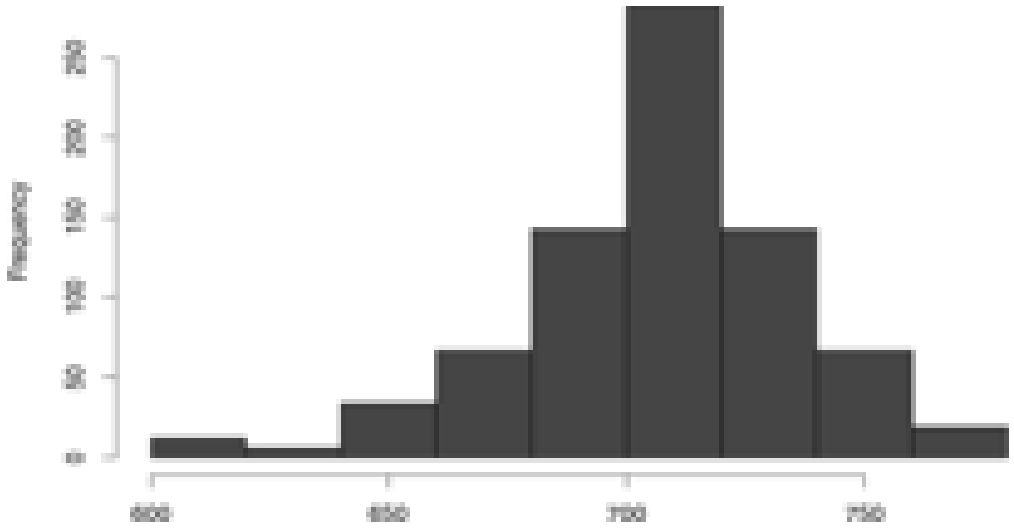
Boxplot



A boxplot displays the prominent quartiles of the data along with outliers

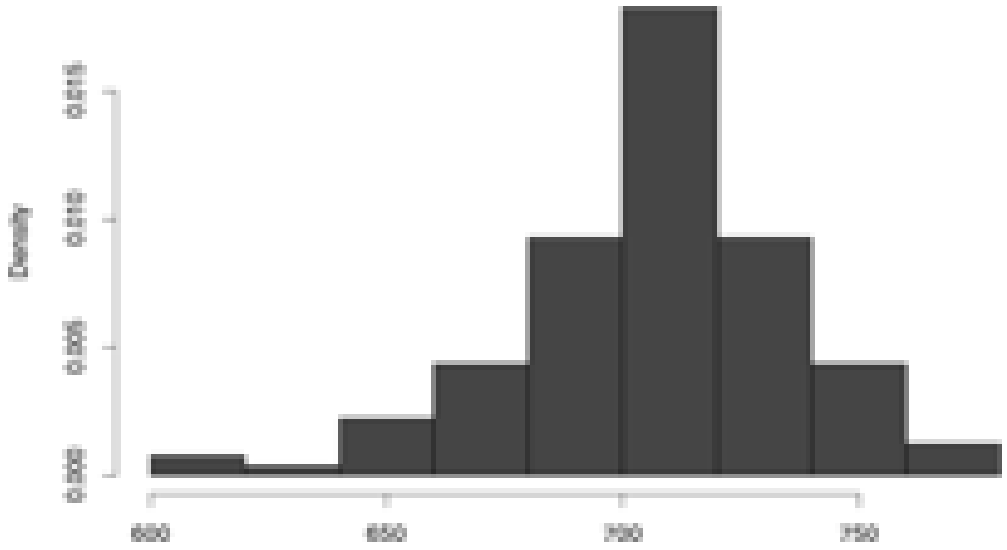
Histogram

- A histogram represents the frequency distribution, i.e., how many observations take the value within a certain interval



Probability distribution

- A frequency distribution can be converted into a relative frequency histogram
- This provides one representation of the probability distribution for the GMAT score of a randomly picked student



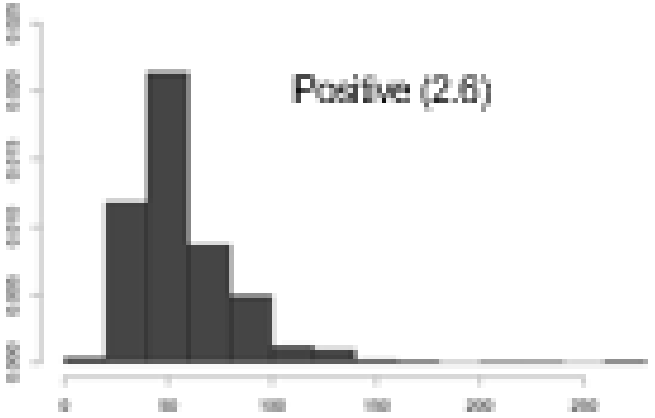
Skewness and Kurtosis

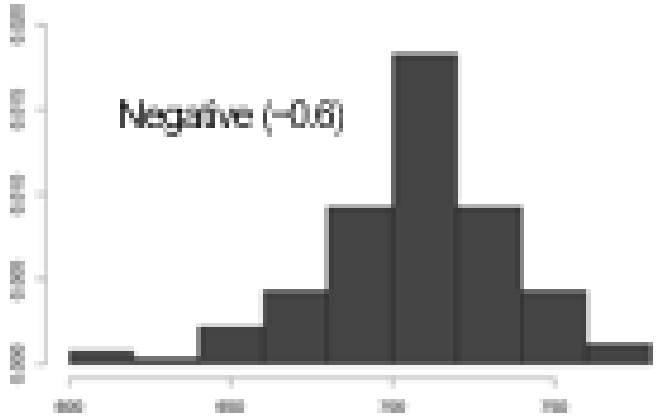
- Two additional summary measures of a random variable f probability distribution
- Can be interpreted as the third and fourth moments just as mean and variance are the first and second moments

Skewness	(Excess) Kurtosis
• A measure of "asymmetry" in the distribution	• A measure of the "peakedness" of the distribution (relative to normal)
• Mathematically, it is given by – $E[(X-\mu)/\sigma]^3$	• Mathematically, it is given by – $E[(X-\mu)/\sigma]^4-3$
• Negative skewness implies mass of the distribution is concentrated on the right	• For symmetric distributions, negative kurtosis implies wider peak and thinner tails

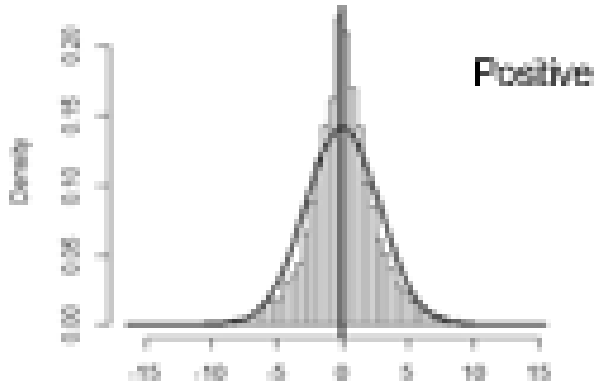
Skewness and Kurtosis (contd.)

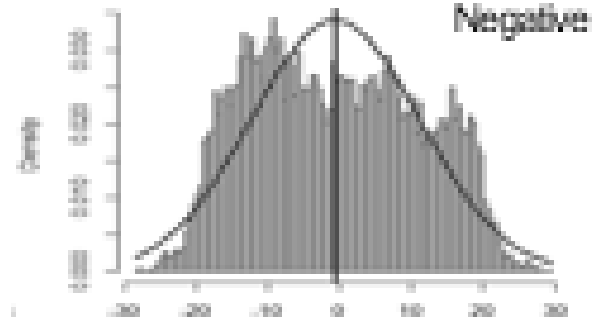
Skewness





(Excess) Kurtosis





Sum of Random Variables

- Let X_1 and X_2 be two random variables with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. Suppose $Y = aX_1 + bX_2$.
  - What is the mean of Y?
  - What is the standard deviation of Y?

- Independent: When the value taken by one random variable does not affect the value taken by the other random variable
  - e.g. Roll of two dice
- Dependent: When the value of one random variable gives us more information about the other random variable
  - e.g. Height and weight of students

Example

- Let X_1 and X_2 be the outcomes associated with a toss of a pair of dice
  - $E(X_1) = E(X_2) = 3.5$
  - $SD(X_1) = SD(X_2) = 1.708$
- Compute the following:
  - $E(X_1 + X_2) =$
  - $SD(X_1 + X_2) =$
- What does the distribution of $X_1 + X_2$ look like?

Summary of Session I

- A random variable describes the probabilities for an uncertain future numerical outcome of a random process
- A probability distribution is a rule that identifies possible outcomes of a random variable and assigns a probability to each
- The expected value *f*mean of a random variable is the weighted average of its values
- Variance is the weighted average of the squared deviations from the mean
- A probability distribution can be pictorially represented by a histogram, box–plot, probability density function
- Expected value of the sum of random variables is the sum of expected values of individual random variables
- Same rule applies to variance only if the variables are independent