# Descriptive Statistics

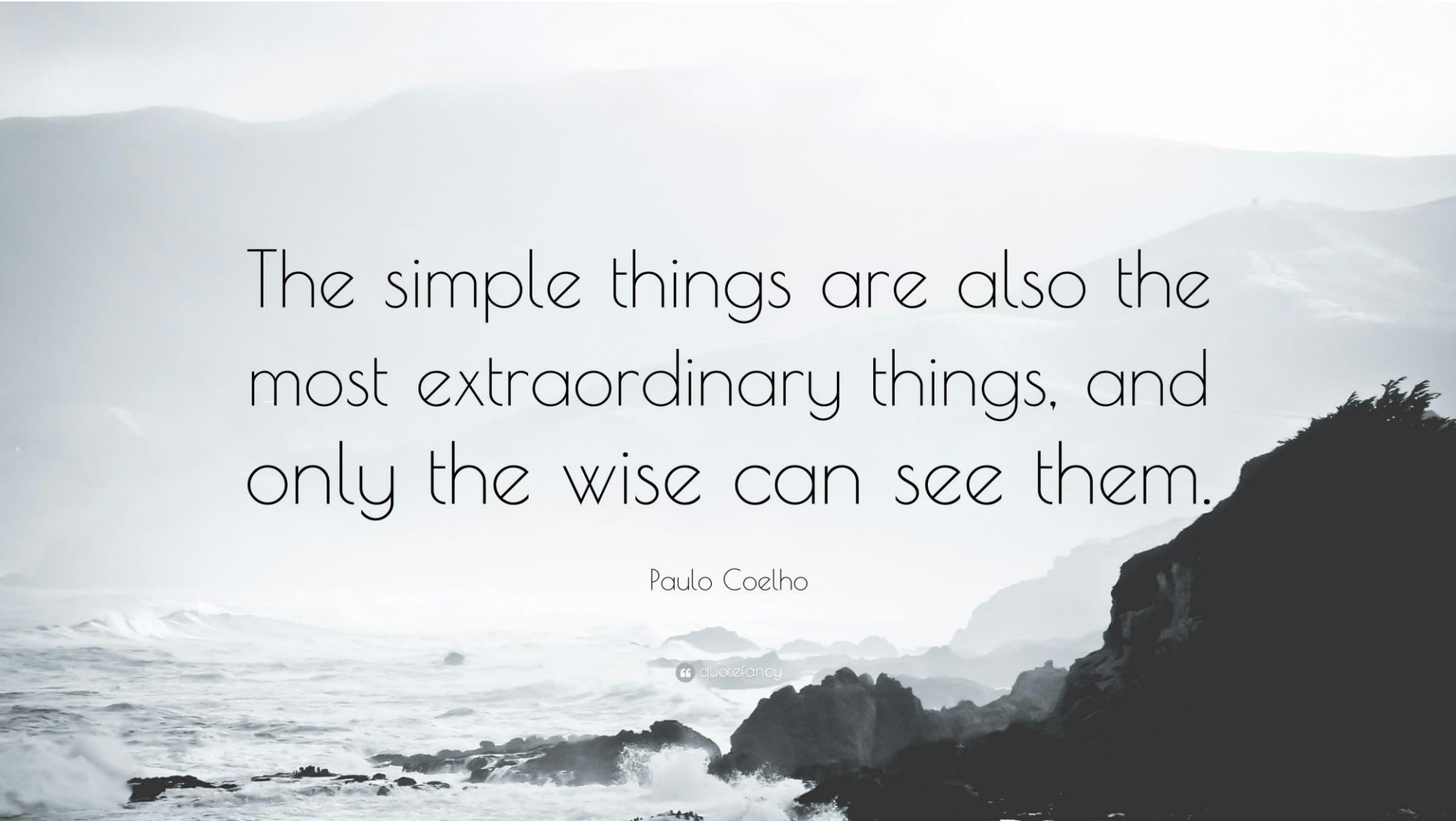
Course Taught at SUAD

**Dr. Tanujit Chakraborty**

Faculty @ Sorbonne

tanujitisi@gmail.com

# Quote of the day…



# Topics of the day…

- Role of Statistics and Data Analysis
- Data summarization
- Concepts of Descriptive Statistics
- Outlier Detection
- Graphical summarization

# Introduction

- We encounter data and make conclusions based on data every day.

- **Statistics** is the scientific discipline that provides methods to help us make sense of data.

- Statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us.

- The field of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

# Definition of Statistics

- Statistics consists of a body of methods for collecting analysing data.

- In order to avoid confusion with the statistical constants of the population (mean ($\mu$), variance ($\sigma^2$), etc.), which are usually referred to as parameters, statistical measures computed from the sample observations along e.g., mean ($\bar{x}$), variance ($s^2$), etc., have been termed as statistics.

- Let $X_1, X_2, \dots, X_n$ be random sampling of size $n$ from a population and let $T(x_1, \dots, x_n)$ be a real valued or vector valued function whose domain includes the sample space of $(X_1, \dots, X_n)$. Then the random variable or random vector $Y = T(X_1, \dots, X_n)$ is called a statistics.

The probability distribution of a statistic $Y$ is called the sampling distribution of $Y$.

# Why Study Statistics?

- Studying statistics will help us to collect data in a sensible way and then use the data to answer questions of interest.
- Studying statistics will allow us to critically evaluate the work of others by providing with the tools we need to make informed judgments.
- Throughout our personal and professional life, we will need to understand and use data to make decisions.
- To do this, we must be able to

  - Decide whether existing data is adequate or whether additional information is required.
  - If necessary, collect more information in a reasonable and thoughtful way.
  - Summarize the available data in a useful and informative manner.
  - Analyse the available data.
  - Draw conclusions, make decisions, and assess the risk of an incorrect decision.

# The Nature and Role of Variability

- Statistical methods allow us to collect, describe, analyse and draw conclusions from data.
- If we lived in a world where all measurements were identical for every individual, these tasks would be simple.
- **Example of No Variability:**

  Imagine a population consisting of all students at a particular university. Suppose that *every* student was enrolled in the same number of courses, spent exactly the same amount of money on textbooks this semester, and favoured increasing student fees to support expanding library services. For this population, there is *no* variability in number of courses, amount spent on books, or student opinion on the fee increase.
- **Example of Variability:**

  Let us consider the Mathematics score of all student of a particular batch.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 44 | 33 | 43 | 43 | 48 | 30 | 41 | 35 | 31 | 45 |
| 31 | 30 | 44 | 41 | 35 | 33 | 45 | 35 | 31 | 41 |

# Statistics and The Data Analysis Process

- The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to making informed conclusions based on the resulting data.

- The process can be organized into the following six steps:

1. Understanding the nature of the problem.
2. Deciding what to measure and how to measure it.
3. Data collection.
4. Data summarization and preliminary analysis.
5. Formal data analysis.
6. Interpretation of results.

# Example

- The admissions director at a large university might be interested in learning why some applicants who were accepted for the fall 2010 term failed to enroll at the university.
- The population of interest to the director consists of all accepted applicants who did not enroll in the fall 2021 term.
- Because this population is large and it may be difficult to contact all the individuals, the director might decide to collect data from only 300 selected students.
- These 300 students constitute a sample.
- Deciding how to select the 300 students and what data should be collected from each student are steps 2 and 3 in the data analysis process.

# Example (Continued)

- The next step in the process involves organizing and summarizing data.
- Methods for organizing and summarizing data, such as the use of tables, graphs, or numerical summaries, make up the branch of statistics called **descriptive statistics**.
- The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected.
- When we generalize in this way, we run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information.
- An important aspect in the development of inferential techniques involves quantifying the chance of an incorrect conclusion.

# Definitions

- ## Population and Sample:

  The entire collection of individuals or objects about which information is desired is called the **population** of interest.

  A **sample** is a subset of the population, selected for study.

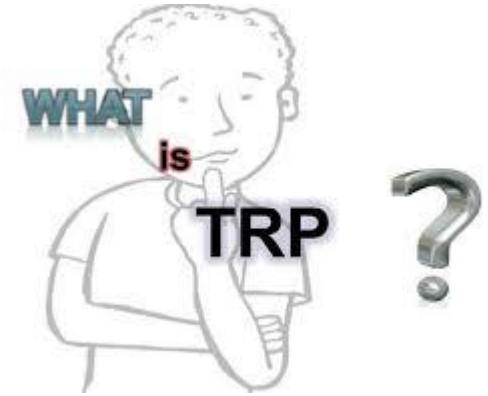
- ## Descriptive Statistics:

  It is the branch of statistics that includes methods for organizing and summarizing data.

- ## Inferential Statistics:

  It is the branch of statistics that involves generalizing from a sample to the population from which the sample was selected and assessing the reliability of such generalizations.

# TRP: An example




- Television rating point (TRP) is a tool provided to judge which programs are viewed the most.
  - This gives us an index of the choice of the people and also the popularity of a particular channel.
- For calculation purpose, a device is attached to the TV sets in few thousand viewers' houses in different geographic and demographic sectors.
  - The device is called as **People's Meter**. It reads the time and the programme that a viewer watches on a particular day for a certain period.
- An average is taken, for example, for a 30-days period.
- The above further can be augmented with a personal interview survey (PIS), which becomes the basis for many studies/decision making.
- Essentially, we are to analyze **data** for TRP estimation.

# Data

**Definition : Data**

A set of data is a collection of observed values representing one or more characteristics of some objects or units.

**Example:** For TRP, data collection consist of the following attributes.

- **Age:** A viewer's age in years
- **Sex:** A viewer's gender coded 1 for male and 0 for female
- **Happy:** A viewer's general happiness
  - NH for not too happy
  - PH for pretty happy
  - VH for very happy
- **TVHours:** The average number of hours a respondent watched TV during a day

# Data : Example

| Viewer# | Age | Sex | Happy | TVHours |
|---|---|---|---|---|
| … | … | … | … | … |
| … | … | … | … | … |
| **55** | 34 | F | VH | 5 |
| … | … | … | … | … |

**Note:**

- A data set is composed of information from a set of units.
- Information from a unit is known as an observation.
- An observation consists of one or more pieces of information about a unit; these are called variables.

# Type of Data

**Variables:**

A characteristic that varies from one person or thing to another is called a variable.

Example: height, weight, sex, marital status etc.

**Quantitative (or Numerical) Variable:**

A variable is numerical (or quantitative) if each observation is a number.

Example: height, weight etc.

**Qualitative (or Categorical) Variable:**

A variable is categorical (or qualitative) if the individual observations are categorical responses.

Example: sex, marital status etc.

# Type of Data

**Quantitative variable** can also be classified as either discrete or continuous.

**Discrete Variable:**

A variable is discrete if it has only a countable number of distinct possible values i.e. a variable is discrete if it can assume only a finite numbers of values.

Example: Number of defects.

**Continuous Variable:**

A numerical variable is called continuous variables if the set of possible values forms an entire interval on the numerical line.

Example: Length, temperature etc.

**Data:** A collection of observations on one or more variables is called data.

# Collecting Data

- Statisticians select their observations so that all relevant groups are represented in the data

  - Data can come from actual observations or from records that are kept for normal purpose.

  - Data can assist decision makers in educated guessed about the causes and therefore the probable effects of certain characteristics in given situations

  - When data are arranged in compact, usable forms, decision makers can take reliable information from the environment and use it to make intelligent decision.

# Population

Definition : **Population**

A population is a data set representing the entire entities of interest.

**Example:** All TV Viewers in the country/world.

**Note:**

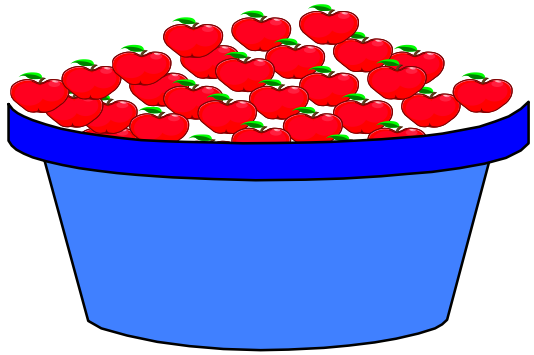
1. All people in the country/world is not a population.
2. For different survey, the population set may be completely different.
3. For statistical learning, it is important to define the population that we intend to study very carefully.

# Sample

Definition : **Sample**

The small number of items taken from the population to make a judgment of the population is called a Sample. The numbers of samples taken to make this judgment is called ***Sample size.***

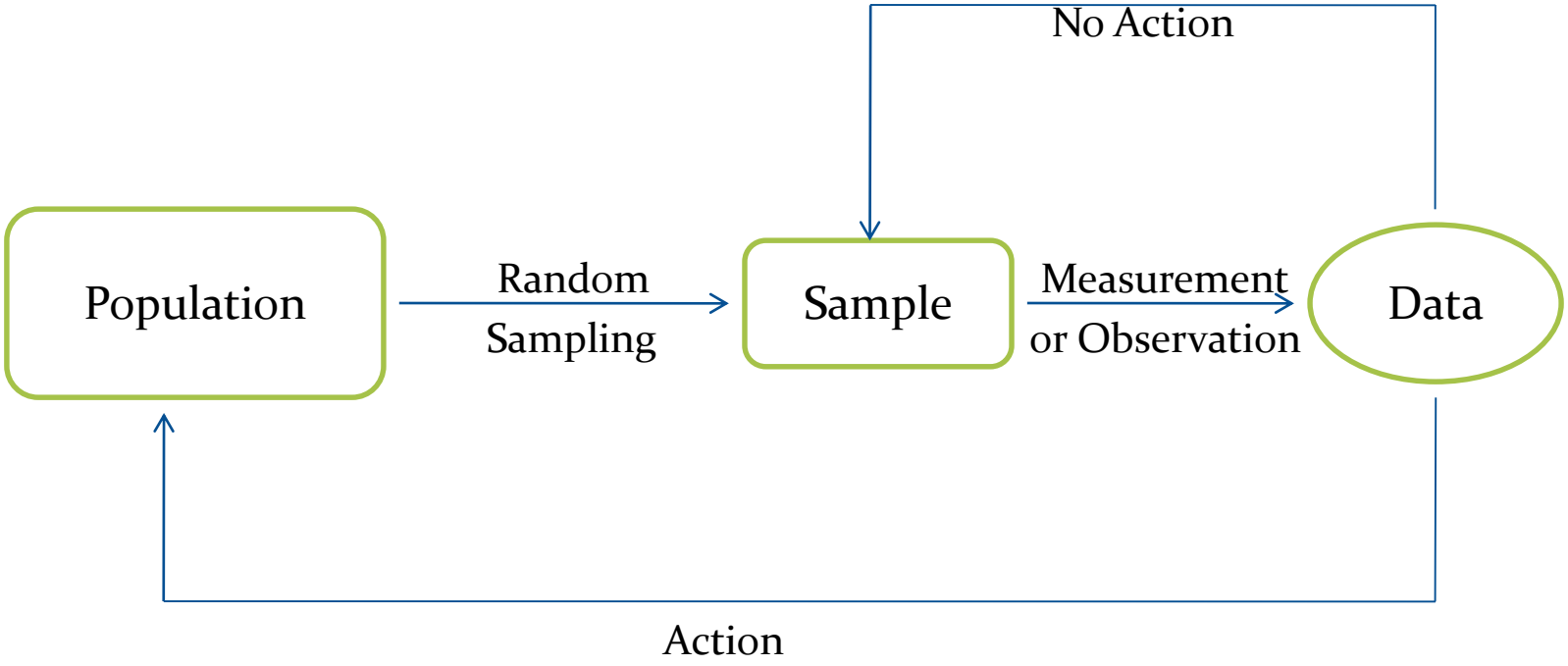
**Example:** All students studying BSc Mathematics and Data Science in SUAD is a sample, whereas all students belong to SUAD is population.



Population

Sample of size Three

# Population, Sample and Data




**Note:** Normally a sample is obtained in such a way as to be representative of the population.

# Statistic

**Definition : Statistic**

A statistic is a quantity calculated from data that describes a particular characteristics of a sample.

**Example:**

- The sample mean (denoted by $\bar{y}$) is the arithmetic mean of a variable of all the observations of a sample.
- **Statistic** (mean ($\bar{x}$), variance ($s^2$), etc.) consists of a body of methods for collecting analysing data.
- The probability distribution of a statistic $Y$ is called the **sampling distribution** of $Y$.

# Parameters and Statistic

- The purpose of statistical inference is to draw conclusions about population characteristics or **parameters** (mean ($\mu$), variance ($\sigma^2$), etc.)

- Let $X_1, X_2, \dots, X_n$ be random sampling of size $n$ from a population and let $T(x_1, \dots, x_n)$ be a real valued or vector valued function whose domain includes the sample space of $(X_1, \dots, X_n)$. Then the random variable or random vector

$$\boldsymbol{Y = T(X_1, \dots, X_n)}$$

  is called a **statistic**.

# Statistical Inference

Definition : **Statistical inference**

Statistical inference is the process of using sample statistic to make decisions about population.

**Example:** In the context of TRP

- Overall frequency of the various levels of happiness.
- Is there a relationship between the age of a viewers and his/her general happiness?
- Is there a relationship between the age of the viewer and the number of TV hours watched?

# Data Summarization

- To identify the typical characteristics of data (i.e., to have an overall picture).
- To identify which data should be treated as noise or outliers.
- The data summarization techniques can be classified into two broad categories:
  - Measures of **location**
  - Measures of **dispersion**

# Measurement of location

- It is also alternatively called as measuring the central tendency.
  - A function of the sample values that summarizes the location information into a single number is known as a measure of location.

- The most popular measures of location are
  - Mean
  - Median
  - Mode
  - Midrange

- These can be measured in three ways
  - Distributive measure
  - Algebraic measure
  - Holistic measure

## Distributive measure

- It is a measure (*i.e. function*) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (*i.e. entire*) data set.

**Example:** sum(), count()

## Algebraic measure

- It is a measure that can be computed by applying an algebraic function to one or more distributive measures.

**Example:** average = $\frac{\text{sum}()}{\text{count}()}$

## Holistic measure

- It is a measure that must be computed on the entire data set as a whole.

**Example**: Calculating median. What about *mode?*

# Mean of a sample

- The mean of a sample data is denoted as $\bar{x}$. Different mean measurements known are:
  - Simple mean
  - Weighted mean
  - Trimmed mean

- In the next few slides, we shall learn how to calculate the mean of a sample.

- We assume that given $x_1$, $x_2$, $x_3$,….., $x_n$ are the sample values.

# Simple mean of a sample

- **Simple mean**

  It is also called simply arithmetic mean or average and is abbreviated as (AM).

Definition : **Simple mean**

If $x_1, x_2, x_3, \ldots, x_n$ are the sample values, the simple mean is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

# Disadvantages of A.M

- It cannot be used if we are dealing with qualitative data.
- It cannot be obtained if a single observation is missing.
- It affected very much by extreme values.
- It may lead to wrong conclusions if the details of the data from which it is computed are not given.
  - Example: Let us consider the following marks obtained by two student A and B in three tests:

| Marks | Test I | Test II | Test III | Average |
|---|---|---|---|---|
| A | 50% | 60% | 70% | 60% |
| B | 70% | 60% | 50% | 60% |

# Weighted mean of a sample

- **Weighted mean**

  It is also called weighted arithmetic mean or weighted average.

Definition : **Weighted mean**

When each sample value $x_i$ is associated with a weight $w_i$, for $i = 1,2,\ldots,n$, then it is defined as

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

**Note:** *When all weights are equal, the weighted mean reduces to simple mean.*

# Weighted Mean: Rationale

- In calculating A.M we suppose that all the items in the distribution have equal importance. But in practice this may not be so.
- If some items in a distribution are more important than other then in order that average computed is representative of the distribution.
- In such cases, proper weightage is to be given to various item; the weight attached to each item being proportional to the importance of the item in the distribution.
- Example: A candidate obtained the following % of marks English 70, Maths 90, Statistics 75, Chemistry 88, Physics 79 and the weights are 1,2,2,3,3. Find the weighted mean score.

# Trimmed mean of a sample

- **Trimmed Mean**

  If there are extreme values (*also called outlier*) in a sample, then the mean is influenced greatly by those values. To offset the effect caused by those extreme values, we can use the concept of trimmed mean

Definition : **Trimmed mean**

Trimmed mean is defined as the mean obtained after chopping off values at the high and low extremes.

# Trimmed Means

- A **trimmed mean** is computed by first ordering the data values from smallest to largest, deleting a selected number of values from each end of the ordered list, and finally averaging the remaining values.
- The **trimming percentage** is the percentage of values deleted from *each* end of the ordered list.
- Sometimes the number of observations to be deleted from each end of the data set is specified.
    - Then the corresponding trimming percentage is calculated as

$$trimming\ percentage = \left(\frac{\text{number deleted from each end}}{n}\right).100$$

- If the number of observations to be deleted from each end resulting from this calculation is not an integer, it can be rounded to the nearest integer (which changes the trimming percentage a bit).

# Example

The following data describe the salaries of Basketball players:

| Players | J.J | T.Th | A.G | L.D | J.N | T.T | B.M |
|---|---|---|---|---|---|---|---|
| 2009 Salary | 6,600,000 | 4,743,598 | 1,000,497 | 10,370425 | 2,455,680 | 6,466,600 | 12,250,000 |

| Players | D.R | K.H | L.H | J.S | J. Jo | J.P | T.G |
|---|---|---|---|---|---|---|---|
| 2009 Salary | 5,184,480 | 9,500,000 | 1,306,455 | 5,456,000 | 1,594,080 | 2,000,000 | 1,039,800 |

# Properties of mean

- **Lemma 1:**

  If $\bar{x}_i$, $i = 1,2,\ldots,m$ are the means of $m$ samples of sizes $n_1, n_2, \ldots, n_m$ respectively, then the mean of the combined sample is given by:-

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

(Distributive Measure)

- **Lemma 2:**

  If a new observation $x_k$ is added to a sample of size $n$ with mean $\bar{x}$, the new mean is given by

$$\bar{x}' = \frac{n\,\bar{x} + x_k}{n+1}$$

# Properties of mean

- **Lemma 3:**

  If an existing observation $\boldsymbol{x_k}$ is removed from a sample of size $\boldsymbol{n}$ with mean $\boldsymbol{\bar{x}}$, the new mean is given by

$$\boldsymbol{\bar{x}' = \frac{n\,\bar{x} - x_k}{n-1}}$$

- **Lemma 4:**

  If $\boldsymbol{m}$ observations with mean $\boldsymbol{\bar{x}_m}$, are added (*removed*) from a sample of size $\boldsymbol{n}$ with mean $\boldsymbol{\bar{x}_n}$, then the new mean is given by

$$\boldsymbol{\bar{x} = \frac{n\,\bar{x}_n \pm m\,\bar{x}_m}{n \pm m}}$$

# Properties of mean

- **Lemma 5:**

  If a constant ***c*** is subtracted (*or added*) from each sample value, then the mean of the transformed variable is linearly displaced by *c*. That is,

$$\bar{x}' = \bar{x} \mp c$$

- **Lemma 6:**

  If each observation is called by multiplying (*dividing*) by a non-zero constant, then the altered mean is given by

$$\bar{x}' = \bar{x} * c$$

  *where, * is* ***x*** *(multiplication) or ÷ (division) operator.*

# Mean with grouped data

Sometimes data is given in the form of classes and frequency for each class.

| | | | | | | |
|---|---|---|---|---|---|---|
| *Class* → | $x_1$- $x_2$ | $x_2$- $x_3$ | ..... | $x_i$- $x_{i+1}$ | ..... | $x_{n-1}$- $x_n$ |
| *Frequency* → | $f_1$ | $f_2$ | ..... | $f_i$ | ..... | $f_n$ |

There three methods to calculate the mean of such a grouped data.

- Direct method
- Assumed mean method
- Step deviation method

## Examples: Compute the mean for the following data sets.

- Data Set 1: (Ungroup Data)

| $x$: | 20 | 37 | 4 | 20 | 0 | 84 | 14 | 36 | 5 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|

- Data Set 2: (Group Data)

| $x$ : | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $f$ : | 5 | 9 | 12 | 17 | 14 | 10 | 6 |

- Data Set 3: (Group Data)

| $Marks$: | $0-10$ | $10-20$ | $20-30$ | $30-40$ | $40-50$ | $50-60$ |
|---|---|---|---|---|---|---|
| $No.of$ $Student\ (f)$: | 12 | 18 | 27 | 20 | 17 | 6 |

# Direct method

- **Direct Method**

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Where, $\mathbf{x_i} = \frac{1}{2}$ **(lower limit + upper limit)** of the $i^{\text{th}}$ class, i.e., $x_i = \frac{x_i + x_{i+1}}{2}$ (also called class size), and $f_i$ is the frequency of the $i^{\text{th}}$ class.

**Note:** $\sum f_i (x_i - \bar{x}) = 0$

# Assumed mean method

- **Assumed Mean Method**

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

where, $A$ is the assumed mean (it is usually a value $x_i = \frac{x_i + x_{i+1}}{2}$ chosen in the middle of the groups $d_i = (A - x_i)$ for each $i$ )

# Step deviation method

- **Step deviation method**

$$\bar{x} = A + \left\{ \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \; h \right\}$$

where,

$A$ = assumed mean

$h$ = class size (*i.e.*, $\mathbf{x_{i+1} - x_i}$ *for the* $i^{th}$ *class*)

$u_i = \frac{x_i - A}{h}$

# Mean for a group of data

- For the above methods, we assume that…
  - All classes are equal sized
  - Groups are with inclusive classes, i.e., $x_i = x_{i-1}$(*linear limit of a class is same as the upper limit of the previous class*)

| 10 - 19 | 20 - 29 | 30 - 39 | 40 – 49 |
|---|---|---|---|

*Data with **ex**clusive classes*

| 9.5 – 19.5 | 19.5 – 29.5 | 29.5 – 39.5 | 39.5 – 49.5 |
|---|---|---|---|

*Data with **in**clusive classes*

# Ogive: Graphical method to find mean

- **Ogive** (pronounced as O-Jive) is a cumulative frequency polygon graph.
  - When cumulative frequencies are plotted against the upper (lower) class limit, the plot resembles one side of an Arabesque or **ogival** architecture, hence the name.
  - There are two types of Ogive plots
    - Less-than (upper class versus cumulative frequency)
    - More than (lower class versus cumulative frequency)

**Example:**
Suppose, there is a data relating the marks obtained by 200 students in an examination

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, .....

(Further, suppose it is observed that the minimum and maximum marks are 410, 479, respectively.)

# Ogive: Cumulative frequency table

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, .....

**Step 1:** Draw a cumulative frequency table

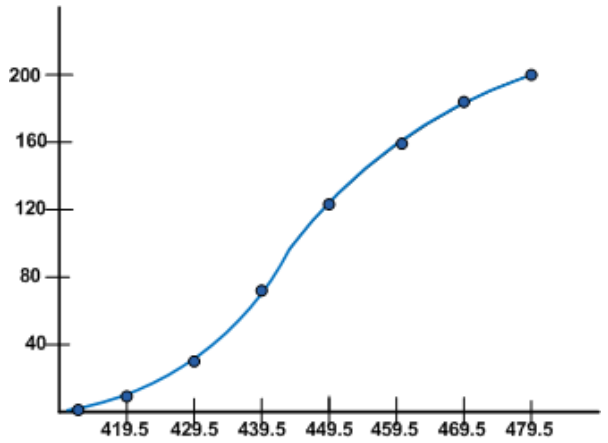
| Marks | Conversion into exclusive series | No. of students | Cumulative Frequency |
|---|---|---|---|
| (x) | | (f) | (C.M) |
| 410-419 | 409.5-419.5 | 14 | 14 |
| 420-429 | 419.5-429.5 | 20 | 34 |
| 430-439 | 429.5-439.5 | 42 | 76 |
| 440-449 | 439.5-449.5 | 54 | 130 |
| 450-459 | 449.5-459.5 | 45 | 175 |
| 460-469 | 459.5-469.5 | 18 | 193 |
| 470-479 | 469.5-479.5 | 7 | 200 |

# Ogive: Graphical method to find mean

| Marks | Conversion into exclusive series | No. of students | Cumulative Frequency |
|---|---|---|---|
| (x) | | (f) | (C.M) |
| 410-419 | 409.5-419.5 | 14 | 14 |
| 420-429 | 419.5-429.5 | 20 | 34 |
| 430-439 | 429.5-439.5 | 42 | 76 |
| 440-449 | 439.5-449.5 | 54 | 130 |
| 450-459 | 449.5-459.5 | 45 | 175 |
| 460-469 | 459.5-469.5 | 18 | 193 |
| 470-479 | 469.5-479.5 | 7 | 200 |

**Step 2:** Less-than Ogive graph

| Upper class | Cumulative Frequency |
|---|---|
| Less than 419.5 | 14 |
| Less than 429.5 | 34 |
| Less than 439.5 | 76 |
| Less than 449.5 | 130 |
| Less than 459.5 | 175 |
| Less than 469.5 | 193 |
| Less than 479.5 | 200 |

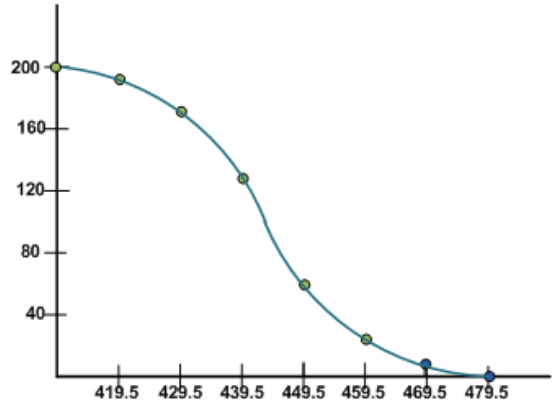


# Ogive: Graphical method to find mean

| Marks | Conversion into exclusive series | No. of students | Cumulative Frequency |
|---|---|---|---|
| (x) | | (f) | (C.M) |
| 410-419 | 409.5-419.5 | 14 | 14 |
| 420-429 | 419.5-429.5 | 20 | 34 |
| 430-439 | 429.5-439.5 | 42 | 76 |
| 440-449 | 439.5-449.5 | 54 | 130 |
| 450-459 | 449.5-459.5 | 45 | 175 |
| 460-469 | 459.5-469.5 | 18 | 193 |
| 470-479 | 469.5-479.5 | 7 | 200 |

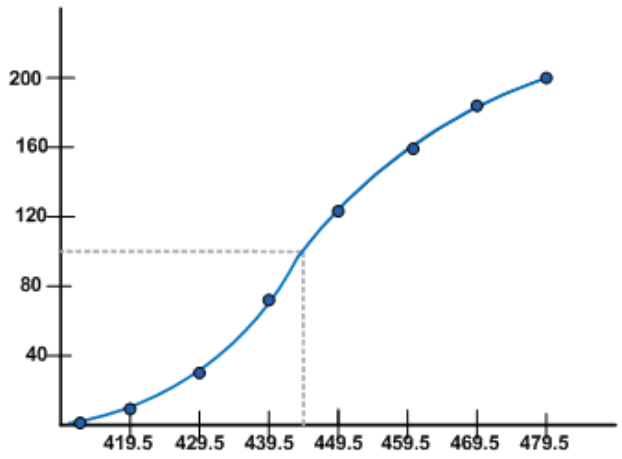
**Step 3:** More-than Ogive graph

| Lower class | Cumulative Frequency |
|---|---|
| More than 409.5 | 200 |
| More than 419.5 | 186 |
| More than 429.5 | 166 |
| More than 439.5 | 124 |
| More than 449.5 | 70 |
| More than 459.5 | 25 |
| More than 469.5 | 7 |

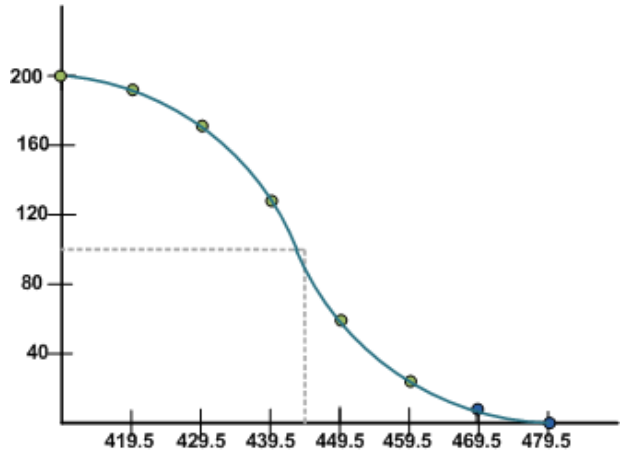


# Information from Ogive

- Mean from Less-than Ogive



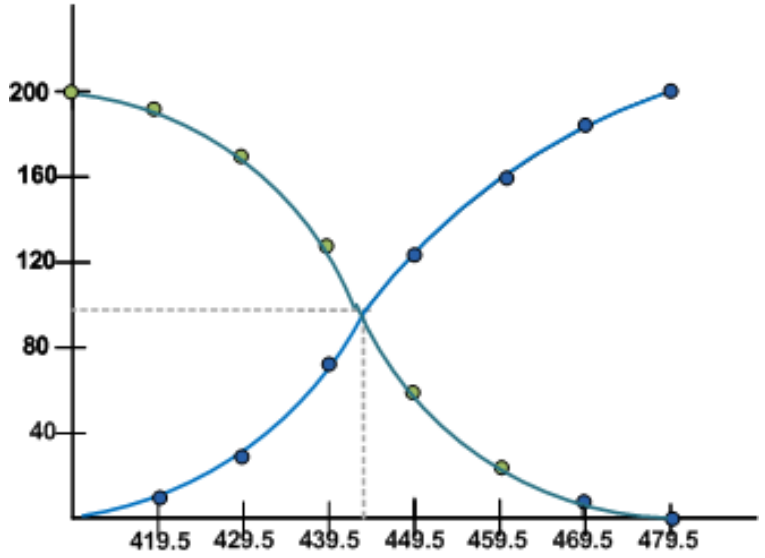

- Mean from More-than Ogive




- A % C freq of .65 for the third class 439.5…..449.5 means that 65% of all scores are found in this class or below.

# Information from Ogive

- Less-than and more-than Ogive approach




A cross point of two Ogive plots gives the mean of the sample

# Some other measures of mean

- There are three mean measures of location:
  - Arithmetic Mean (AM)
  - Geometric mean (GM)
  - Harmonic mean (HM)

  These three means are called Pythagorean means.

# Some other measures of mean

- Arithmetic Mean (**AM**)
  - $S: \{x_1, x_2\}$
  - $\bar{x} = \frac{x_1 + x_2}{2}$
  - $\bar{x} - x_1 = x_2 - \bar{x}$
- Geometric mean (**GM**)
  - $S: \{x_1, x_2\}$
  - $\tilde{x} = \sqrt{x_1 . x_2}$
  - $\frac{x_1}{\tilde{x}} = \frac{\tilde{x}}{x_2}$
- Harmonic Mean (**HM**)
  - $S: \{x_1, x_2\}$
  - $\hat{x} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$
  - $\frac{2}{\hat{x}} = \frac{1}{x_1} + \frac{1}{x_2}$

# Geometric mean

Definition : **Geometric mean**

Geometric mean of $\boldsymbol{n}$ observations (*none of which are zero*) is defined as:

$$\tilde{x} = \left( \prod_{i=1}^{n} x_i \right)^{1/n}$$

where, $n \neq 0$

**Note**

- GM is the arithmetic mean in "log space". This is because, alternatively,

$$log\tilde{x} = \frac{1}{n} \sum_{i=1}^{n} log\, x_i$$

- This summary of measurement is meaningful only when all observations are $> 0$.
- If at least one observation is zero, the product will itself be zero! For a negative value, root is not real

# Applications of GM

- The geometric mean is most useful to calculate the compounded growth rate where values in the sample are **not** independent of each other

**Example: Calculation of growth rate**

Consider a stock that grows by 10% in year one, declines by 20% in year two, and then grows by 30% in year three. What is the growth rate?

$$GM = \sqrt[3]{(1+0.1)(1-0.2)(1+0.3)}$$

$$= 0.046$$

= 4.6% annually.

# Applications of GM

- Other some applications:
  - if values tend to make large fluctuations.
  - when the values that are multiplied together are exponential value
    - the statistical rates of human population growth in consecutive 10 or 20 years, etc.

Why Geometric Mean is “geometric”?

# Harmonic mean

Definition : **Harmonic mean**

If all observations are non zero, the reciprocal of the arithmetic mean of the reciprocals of observations is known as harmonic mean.

For ungrouped data

$$\hat{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

For grouped data

$$\hat{x} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \left(\frac{f_i}{x_i}\right)}$$

where, $f_i$ is the frequency of the $i^{th}$ class with $x_i$ as the center value of the $i^{th}$ class.

# Applications of HM

- In fact, harmonic mean is the reciprocal of mean of sum of reciprocals
  - Harmonic means are often used in averaging things like rates.

**Example: Calculate travel speed given durations of several trips**

- A car in first 1 hour travels 60 kmph, in next 2 hours it travels with 90 kmph and next in 3 hours it travels with 80 kmph

Calculation using AM:

$$\text{Average speed} = \frac{Total\ disatnec}{Total\ time} = \frac{60+180+240}{1+2+3} = \frac{480}{6} = 80\ kmph$$

Calculation using HM

$$\text{Average speed} = \frac{1+2+3}{\frac{1}{60}+\frac{2}{90}+\frac{3}{80}} = \frac{6}{0.0167+0.0222+0.0375} = \frac{6}{0.0764} = 78.5\ kmph$$

# Applications of weighted HM

- The weighted harmonic mean of $x_1$, $x_2$, $x_3$ with the corresponding weights $w_1$, $w_2$, $w_3$ is given as:

$$whm = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

**Example: Calculation of price earning ratio (per)**

Consider two firms: one has a market capitalization of $100 billion and earnings of $4 billion (per of 25) and another with a market capitalization of $1 billion and earnings of $4 million (per of 250). In an index made of the two stocks, with 10% invested in the first and 90% invested in the second, what is the per of the index?

# Applications of weighted HM

- **Example: Calculation of price earning ratio (per)**

  Consider two firms: one has a market capitalization of $100 billion and earnings of $4 billion (per of 25) and another with a market capitalization of $1 billion and earnings of $4 million (per of 250). In an index made of the two stocks, with 10% invested in the first and 90% invested in the second, what is the per of the index?

  Calculation 1: Using weighted average mean

$$wam = 0.1 \times 25 + 0.9 \times 250 = 227.5$$

  Calculation 2: Using weighted harmonic mean

$$whm = \frac{0.1 + 0.9}{\frac{0.1}{25} + \frac{0.9}{250}} = 131.6$$

  As can be seen, the weighted arithmetic mean significantly overestimates the mean price-earnings ratio.

# Other way of using means

# Significant of different mean calculations

- There are two things involved when we consider a sample
  - Observation
  - Range

**Example:** Rainfall data

| **Rainfall (in mm)** | $r_1$ | $r_2$ | ... | $r_n$ |
|---|---|---|---|---|
| **Days (in number)** | $d_1$ | $d_2$ | ... | $d_n$ |

- Here, rainfall is the observation and day is the range for each element in the sample
- Here, we are to measure the mean "**rate of rainfall**" as the measure of location

# Significant of different mean calculations

- **Case 1: Range remains same for each observation**

**Example:** Having data about amount of rainfall per week, say.

| **Rainfall (in mm)** | 35 | 18 | … | 22 |
|---|---|---|---|---|
| **Days (in number)** | 7 | 7 | … | 7 |

# Significant of different mean calculations

- **Case 2: Ranges are different, but observation remains same**

**Example:** Same amount of rainfall in different number of days, say.

| **Rainfall (in mm)** | 50 | 50 | … | 50 |
|---|---|---|---|---|
| **Days (in number)** | 1 | 2 | … | 7 |

# Significant of different mean calculations

- **Case 3: Ranges are different, as well as the observations**

**Example:** Different amount of rainfall in different number of days, say.

| Rainfall (in mm) | 21 | 34 | … | 18 |
|---|---|---|---|---|
| Days (in number) | 5 | 3 | … | 7 |

# Rule of thumbs for means

- **AM:** When the range remains same for each observation

  Example: Case 1

| **Rainfall (in mm)** | **35** | **18** | **…** | **22** |
|---|---|---|---|---|
| **Days (in number)** | 7 | 7 | … | 7 |

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i$$

# Rule of thumbs for means

- **HM:** When the range is different but each observation is same
  - Example: Case 2

| **Rainfall (in mm)** | **50** | **50** | **…** | **50** |
|---|---|---|---|---|
| **Days (in number)** | 1 | 4 | … | 7 |

$$\tilde{x} = \frac{n}{\sum_1^n \frac{1}{x_i}}$$

$$x_i = \frac{r_i}{d_i}$$

# Rule of thumbs for means

- **GM:** When the ranges are different as well as the observations
  - Example: Case 3

| **Rainfall (in mm)** | **21** | **34** | **…** | **18** |
|---|---|---|---|---|
| **Days (in number)** | 5 | 3 | … | 7 |

$$\hat{x} = \left( \prod_{1}^{n} x_i \right)^{\frac{1}{n}} \qquad \text{where } x_i = r_i \times d_i$$

# Rule of thumbs for means

- The important things to recognize is that all three means are simply the arithmetic means in disguise!

- Each mean follows the “additive structure”.
  - Suppose, we are given some abstract quantities $\{x_1, x_2, \ldots, x_n\}$
  - Each of the three means can be obtained with the following steps
    1. Transform each $x_i$ into some $y_i$
    2. Taking the arithmetic mean of all $y_i$’s
    3. Transforming back the to the original scale of measurement

# Rule of thumbs for means

- For arithmetic mean
  - Use the **transformation $y_i = x_i$**
  - Take the arithmetic mean of all $y_i$ *s* to get $\bar{y}$
  - Finally, $\bar{x} = \bar{y}$
- For geometric mean
  - Use the **transformation $y_i = \log(x_i)$**
  - Take the arithmetic mean of all $y_i$ *s* to get $\bar{y}$
  - Finally, $\hat{x} = e^{\bar{y}}$
- For harmonic mean
  - Use the **transformation $y_i = \frac{1}{x_i}$**
  - Take the arithmetic mean of all $y_i$ *s* to get $\bar{y}$
  - Finally, $\tilde{x} = \frac{1}{\bar{y}}$
- If your sample contains all values which are same (i.e., invariant), then

AM ≥ GM ≥ HM

Try to Prove it for n positive real numbers.

Home Exercise (See https://arxiv.org/pdf/2003.02664.pdf)

Exercise 2: Suppose there are three observations with AM = 4, GM = 3.63, HM = 2.67. Find the three observations?

AM = GM = HM

# Median

- Median of a distribution is the value of the variable which divides it into two equal parts.

- Median is not at all affected by extreme values.

- Ungrouped Data:
  - If the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude.
  - In case of even number of observations, there are two middle values and median is obtained by taking the A.M of the middle values.

# Median of a sample

Definition : **Median of a sample**

Median of a sample is the middle value when the data are arranged in increasing (*or decreasing*) order. Symbolically,

$$\hat{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left\{x_{n/2} + x_{(\frac{n}{2}+1)}\right\} & \text{if } n \text{ is even} \end{cases}$$

- Median is not at all affected by extreme values.

# Calculating median from a set of samples

- Consider the case of three sets of data:
  - Set 1: $x_{11}, x_{12}, x_{13}, \ldots, x_{1m}$
  - Set 2: $x_{21}, x_{22}, x_{23}, \ldots, x_{2n}$
  - Set 3: $x_{31}, x_{32}, x_{33}, \ldots, x_{3p}$

What is median?
Calculate in a memory-constrained computing environment.
Distributive/ algebraic/ holistic approach?

# Median of a sample

Definition : **Median of a grouped data**

Median of a grouped data is given by

$$\hat{x} = l + \left\{ \frac{\frac{N}{2} - cf}{f} \; h \right\}$$

where $h$ = width of the median class

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

$f_i$ is the frequency of the $i^{th}$ class, and $n$ is the total number of groups

$cf$ = the cumulative frequency

$N$ = the total number of samples

$l$ = lower limit of the median class

**Note**

A class is called median class if its cumulative frequency is just greater than $N/2$

# Examples:

- Ex 1: (Ungroup data)

  (i) No. of observations: 25 20 15 35 18

  (ii) No. of observations: 8 20 50 25 15 30

- Ex 2: (Group data)

| $x$: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $f$: | 8 | 10 | 11 | 16 | 20 | 25 | 15 | 9 | 6 |

- Ex 3: (Group data)

| Wages (x) | 2000-3000 | 3000-4000 | 4000-5000 | 5000-6000 | 6000-7000 |
|---|---|---|---|---|---|
| No. of Employees (f) | 3 | 5 | 20 | 10 | 5 |

# Comparing The Mean and The Median

- When the histogram is symmetric, the point of symmetry is both the dividing point for equal areas and the balance point, and the mean and the median are equal.

- However, when the histogram is unimodal with a longer upper tail (positively skewed), the outlying values in the upper tail pull the mean up, so it generally lies above the median.

- When a unimodal histogram is negatively skewed, the mean is generally smaller than the median.

# Example

The accompanying data on number of minutes used for cell phone calls in one month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents **(Tele-Truth, March 2009)**:

189 0 189 177 106 201 0 212 0 306 0 0 59 224 0 189 142 83 71 165 236 0 142 236 130

a. Would you recommend the mean or the median as a measure of center for this data set? Give a brief explanation of your choice.

b. Compute a trimmed mean by deleting the three smallest observations and the three largest observations in the data set and then averaging the remaining 19 observations. What is the trimming percentage for this trimmed mean?

c. What trimming percentage would you need to use in order to delete all of the 0 minute values from the data set? Would you recommend a trimmed mean with this trimming percentage? Explain why or why not.

# Mode of a sample

- Mode is defined as the observation which occurs most frequently.
- For example, number of wickets obtained by bowler in 10 test matches are as follows.

1 2 0 3 2 4 1 1 2 2

- In other words, the above data can be represented as:-

| value | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| # of matches | 1 | 3 | 4 | 1 | 1 |

- Clearly, the mode here is “2”.
- If a distribution has two modes, then it is called **bimodal**.

# Mode of a grouped data

Definition : **Mode of a grouped data**

Select the modal class (it is the class with the highest frequency). Then the mode $\tilde{x}$ is given by:

$$\tilde{x} = l + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) h$$

where,

$h$ is the class width

$\Delta_1$ is the difference between the frequency of the modal class and the frequency of the class just after the modal class

$\Delta_2$ is the difference between the frequency of the modal class and the class just before the modal class

$l$ is the lower boundary of the modal class

**Note**

If each data value occurs only once, then there is no mode!

# Relation between mean, median and mode

- There is an empirical relation, valid for moderately skewed data

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

- A given set of data can be categorized into three categories:-
  - Symmetric data
  - Positively skewed data
  - Negatively skewed data

- To understand the above three categories, let us consider the following
- Given a set of $\boldsymbol{m}$ objects, where any object can take values $\boldsymbol{v_1, v_2, \ldots, v_k}$. Then, the frequency of a value $\boldsymbol{v_i}$ is defined as

$$\boldsymbol{Frequency(v_i) = \frac{Number\ of\ objects\ with\ value\ v_i}{n}}$$

$$for\ i = 1, 2, \ldots, k$$

## Example:

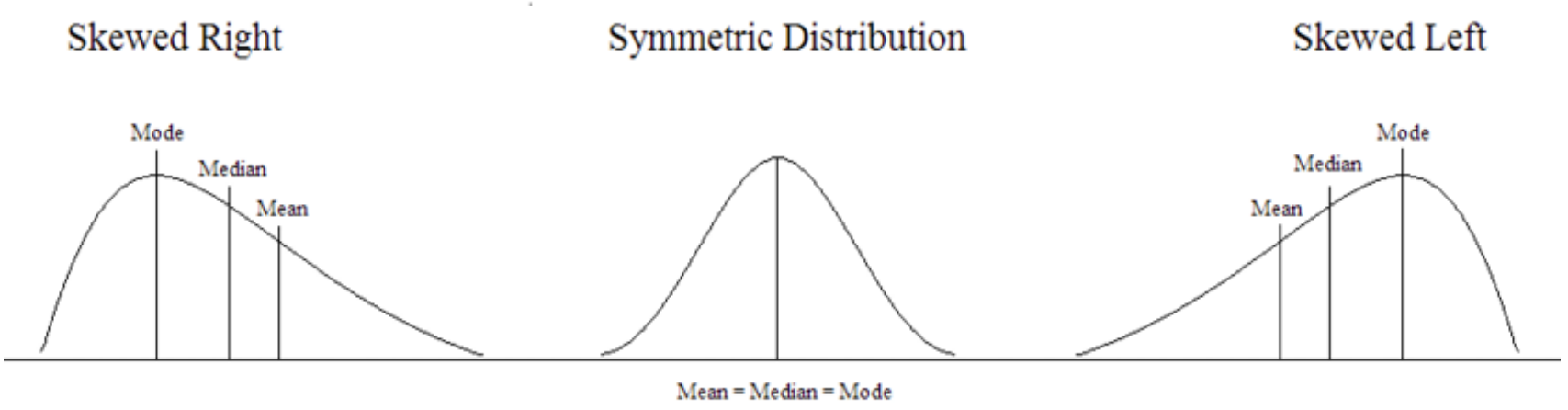
The accompanying data on number of minutes used for cell phone calls in one month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents **(Tele-Truth, March 2009)**:

189 0 189 177 106 201 0 212 0 306 0 0 59
224 0 189 142 83 71 165 236 0 142 236 130

Would you recommend the mean or the median or the mode as a measure of center for this data set? Give a brief explanation of your choice.

# Symmetric & Skewed data

- For symmetric data, all mean, median and mode lie at the same point.
- Positively Skewed Data: Mode occurs at a value smaller than the median.
- Negatively Skewed Data: Mode occurs at a value greater than the median.



# Midrange

- It is the average of the largest and smallest values in the set.

- Steps

1. A percentage ‘p’ between 0 and 100 is specified.
2. The top and bottom of (p/2)% of the data is thrown out
3. The mean is then calculated in the normal way

- Thus, the median is trimmed mean with p = 100% while the traditional mean corresponds to p = 0%

**Note**

- Trimmed mean is a special case of Midrange.

# Categorical Data

The **sample proportion of successes,** denoted by $\hat{p}$, is

$$\hat{p} = sample\ proportion\ of\ successes = \frac{\text{number of S's in the sample}}{n}$$

where $S$ is the label used for the response designated as success.

# Example

Suppose that 10 patients with meningitis received treatment with large doses of penicillin. Three days later, temperatures were recorded, and the treatment was considered successful if there had been a reduction in a patient's temperature. Denoting success by S and failure by F, the 10 observations are

S S F S S S F F S S

a) What is the value of the sample proportion of successes?

b) Replace each S with a 1 and each F with a 0. Then calculate $\bar{x}$ for this numerically coded sample. How does $\bar{x}$ compare to $\hat{p}$?

c) Suppose that it is decided to include 15 more patients in the study. How many of these would have to be S's to give $\hat{p} = .80$ for the entire sample of 25 patients?

# Measures of dispersion

- Location  measure are far too insufficient to understand data.
- Another set of commonly used summary statistics for continuous data are those that measure the dispersion.
- A dispersion measures the extent of spread of observations in a sample.

- Some important measure of dispersion are:
  - Range
  - Variance and Standard Deviation
  - Mean Absolute Deviation (MAD)
  - Absolute Average Deviation (AAD)
  - Interquartile Range (IQR)

# Measures of dispersion

**Example**

- Suppose, two samples of fruit juice bottles from two companies ***A*** and ***B***. The unit in each bottle is measured in litre.

| | | | | | |
|---|---|---|---|---|---|
| **Sample A** | 0.97 | 1.00 | 0.94 | 1.03 | 1.06 |
| **Sample B** | 1.06 | 1.01 | 0.88 | 0.91 | 1.14 |

- Both samples have same mean. However, the bottles from company A with more uniform content than company B.
- We say that the dispersion (or variability) of the observation from the average is less for A than sample B.
  - The variability in a sample should display how the observation spread out from the average
  - In buying juice, customer should feel more confident to buy it from A than B

# Range of a sample

Definition : **Range of a sample**

Let $\mathbf{X = x_1, \ldots, x_n}$ be **n** sample values that are arranged in increasing order.

The range **R** of these samples are then defined as:

$$\mathbf{R = max(X) - min(X) = x_n - x_1}$$
$$\mathbf{= Largest\ observation - Smallest\ observation}$$

- Range identifies the maximum spread, it can be misleading if most of the values are concentrated in a narrow band of values, but there are also a relatively small number of more extreme values.
- The variance is another measure of dispersion to deal with such a situation.

# Variance and Standard Deviation

Definition : **Variance and Standard Deviation**

Let $\mathbf{X = \{x_1, \ldots, x_n\}}$ are sample values of **n** samples. Then, variance denoted as $\boldsymbol{\sigma^2}$ is defined as :-

$$\boldsymbol{\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{S_{xx}}{n-1}}$$

where, $\bar{x}$ denotes the mean of the sample

The standard deviation, $\boldsymbol{\sigma}$, of the samples is the square root of the variance $\boldsymbol{\sigma^2}$

The **sample standard deviation** is the positive square root of the sample variance and is denoted by $s$.

**Why *(n-1)* is in the denominator instead of *n*?**

# Variance and Standard Deviation

- **Lemma:** If data are transformed as $\mathbf{x' = \frac{(x-a)}{c}}$, the variance is transformed as

$$\boldsymbol{\sigma'^2 = \frac{1}{c^2}\sigma^2}$$

**Proof**

The new mean $\boldsymbol{\bar{x}' = \frac{\bar{x}-a}{c}}$

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i' - \bar{x}'\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(x_i-a)}{c} - \frac{(\bar{x}-a)}{c}\right\}^2$$

$$= \frac{1}{c^2 n}\sum_{i=1}^{n}\{(x_i - a) - (\bar{x} - a)\}^2$$

$$= \frac{1}{c^2 n}\sum_{i=1}^{n}\{x_i - \bar{x}\}^2$$

$$= \frac{1}{c^2}\sigma^2$$ [PROVED]

# Exercise

**The Insurance Institute for Highway Safety (www.iihs.org, June 11, 2009)** published data on repair costs for cars involved in different types of accidents. In one study, seven different 2009 models of mini- and micro-cars were driven at 6 mph straight into a fixed barrier. The following table gives the cost of repairing damage to the bumper for each of the seven models:

| Model | Repair Cost |
|---|---|
| Smart Fortwo | $1,480 |
| Chevrolet Aveo | $1,071 |
| Mini Cooper | $2,291 |
| Toyota Yaris | $1,688 |
| Honda Fit | $1,124 |
| Hyundai Accent | $3,476 |
| Kia Rio | $3,701 |

# Exercise (Continuation)

(a) Compute the values of the variance and standard deviation. The standard deviation is fairly large. What does this tell you about the repair costs?

(b) The **Insurance Institute for Highway Safety** (referenced in the previous exercise) also gave bumper repair costs in a study of six models of minivans **(December 30, 2007)**. Write a few sentences describing how mini- and micro-cars and minivans differ with respect to typical bumper repair cost and bumper repair cost variability.

| Model | Repair Cost |
|---|---|
| Honda Odyssey | $1,538 |
| Dodge Grand Caravan | $1,347 |
| Toyota Sienna | $840 |
| Chevrolet Uplander | $1,631 |
| Kia Sedona | $1,176 |
| Nissan Quest | $1,603 |

# Coefficient of Variation

- **Basic properties**
  - **σ** measures spread about mean and should be chosen only when the mean is chosen as the measure of central tendency
  - **σ = 0** only when there is no spread, that is, when all observations have the same value, otherwise **σ > 0**

Definition : **Coefficient of variation**

A related measure is the coefficient of variation **CV**, which is defined as follows

$$\mathbf{CV} = \frac{\boldsymbol{\sigma}}{\bar{\mathbf{x}}} \times \mathbf{100}$$

This gives a ratio measure to spread.

# Coefficient variation

- **Significance of CV**
  - It is a statistical measure of the dispersion of data points in a data series around the mean
  - **CV = p%** implies that standard deviation is p% to that of the mean of a sample.

**Example:**

- Suppose, there are three series of data representing amount of returns that investors receives from three farms F1, F2 and F3 in a year.
- CV(F1), CV(F2), CV(F3) indicate volatilities/ risks in comparison of return expected from investment

$$\mathbf{CV} = \frac{\mathbf{Volatilty}}{\mathbf{Expected\ Return}} \times \mathbf{100}$$

# Coefficient variation

Mr. X is looking for a safe investment that provides safe and stable returns. There are following options:

a) **Stocks:** The volatility of the stock is 10% and the expected return is 14%.
b) **Mutual funds:** It offers an expected return of 13% with a volatility of 7%.
c) **Fixed deposits:** This scheme offers an expected return of 3% with 2% volatility.

In order to select the most suitable investment opportunity, Mr. X should choose which investment scheme?

# Mean Absolute Deviation (MAD)

- Since, the mean can be distorted by outlier, and as the variance is computed using the mean, it is thus sensitive to outlier. To avoid the effect of outlier, there are two more robust measures of dispersion known. These are:

  - Mean Absolute Deviation (MAD)

$$\textbf{MAD (X) = median } (\{|\mathbf{x_1} - \bar{\mathbf{x}}|, \ldots\ldots, |\mathbf{x_n} - \bar{\mathbf{x}}|\})$$

  - Absolute Average Deviation (AAD)

$$\textbf{AAD(X)} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} |\mathbf{x_i} - \bar{\mathbf{x}}|$$

where, X = $\{\mathbf{x_1}, \mathbf{x_2}, \ldots\ldots, \mathbf{x_n}\}$*is the sample values of n observations*

# Interquartile Range

- Like MAD and AAD, there is another robust measure of dispersion known, called as Interquartile range, denoted as IQR
- To understand IQR, let us first define *percentile* and *quartile*
- **Percentile**
  - The percentile of a set of ordered data can be defined as follows:
    - Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the $\mathbf{p^{th}}$ percentile $\mathbf{x_p}$ is a value of **x** such that **p**% of the observed values of **x** are less than $\mathbf{x_p}$
    - Example: The $\mathbf{50^{th}}$ percentile is that value $\mathbf{x_{50\%}}$ such that **50%** of all values of **x** are less than $\mathbf{x_{50\%}}$.
  - **Note:** The median is the $\mathbf{50^{th}}$ percentile.

# Interquartile Range

- **Quartile**
  - The most commonly used percentiles are quartiles.
    - The first quartile, denoted by $\mathbf{Q_1}$ is the $\mathbf{25^{th}}$ percentile.
    - The third quartile, denoted by $\mathbf{Q_3}$ is the $\mathbf{75^{th}}$ percentile
    - The median, $\mathbf{Q_2}$ is the $\mathbf{50^{th}}$ percentile.

- The quartiles including median, give some indication of the center, spread and shape of a distribution.

- The distance between $\mathbf{Q_1}$ and $\mathbf{Q_3}$ is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (**IQR**) and is defined as

$$\mathbf{IQR = Q_3 - Q_1}$$

# Example

The accompanying data on number of minutes used for cell phone calls in 1 month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents **(TeleTruth, March 2009)**:

189 0 189 177 106 201 0 212 0 306 0 0

59 224 0 189 142 83 71 165 236 0 142 236 130

a) Compute the values of the quartiles and the interquartile range for this data set.

b) Explain why the lower quartile is equal to the minimum value for this data set. Will this be the case for every data set? Explain.

# Application of IQR

- **Outlier detection using five-number summary**
  - A common rule of the thumb for identifying suspected outliers is to single out values falling at least **1.5 × IQR** above $\mathbf{Q_3}$ and below $\mathbf{Q_1}$.
  - In other words, extreme observations occurring within **1.5 × IQR** of the quartiles

# Application of IQR

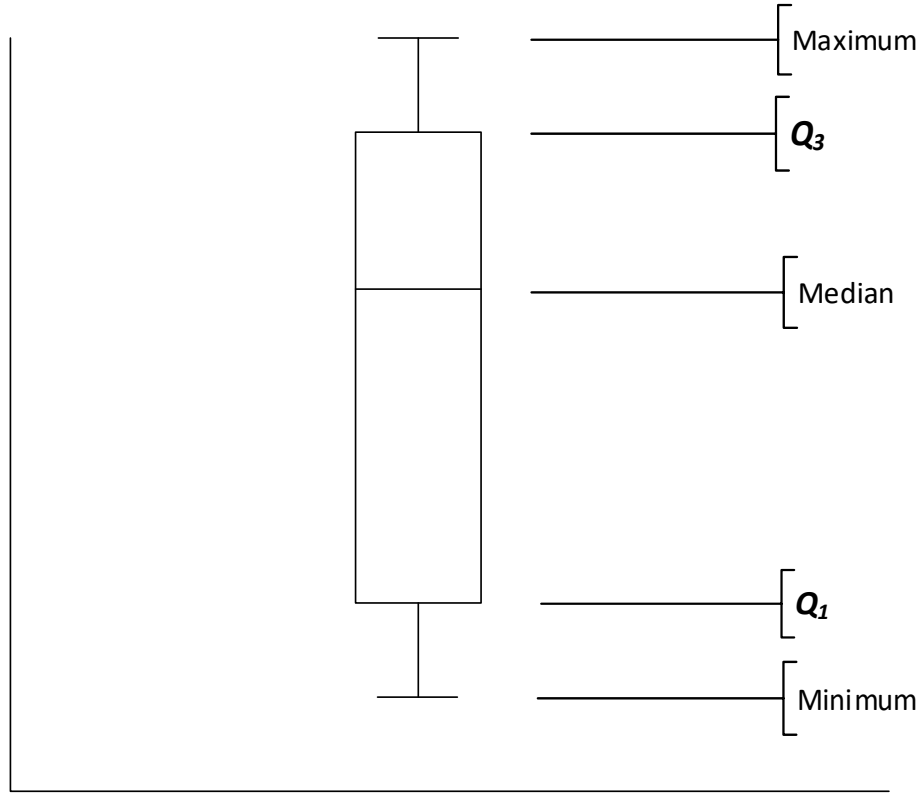
- **Five Number Summary**
  - Since, $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ together contain no information about the endpoints of the data, a complete summary of the shape of a distribution can be obtained by providing the lowest and highest data value as well. This is known as the five-number summary
  - The five-number summary of a distribution consists of :
    - The Median $\mathbf{Q_2}$
    - The first quartile $\mathbf{Q_1}$
    - The third quartile $\mathbf{Q_3}$
    - The smallest observation
    - The largest observation

  These are, when written in order gives the **five-number summary**:

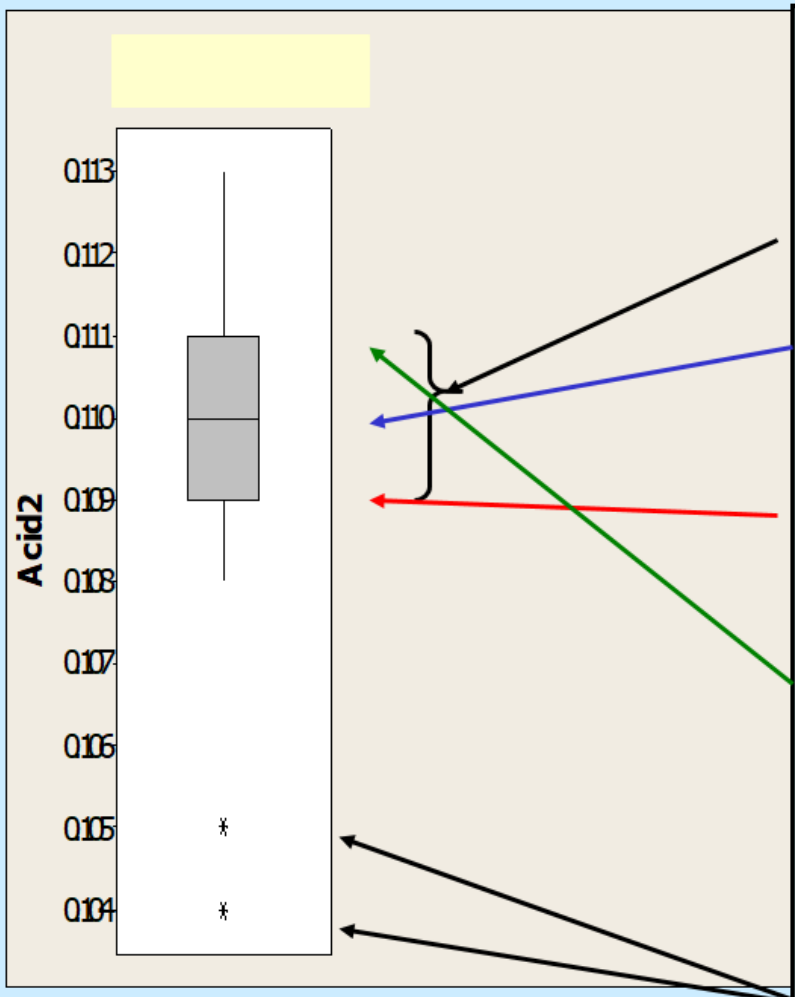
  Minimum, $\mathbf{Q_1}$, Median ($\mathbf{Q_2}$), $\mathbf{Q_3}$, Maximum

# Box plot

- **Graphical view of Five number summary**



# Box plot




**A box and whisker plot provides a 5 point summary of the data.**

1)The box represents the middle 50% of the data.

2) The median is the point where 50% of the data is above it and 50% below it.

3) The 1st quartile is where, 25% of the data fall below it.

4) The 3rd quartile is where, 75% of the data is below it.

5) The whiskers cannot extend any further than 1.5 times the length of the inner quartiles.

If you have data points outside this, they will show up as outliers.

# Construction of a Boxplot

- An observation is an **outlier** if it is more than $1.5(IQR)$ away from the nearest quartile (the nearest end of the box).
- An outlier is **extreme** if it is more than $3(IQR)$ from the nearest quartile and it is **mild** otherwise.
- Construction of a Modified Boxplot
  - Draw a horizontal (or vertical) measurement scale.
  - Construct a rectangular box with a left (or lower) edge at the lower quartile and right (or upper) edge at the upper quartile.
  - The box width is then equal to the iqr.
  - Draw a vertical (or horizontal) line segment inside the box at the location of the median.
  - Determine if there are any mild or extreme outliers in the data set.
  - Draw whiskers that extend from each end of the box to the most extreme observation that is *not* an outlier.
  - Draw a solid circle to mark the location of any mild outliers in the data set.
  - Draw an open circle to mark the location of any extreme outliers in the data set.

# Example

The following data is on the % of the population with a bachelor's or higher degree in 2007 for each of the 50 U.S. states.

21 27 26 19 30 35 35 26 47 26 27 30 24 29 22 24
29 20 20 27 35 38 25 31 19 24 27 27 23 34 34 25
32 26 26 24 22 28 26 30 23 25 22 25 29 33 34 30
17 25 23

Construct a boxplot.

# Exercise

Fiber content (in grams per serving) and sugar content (in grams per serving) for 18 high fiber cereals **(www.consumerreports.com)** are shown below.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Fiber Content** | 7 | 10 | 10 | 7 | 8 | 7 | 12 | 12 | 8 |
| | 13 | 10 | 8 | 12 | 7 | 14 | 7 | 8 | 8 |
| **Sugar Content** | 11 | 6 | 14 | 13 | 0 | 18 | 9 | 10 | 19 |
| | 6 | 10 | 17 | 10 | 10 | 0 | 9 | 5 | 11 |

a) Find the median, quartiles, and interquartile range for the fiber content data set.

b) Find the median, quartiles, and interquartile range for the sugar content data set.

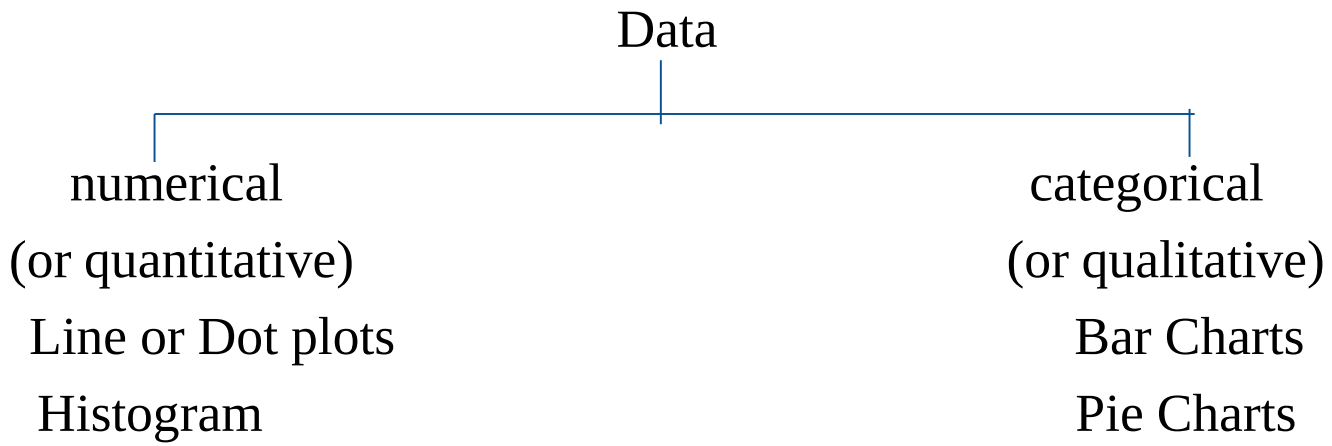
c) Are there any outliers in the sugar content data set?

d) Explain why the minimum value for the fiber content data set and the lower quartile for the fiber content data set are equal.

e) Construct a comparative boxplot and use it to comment on the differences and similarities in the fiber and sugar distributions.

# Graphical Representation of Data

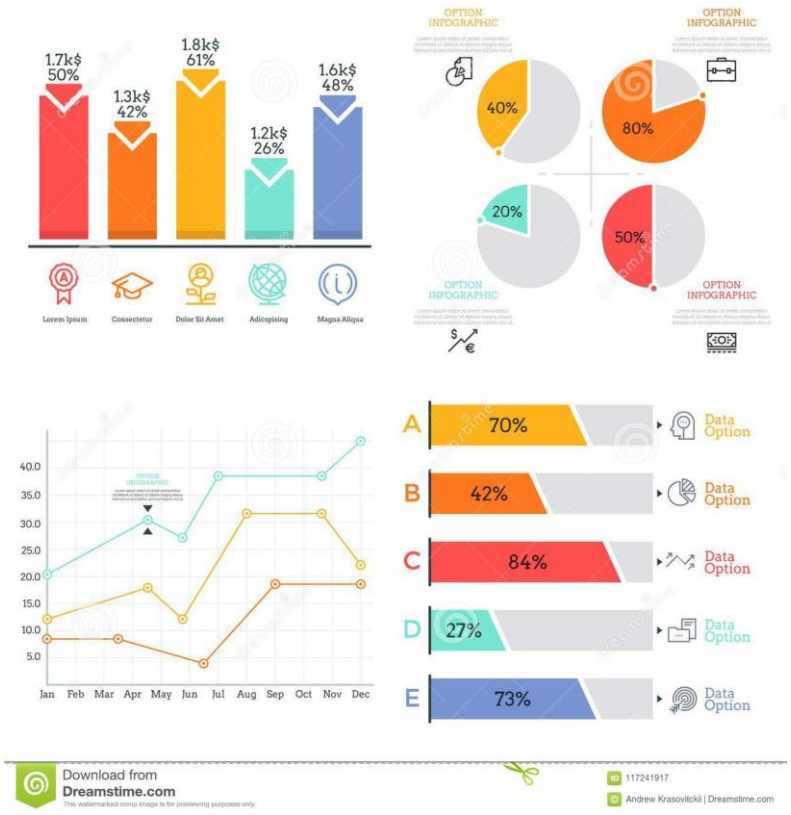
- Visualization techniques are ways of creating and manipulating graphical representations of data.
- We use these representations in order to gain better insight and understanding of the problem we are studying - pictures can convey an overall message much better than a list of numbers.



# Frequency Distributions & Different Charts

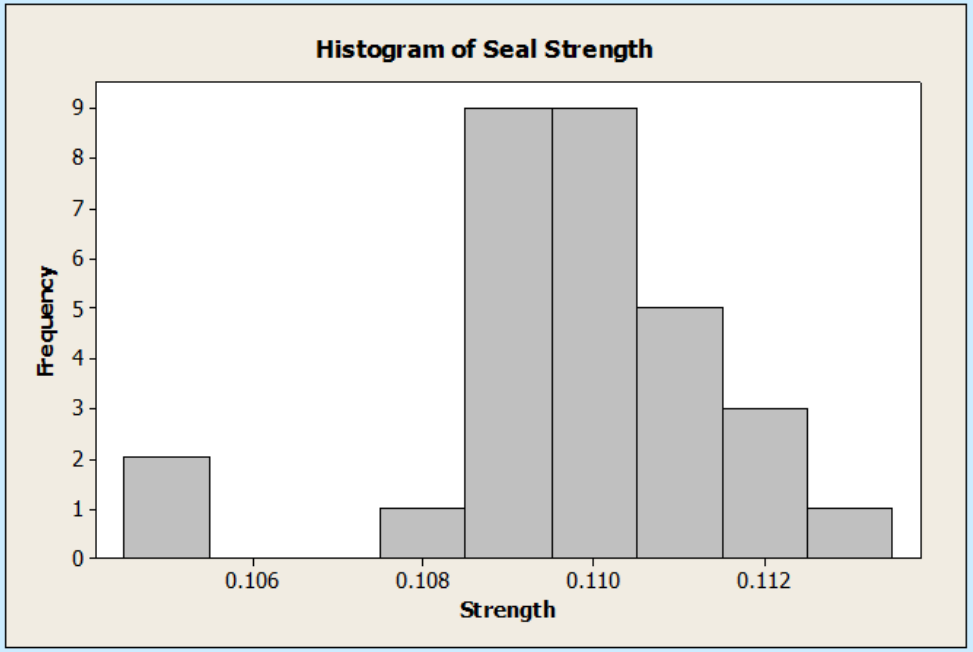
- When we deal with large sets of data, a good overall picture and sufficient information can be often conveyed by distributing the data into a number of classes or class intervals.

- To determine the number of elements belonging to each class, called class frequency.



# Histogram

Histogram is a basic graphing tool that displays the relative frequency or occurrence of continuous data values showing which values occur most and least frequently.




A histogram illustrates the

**Shape**,
Centering, and
Spread

of data distribution

and indicates whether there are any outliers.

# How to use Histogram?

- Collect at least 50 or more observations .
- Determine the maximum (L) and the minimum (S) of the data.
- Obtain the range of data as R=L-S
- Decide the number of classes (K) from the following table:

| NO. OF DATA POINTS | NO. OF CLASSES (APPROX.) |
|---|---|
| 50-100 | 5-10 |
| 100-250 | 7-12 |
| 250 & above | 10-20 |

- Decide the width of class interval (h) with convenient rounding as

$$h = R/K.$$

- Check the least count.
- Make the horizontal (X) axis with the class intervals in the scale of data points.
- Make the vertical (Y) axis with the frequency scale as absolute number or percent of total observations.

# Exercise:

Consider the data given below on the scores in mathematics for 56 students and draw the Histogram.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 44 | 33 | 43 | 43 | 53 | 44 | 57 | 47 | 33 | 29 | 59 | 37 | 57 | 47 |
| 61 | 49 | 31 | 45 | 38 | 46 | 50 | 48 | 45 | 46 | 55 | 46 | 47 | 49 |
| 60 | 55 | 54 | 35 | 43 | 39 | 41 | 47 | 46 | 46 | 57 | 57 | 42 | 65 |
| 43 | 55 | 52 | 53 | 45 | 35 | 51 | 35 | 46 | 51 | 40 | 50 | 43 | 54 |

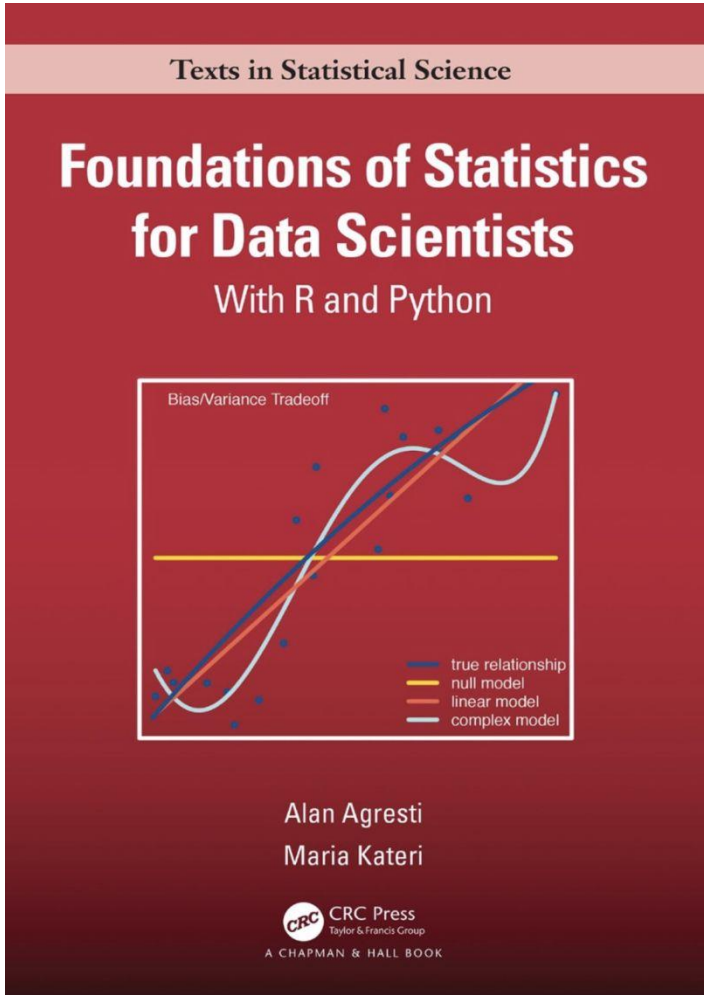
# Exercise:

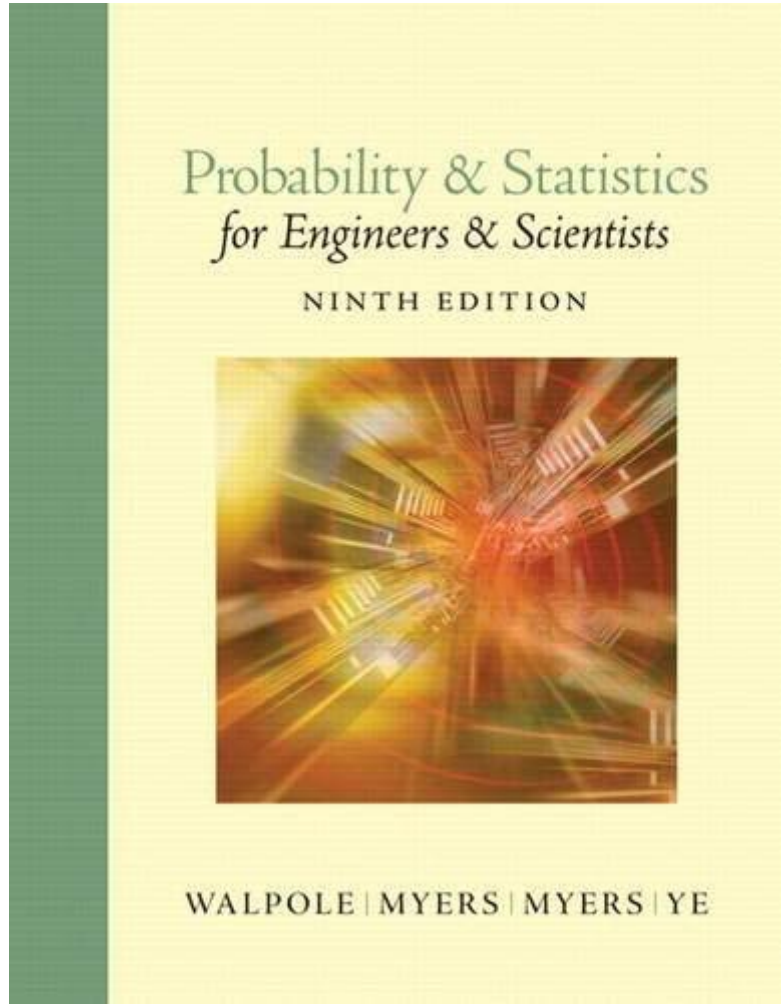
States differ widely in the percentage of college students who are enrolled in public institutions. **The National Center for Education Statistics** provided the accompanying data on this percentage for the 50 U.S. states for fall 2007. Is the histogram approximately symmetric, positively skewed, or negatively skewed? Would you describe the histogram as unimodal, bimodal, or multimodal?

**Percentage of College Students Enrolled in Public Institutions**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 96 | 86 | 81 | 84 | 77 | 90 | 73 | 53 | 90 | 96 | 73 |
| 93 | 76 | 86 | 78 | 76 | 88 | 86 | 87 | 64 | 60 | 58 |
| 89 | 86 | 80 | 66 | 70 | 90 | 89 | 82 | 73 | 81 | 73 |
| 72 | 56 | 55 | 75 | 77 | 82 | 83 | 79 | 75 | 59 | 59 |
| 43 | 50 | 64 | 80 | 82 | 75 | | | | | |

# References:


Texts in Statistical Science
Foundations of Statistics for Data Scientists
With R and Python
Bias/Variance Tradeoff
true relationship
null model
linear model
complex model
Alan Agresti
Maria Kateri
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK


Probability & Statistics
for Engineers & Scientists
NINTH EDITION
WALPOLE | MYERS | MYERS | YE