Day 2: Business Data Analytics

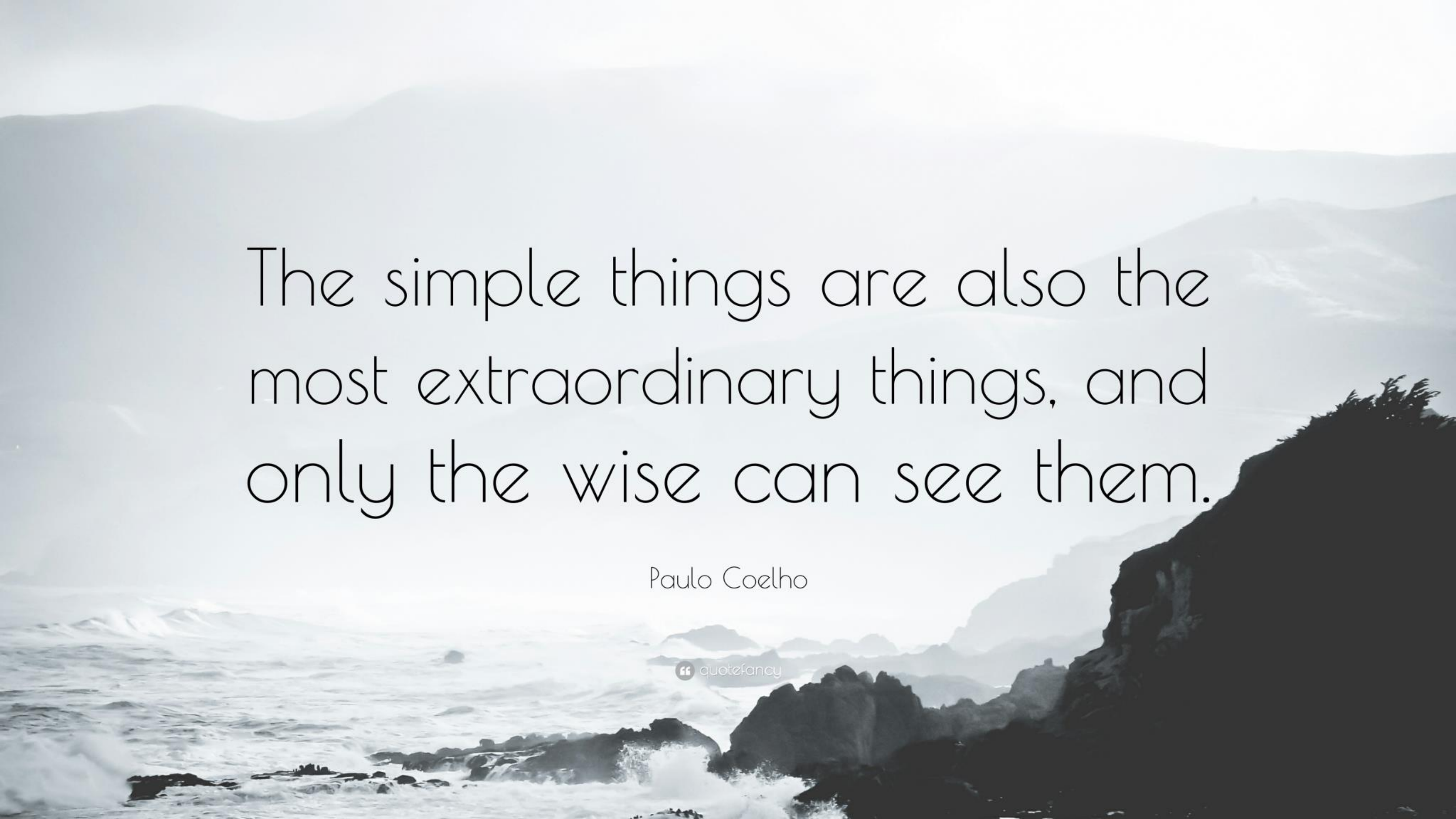
Descriptive Statistics

Dr. Tanujit Chakraborty

Faculty @ Sorbonne

tanujitisi@gmail.com

Quote of the day…



The simple things are also the most extraordinary things, and only the wise can see them.

Paulo Coelho

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.

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**.

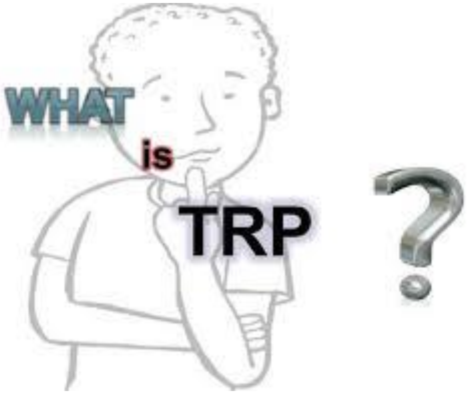
- The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected.

- 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.

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:

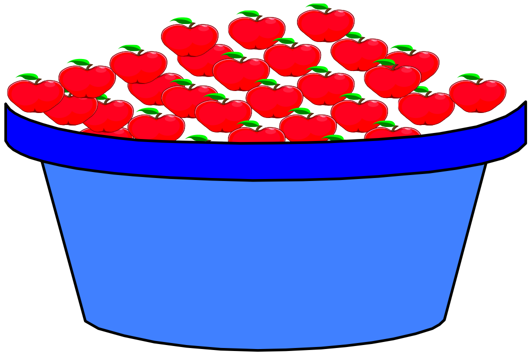
1. All people in the country/world is not a population.
2. For different survey, the population set may be completely different.
3. For statistical learning, it is important to define the population that we intend to study very carefully.

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.***

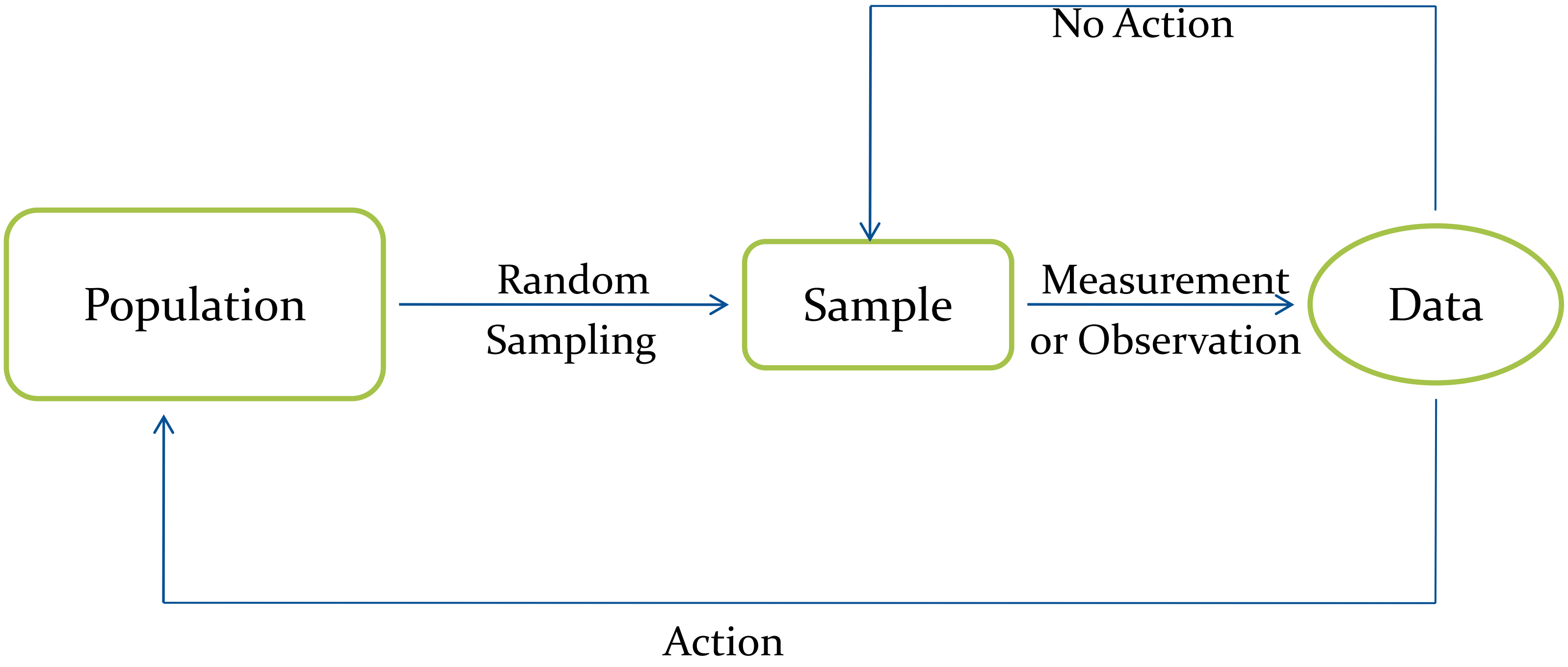
Example: All students studying BSc Mathematics and Data Science in SUAD is a sample, whereas all students belong to SUAD is population.



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 (μ), variance (σ^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**
  - Measures of **Shapes**

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

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,….., x_n are the sample values, the simple mean is defined as

$$\boldsymbol{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_{\mathrm{i}}}$$

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%

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

Examples: Compute the mean for the following data sets.

- Data Set 1: (Ungroup Data)

$$x:\quad 20 \quad 37 \quad 4 \quad 20 \quad 0 \quad 84 \quad 14 \quad 36 \quad 5 \quad 19$$

- 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

Mean with Grouped Data

- **Direct Method**

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

Where, $\mathbf{x_i} = \frac{1}{2}$ **(lower limit + upper limit)** of the i^{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^{th} class.

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

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,

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

- Median is not at all affected by extreme values.

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

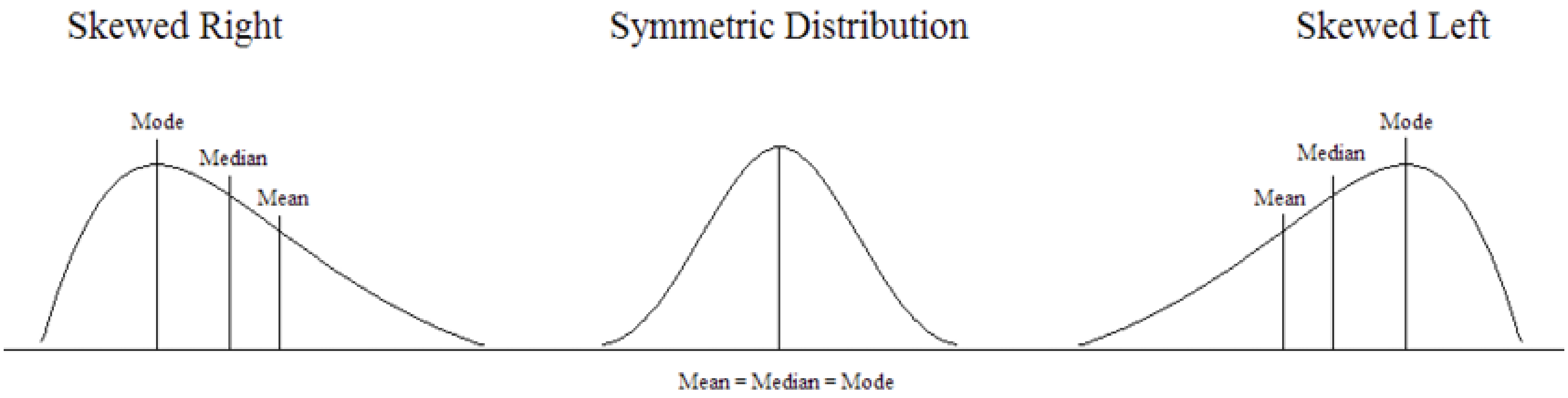
- In other words, the above data can be represented as:-

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**.

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.



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.

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

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

$$\mathbf{R = max(X) – min(X) = x_n - x_1}$$

= 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 **σ²** is defined as :-

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

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

The standard deviation, **σ**, 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*?

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

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

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

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

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 **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 **50th** 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 **50th** percentile.

Interquartile Range

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

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

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

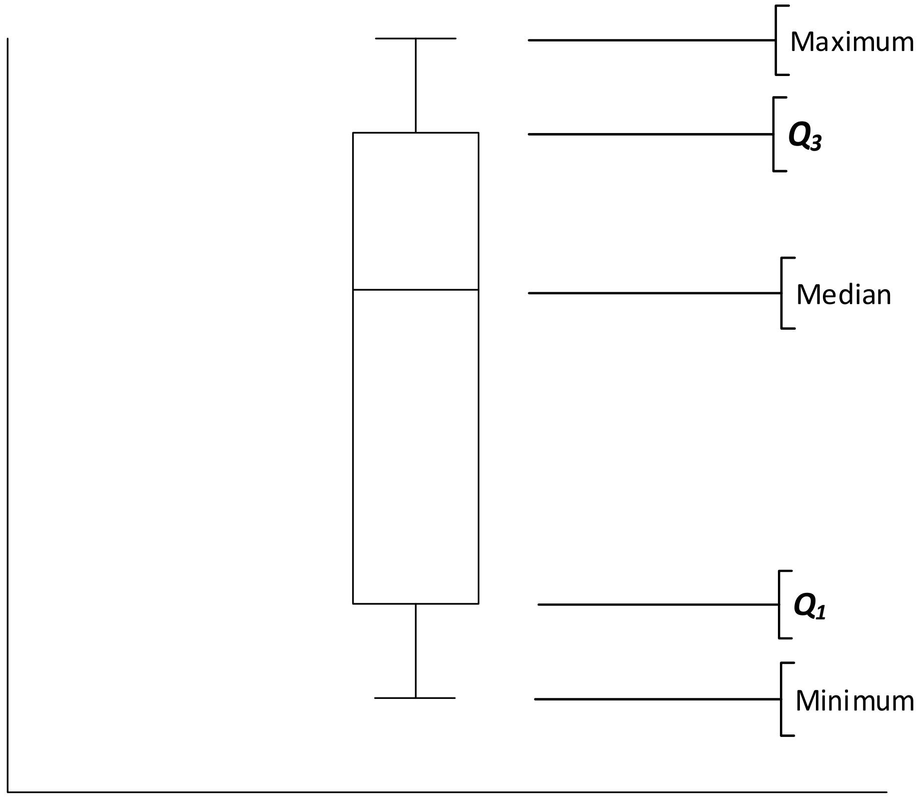
- **Five Number Summary**
  - Since, $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ together contain no information about the endpoints of the data, a complete summary of the shape of a distribution can be obtained by providing the lowest and highest data value as well. This is known as the five-number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

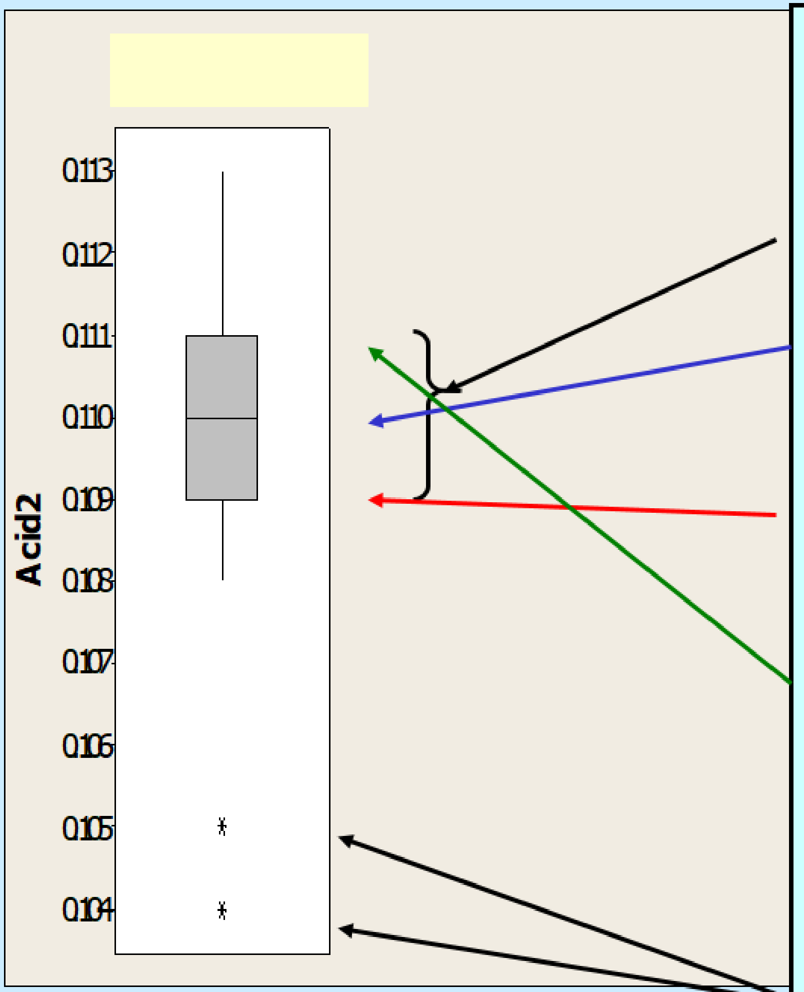
Minimum, $\mathbf{Q_1}$, Median ($\mathbf{Q_2}$), $\mathbf{Q_3}$, Maximum

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.

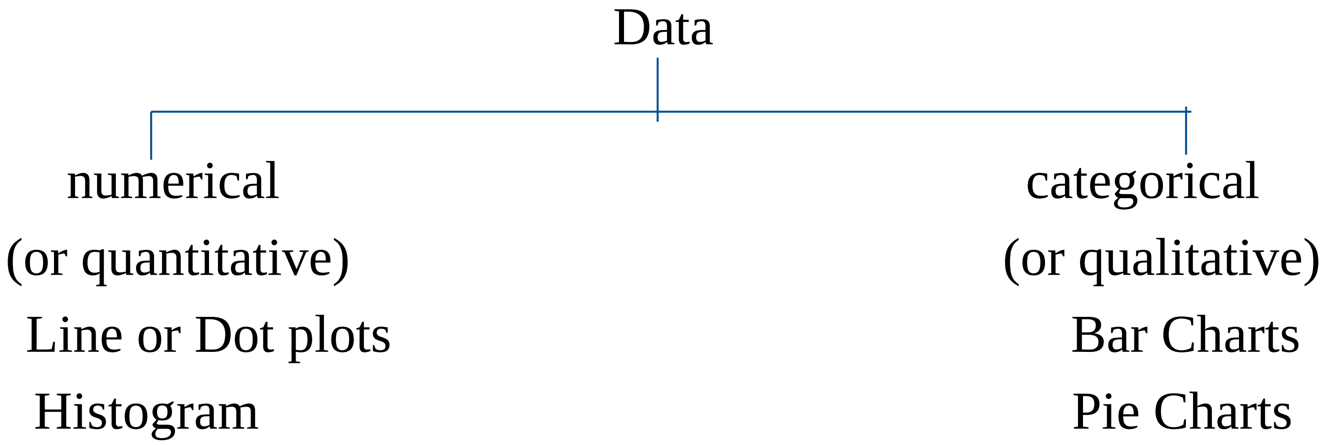
If you have data points outside this, they will show up as outliers.

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.

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.



Data

numerical

(or quantitative)

Line or Dot plots

Histogram

categorical

(or qualitative)

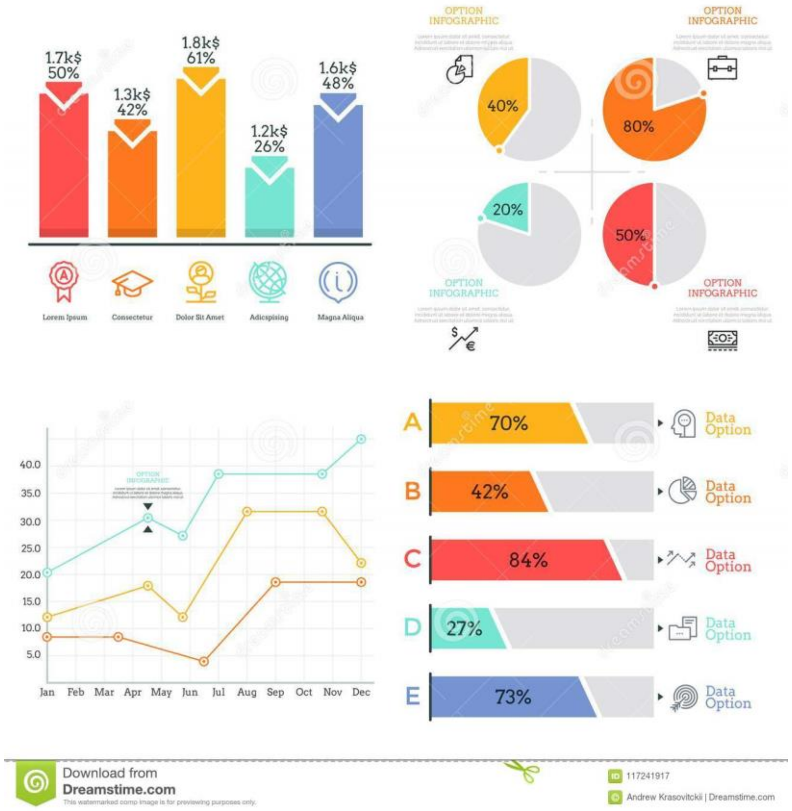
Bar Charts

Pie Charts

Frequency Distributions & Different Charts

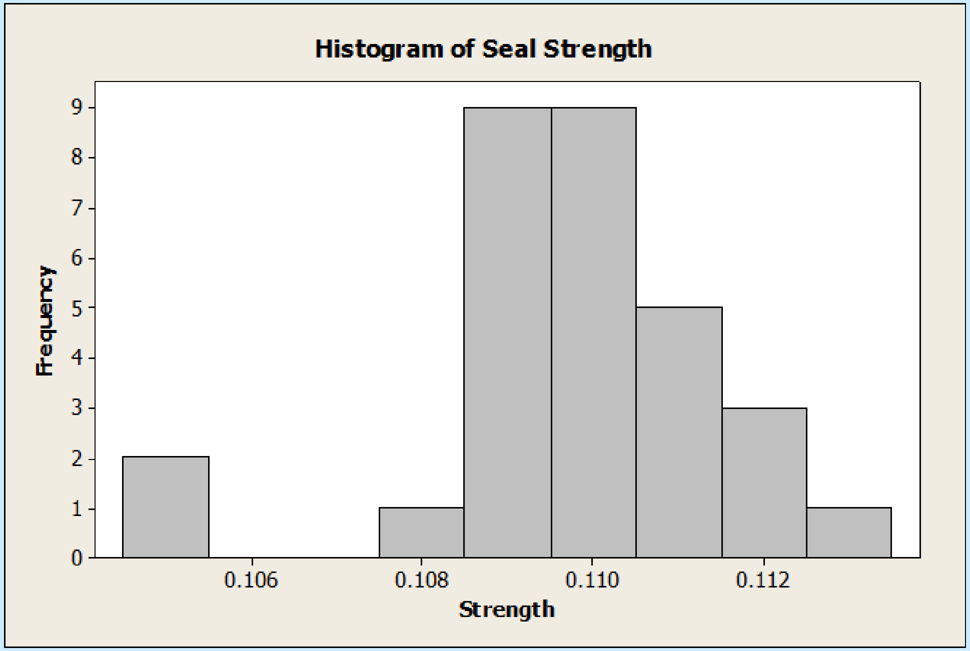
- When we deal with large sets of data, a good overall picture and sufficient information can be often conveyed by distributing the data into a number of classes or class intervals.

- 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:

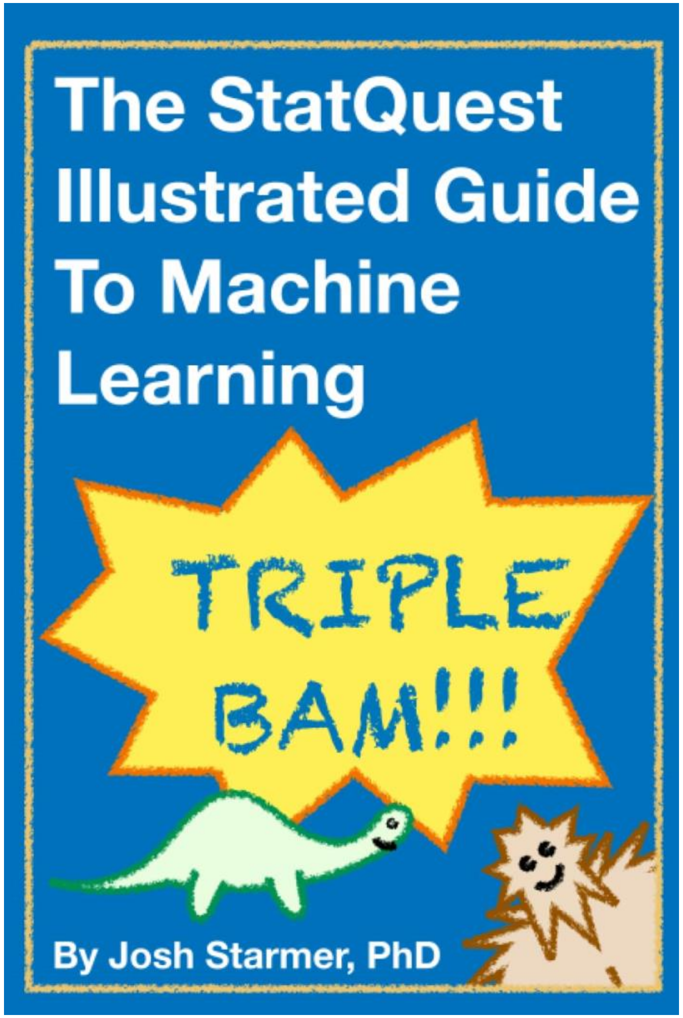
NO. OF DATA POINTS	NO. OF CLASSES (APPROX.)
50-100	5-10
100-250	7-12
250 & above	10-20

- 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.

References:



https://statquest.org/