# RESEARCH METHODOLOGY AND DATA ANALYSIS

**Workshop conducted by KSTA and VTU**


***Today's Topic***

***Statistics for Data Analysis***


**Dr. Tanujit Chakraborty**

*Assistant Professor of Statistics* @ Sorbonne University

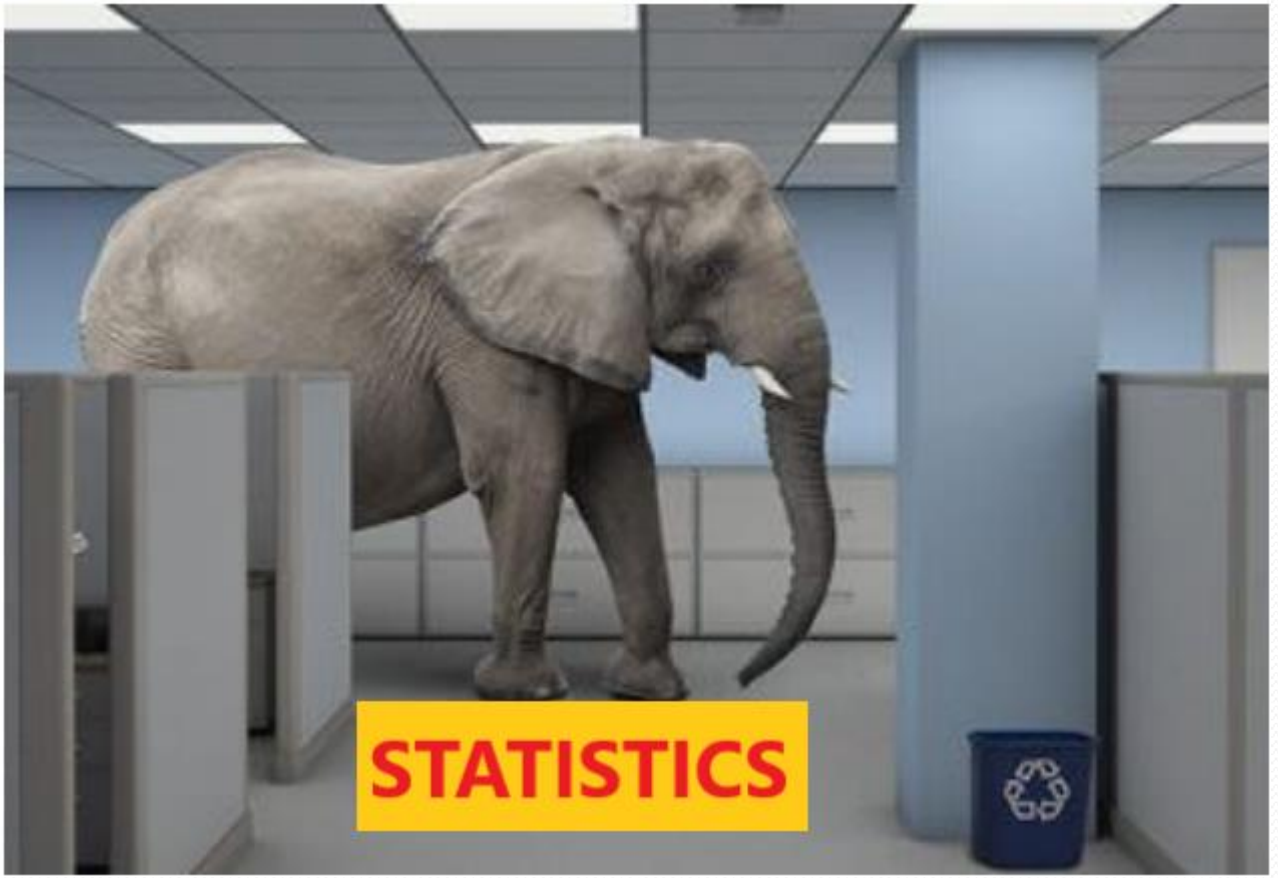
*Mphasis Research Fellow* @ IIIT Bangalore

GitHub: https://github.com/mad-stat/KSTA_Workshop

# Today's Topics…

- Role of Statistics and Data Analysis
- Data summarization
- Concepts of Descriptive Statistics
- Statistical Inference
  - Point Estimation
  - Interval Estimation
  - Hypothesis Testing
  - Analysis of Variance
- Relationship Analysis
  - Correlation Analysis
  - Linear Regression

# What (or Why) is Statistics?




*"Statistics is the universal tool of inductive inference, research in natural and social sciences, and technological applications. Statistics, therefore, must always have purpose, either in the pursuit of knowledge or in promotion of human welfare."*

– Prof. Prasanta Chandra Mahalanobis, *Father of Indian Statistics*

# 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.

# 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.

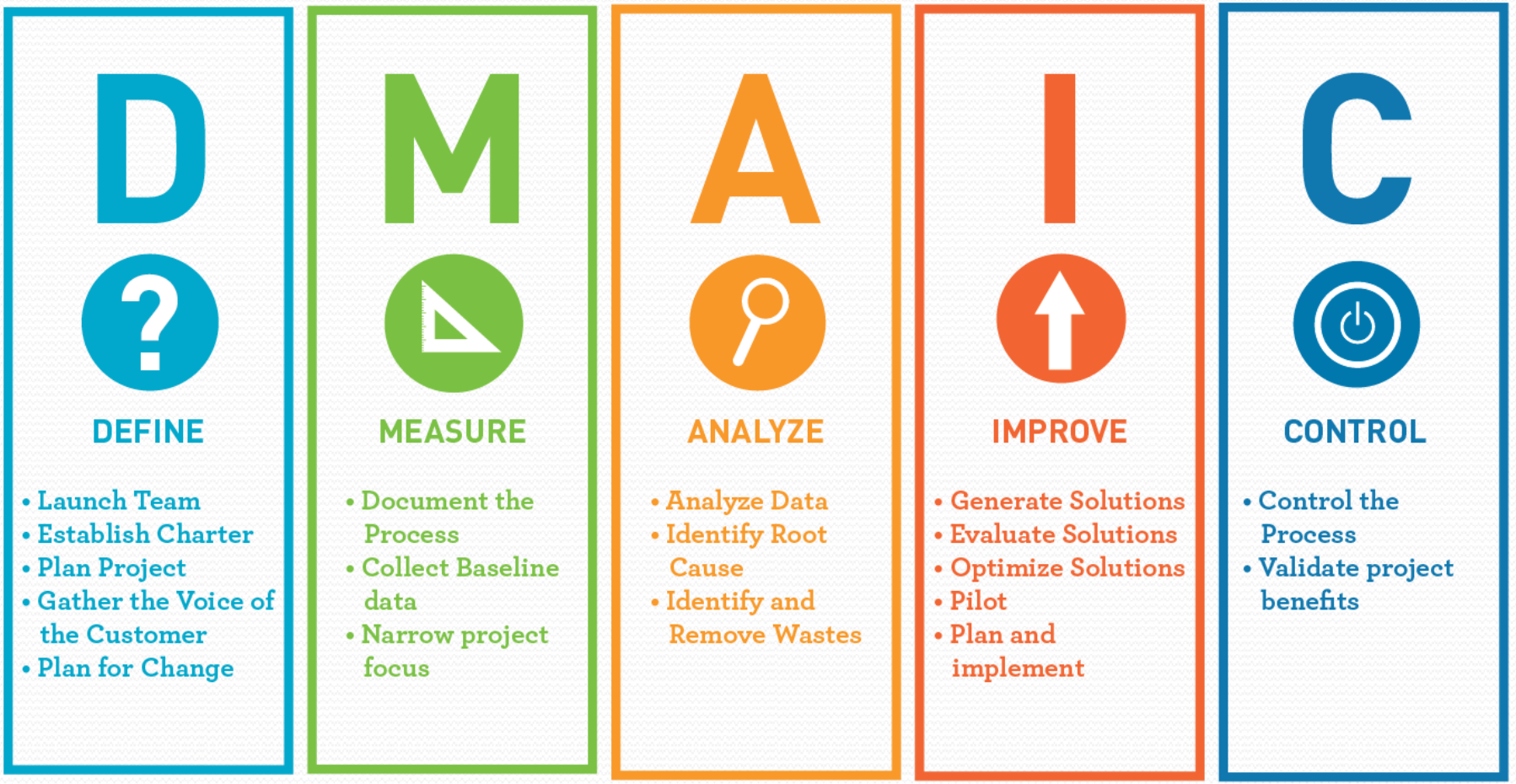
- Example of Variability:

  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.



# 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.

# 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

These can be measured in three ways

- Distributive measure
- Algebraic measure
- Holistic measure

# Simple mean / Arithmetic Mean / Average

Definition : **Simple mean**

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

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

Disadvantages of **Simple mean**

- 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.

# 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

Definition : **Median of a sample**

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

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

# 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 **$x_p$** is a value of **x** such that **p**% of the observed values of **x** are less than **$x_p$**
- Example: The **$50^{th}$** percentile is that value **$x_{50\%}$** such that **50%** of all values of **x** are less than **$x_{50\%}$**.

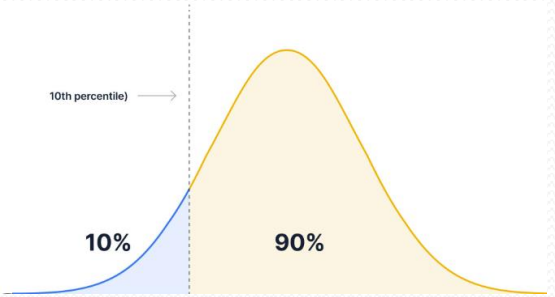
- **Note:** The median is the **$50^{th}$** percentile

# Quartile

The most commonly used percentiles are quartiles.

- The first quartile, denoted by **$Q_1$** is the **$25^{th}$** percentile.
- The third quartile, denoted by **$Q_3$** is the **$75^{th}$** percentile
- The median, **$Q_2$** is the **$50^{th}$** percentile.

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



# 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**.

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

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

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

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 $s$.

# Interquartile Range

- Interquartile range, denoted as IQR is a robust measure of dispersion.
- 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

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

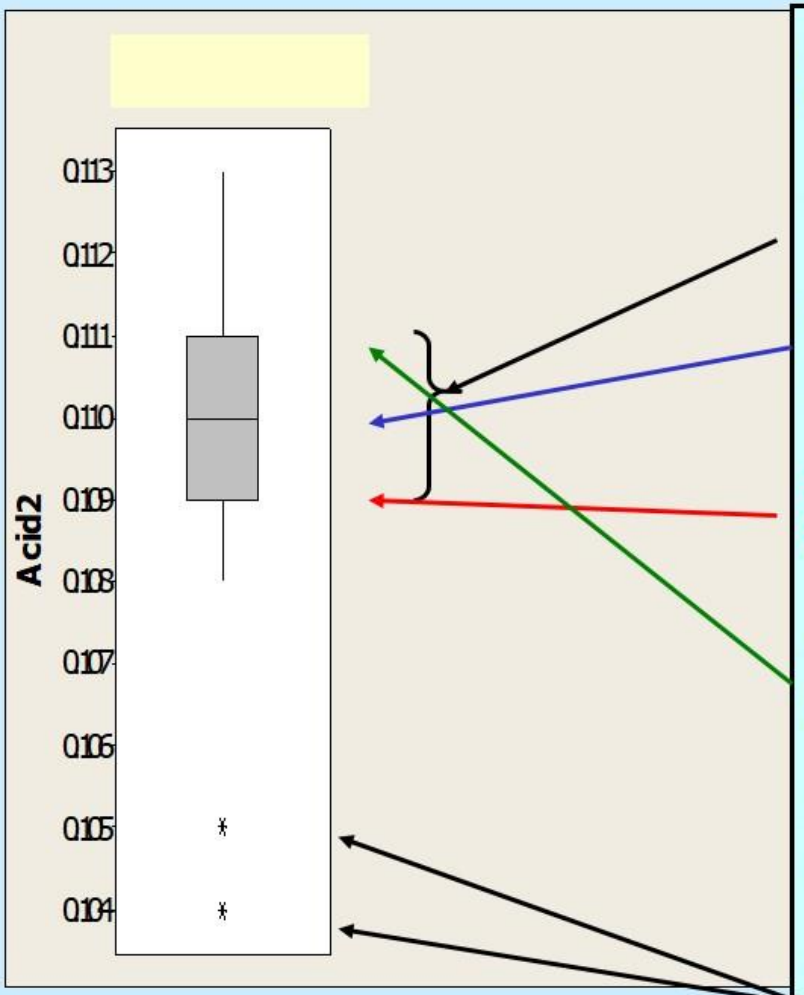
# Five point summary

- 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




**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.

# Statistical Inference

- Descriptive analysis is valid only for the data set under consideration and cannot necessarily be generalized to other data.
- Statistical Inference allows us to infer from the sample data about the population of interest.
- It is not feasible to consider entire population for a analysis, hence we need to collect a representative sample.
- There are two facts, which are key to statistical inference.

  Population parameters are fixed number whose values are usually unknown.

  Sample statistics are known values for any given sample, but vary from sample to sample, even taken from the same population.

- In fact, it is unlikely for any two samples drawn independently, producing identical values of sample **statistic**.

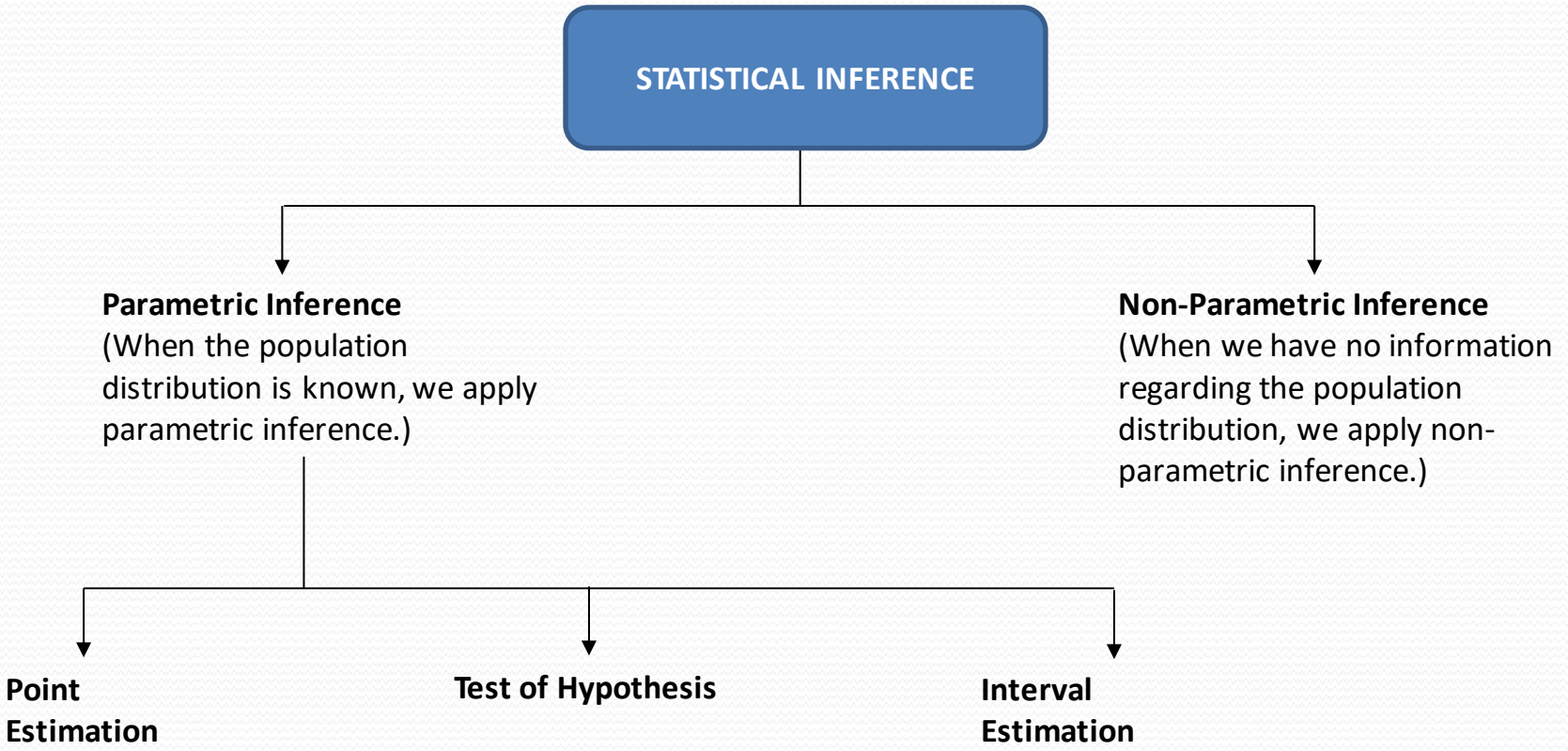



Fig.: Taxonomy of Statistical Inference

# Basic Inferential Approaches

**Approach 1: Point and Confidence interval measurement**

We estimate one (or more) parameter(s) using sample statistics.

- This estimation can be done in the form of a single estimated value (Point Estimation)
- This estimation usually done in the form of an interval (Interval Estimation).

Accuracy of the decision is expressed as the **level of confidence** we have in the interval.

**Approach 2: Hypothesis testing**

We conduct **test on hypothesis**.

- We hypothesize that one (or more) parameter(s) has (have) some specific value(s) or relationship.

Make our decision about the parameter(s) based on one (or more) sample statistic(s)

Accuracy of the decision is expressed as the probability that the decision is incorrect.

# Method of Point Estimatoion -Maximum Likelihood Estimation (MLE)

**Objective:**

To introduce the idea of working out the most likely cause of an observed result by considering the likelihood of each of several possible causes and picking the cause with the highest likelihood.

# Maximum Likelihood Estimation (MLE)

## Who is the daddy?

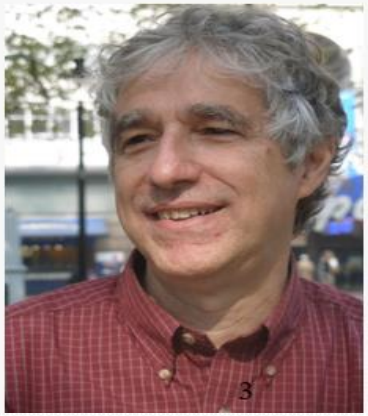
- Here is a picture of a boy and his mother…

# Maximum Likelihood Estimation (MLE)

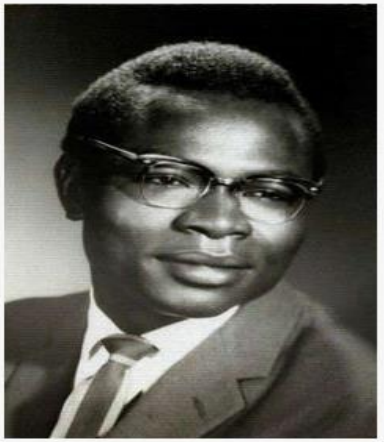
## Who's the daddy?

- Which of the following men, do you think, is the child's father?

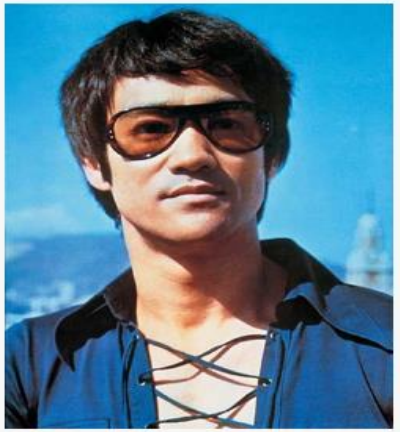
A



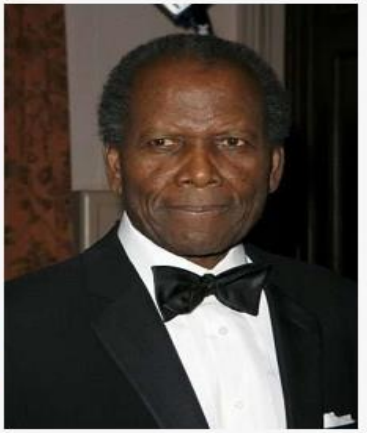
B



C



D



# Maximum Likelihood Estimation (MLE)

- The child appears to be of mixed race parentage.
- His mother is white.
- Therefore, of the four possible daddies, daddy A is the least likely.
- Daddy C is the next least likely because of the child's appearance.
- This leaves B or D; but which one did you go for?
- If only we had more information, e.g. blood types or DNA or history of the two men, we could be more exact in our view of which of the two is more likely....
- But we could never be certain that anyone's daddy is the real daddy! We simply accept the most likely choice. Every daddy is an MLE daddy!

# Did you guess…

- The child in the picture is President Barack Obama and his mother.
- Daddy D is the actor Sydney Poitier.
- C is the actor Bruce Lee.
- A is Professor Alan Agresti- Statistics Icon (writer of Foundation of Statistics for Data Scientists).
- B is Barack Obama, Sr.

# Maximum Likelihood Estimation (MLE)

Likelihood Function: Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a population with density function $f(x, \theta)$.

Then the likelihood function of the sample values $x_1, x_2, \dots, x_n$ usually denoted by $L = L(\theta)$ is their joint density function, given by:

$$L = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

For a given sample $x_1, x_2, \dots, x_n$, $L$ becomes a function of the variable $\theta$, the parameter.

MLE Principle: The principle of ML consists in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, say which maximizes the likelihood function $L(\theta)$ for variations in parameter i.e. we wish to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ so that

$$L(\hat{\theta}) > L(\theta) \quad \forall \ \theta \in \Theta.$$

$\hat{\theta}$ is called Maximum Likelihood Estimator (MLE).

# Interval Estimation

A point estimate on its own does not take into account the accuracy of the estimate.

The deviation between the point estimate and the true parameter (i.e. $|\bar{x} - \mu|$) can be considerable, especially when the sample size is small.

To incorporate the information about the accuracy of an estimate in the estimated value, a **confidence interval** can be constructed.

It is a **random interval** with **lower** and **upper bounds,** $\boldsymbol{I_l(X)}$ and $\boldsymbol{I_u(X)}$, such that the unknown parameter $\theta$ is covered by a prespecified probability of at least $1 - \alpha$:

$$\boldsymbol{P_\theta\left(I_l(X) \leq \theta \leq I_u(X)\right) \geq 1 - \alpha.}$$

The probability $1 - \alpha$ is called the **confidence level.**

$\boldsymbol{I_l(X)}$ is called the **lower confidence limit** and $\boldsymbol{I_u(X)}$ is called the **upper confidence limit.**

Note that the **bounds** are **random** and the **parameter** is a **fixed value**, i.e. the true parameter is covered by the interval with probability $\boldsymbol{1 - \alpha}$.

# Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

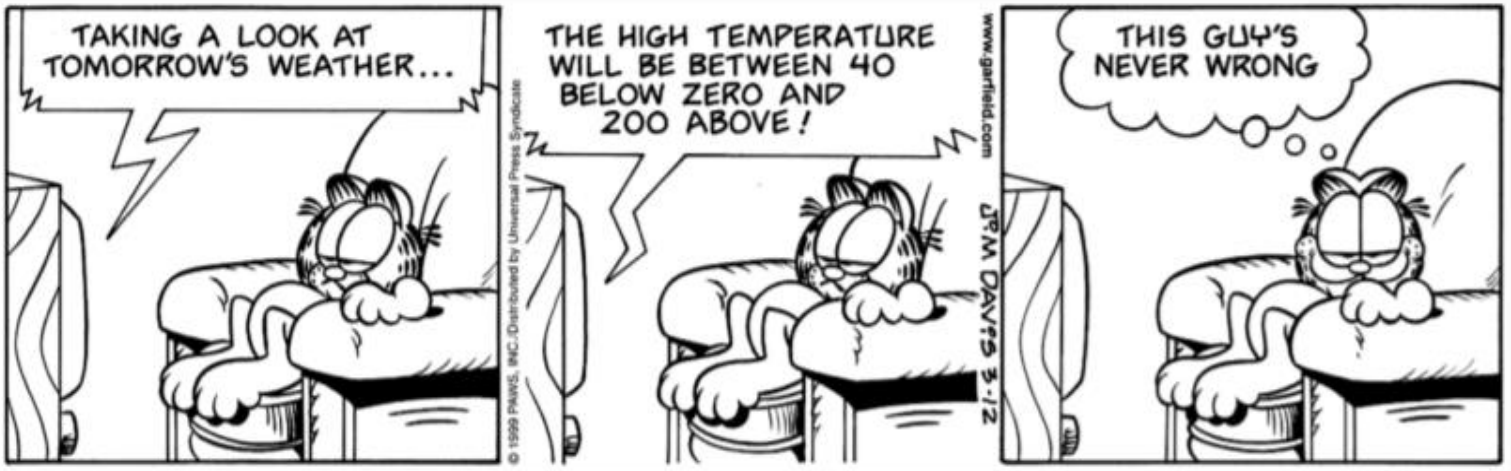
- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

# Width of an Interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

*A wider interval.*

Can you see any drawbacks to using a wider interval?




*If the interval is too wide it may not be very informative.*

# Introduction to Hypothesis Testing

Sometimes, we are not interested in precise estimates of a parameter, but we only want to examine whether a statement about a parameter of interest or the research hypothesis is true or not.

Another related issue is that once an analyst estimates the parameters on the basis of a random sample, he would like to infer something about the value of the parameter in the population.

Statistical hypothesis tests facilitates the comparison of estimated values with hypothetical values.

# Hypothesis Testing

Statistical inference

Null hypothesis

Sample

Alternative hypothesis

A **hypothesis** about the value of a population parameter is an **assertion** about its value.

# Hypothesis Testing

A useful analogy for statistical decision making is the decision process used by a jury in a legal proceeding.

In a criminal trial, the null hypothesis is that the defendant is "not guilty" ( "not" in the null hypothesis).

The alternative is, of course, that the defendant is guilty. The "benefit of the doubt" goes to the null hypothesis (i.e., the evidence must be evaluated by beginning with the assumption that the defendant is not guilty).

The "burden of proof" is on the prosecution to demonstrate that the assumption of not guilty is incorrect (i.e., evidence must be gathered to support the claim of guilty. The defendant need not prove innocence).

Jurors must only vote for guilty if they are convinced "beyond reasonable doubt" that the defendant is guilty. Basically, the decision process for a juror goes something like this.

# Hypothesis Testing

Assume the defendant is not guilty. Evaluate the evidence in light of this assumption. Could all the evidence presented simply be a coincidence? (That is, could the degree that the evidence points to guilt of the defendant just be explained as coincidence?)

What is the chance of that? If that chance is small, then vote for guilty. Otherwise don't vote for guilty. How small is a small chance? Whatever the juror interprets as "beyond reasonable doubt."

In statistical hypothesis testing, the principles of the trial carry through quite similarly. The null hypothesis gets the "benefit of the doubt."

The "burden of proof" is on the alternative hypothesis. The experimenter must have a test statistic (evidence) which convinces us that the alternative is true rather than the null hypothesis.

The same question must be asked as we make a statistical decision as that asked above by the juror.

# Statistical Hypothesis

If the hypothesis is stated in terms of population parameters (such as mean and variance), the hypothesis is called statistical hypothesis.

Data from a sample (which may be an experiment) are used to test the validity of the hypothesis.

A procedure that enables us to agree (or disagree) with the statistical hypothesis is called a **test of the hypothesis.**

**Example :**

1. To determine whether the wages of men and women are equal.
2. A product in the market is of standard quality.
3. Whether a particular medicine is effective to cure a disease.

# The Hypotheses

The main purpose of statistical hypothesis testing is to choose between two competing hypotheses.

**Example :** One hypothesis might claim that wages of men and women are equal, while the **alternative** might claim that men make more than women.

Hypothesis testing start by making a set of two statements about the parameter(s) in question.

The hypothesis actually to be tested is usually given the symbol $H_0$ and is commonly referred as the **null hypothesis.**

The other hypothesis, which is assumed to be true when null hypothesis is false, is referred as the **alternate hypothesis** and is often symbolized by $H_1$

The two hypotheses are exclusive and exhaustive.

# The Hypotheses

**Example:**

Ministry of Human Resource Development (MHRD), Government of India takes an initiative to improve the country's human resources.

To measure the engineering aptitudes of graduates, MHRD conducts GATE examination for a mark of 1000 in every year. A sample of 300 students who gave GATE examination in 2020 were collected and the mean is observed as 220.

In this context, statistical hypothesis testing is to determine the mean mark of the all GATE-2020 examinee.

The two hypotheses in this context are:

$$H_0: \mu = 220$$

$$H_1: \mu < 220$$

# The Hypotheses

**Note:**

1. As null hypothesis, we could choose $H_0: \mu \leq 220$ or $H_0: \mu \geq 220$
2. It is customary to always have the null hypothesis with an equal sign.
3. As an alternative hypothesis there are many options available with us.

**Examples:**

I. $H_1: \mu > 220$

*II.* $H_1: \mu < 220$

*III.* $H_1: \mu \neq 220$

4. The two hypothesis should be chosen in such a way that they are **exclusive** and **exhaustive**.

   One or other must be true, but they cannot both be true.

# The Hypotheses

**One-tailed test**

A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in $H_0$ is called a one-sided (or one-tailed) test.

Example: $H_0 : \mu = 100$ $H_1 : \mu > 100$

**Two-tailed test**

An alternative hypothesis that specifies that the parameter can lie on their sides of the value specified by $H_0$ is called a two-sided (or two-tailed) test.

Example: $H_0 : \mu = 100$ $H_1 : \mu <> 100$

# The Hypotheses

**Simple and Composite Hypothesis:**

A statistical hypothesis is some statement about a population, which we want to verify on the basis of information available from a sample.
If the statistical hypothesis specifies the population completely then it is termed as a simple hypothesis otherwise it is called composite hypothesis.
Example: If $X_1, X_2, \ldots, X_n$ is a random sample of size n from $N(\mu, \sigma^2)$ population, then the hypothesis

$$H_0: \quad \mu = \mu_0, \qquad \sigma^2 = \sigma_0^2$$

Example of Composite Hypothesis:

*i.* $\mu = \mu_0$
*ii.* $\mu < \mu_0, \quad \sigma^2 = \sigma_0^2$
*iii.* $\mu < \mu_0, \quad \sigma^2 > \sigma_0^2$ etc.

# Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

**Type I error :** A type I error occurs when we incorrectly reject $H_0$ (i.e., we reject the null hypothesis, when $H_0$ is true).

$\boldsymbol{\alpha}$**:** denotes the probability of making a Type I error i.e., $\boldsymbol{\alpha} = \mathbf{P}(\text{Rejecting } H_0 | H_0 \text{ is true})$

**Type II error:** A type II error occurs when we incorrectly fail to reject $H_0$ (i.e., we accept $H_0$ when it is not true).

$\boldsymbol{\beta}$**:** denotes the probability of making a Type II error i.e., $\boldsymbol{\beta} = \mathbf{P}(\text{Accepting } H_0 | H_0 \text{ is false})$

**Note:**

- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not independent of each other as one increases, the other decreases
- When the sample size increases, both to decrease since sampling error is reduced.
- In general, we focus on Type I error, but Type II error is also important, particularly when sample size is small.
- So the test statistics are obtained by fixing $\alpha$ and then minimizing $\beta$.
- The probability $1 - \boldsymbol{\beta} = \mathbf{P}(\text{H}_0 \text{ is rejected} | \text{H}_0 \text{ is Not True})$ is called the power of the test.

| Decision | Observation | |
|---|---|---|
| | $H_0$ is true | $H_0$ is false |
| $H_0$ is accepted | Decision is correct | Type II error |
| $H_0$ is rejected | Type I error | Decision is correct |

# Hypothesis Testing Strategies

The hypothesis testing determines the validity of an assumption (technically described as null hypothesis), with a view to choose between two conflicting hypothesis about the value of a population parameter.

There are two types of tests of hypotheses

- ✓ Non-parametric tests (also called distribution-free test of hypotheses)
- ✓ Parametric tests (also called standard test of hypotheses).

# Hypothesis Tests…

A hypothesis test is a form of statistical inference where we attempt to answer a specific question about the distribution of some measurement in the population.

Basically, we want to know "Yes or No, does the following statement hold true for the distribution of our measurement in our population?" (Most often, our hypothesis will be about one or two population means or one or two population proportions).

When we make such a decision, we must keep in mind that we could make an incorrect decision. Hence, we want to design our decision procedures so as to minimize the chances of errors. We will formally see how this is done in the context of statistical decisions below.

We will restrict our attention to the case of two alternatives. The first will be called the and will be represented by the symbol $H_0$ (**null hypothesis).** The other will be called the **alternative hypothesis** and will be represented by the symbol $H_a$.

# Hypothesis Tests…

The alternative hypothesis is usually some positive statement about the experiment being analyzed and is often called the "research hypothesis" because it is what the researcher is trying to show in the experiment.

The null hypothesis is often a representation of the status quo or other representation of lack of positive results in the experiment. Hence the aspect of "null" about the null hypothesis.

Example: In an analgesic drug experiment the alternate hypothesis might be that the new drug relieves pain faster than the old drug. The null hypothesis would then be that the new drug does not relieve pain faster.

Example: In a teaching experiment, the alternative hypothesis might be that the new method results in higher scores than the old method. The null hypothesis would be that this is not so.

# How to Conduct a Statistical Test

- Define the distributional assumption for the random variables of interest, and specify them in terms of population parameters.
- Formulate $H_0$ and $H_1$.
- Fix a significance value (Type I error) $\alpha$ $(say\ 0.05)$.
- Construct a test statistic $T(\boldsymbol{X}) = T(X_1, X_2, \dots, X_n)$. The distribution of $T$ has to be known under the null hypothesis $H_0$.
- Construct a critical region $W$ for the statistic $T$, i.e. a region where – if $T$ falls in this region- $H_0$ is rejected, such that

$$P_{H_0}(T(\boldsymbol{X}) \in W) \leq \alpha.$$

- Calculate $t(x) = T(x_1, x_2, \dots, x_n)$ based on the realized sample values

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n.$$

- Decision Rule:
  i. $t(x) \in W$: $H_0\ is\ rejected \Rightarrow\ H_1\ is\ statistically\ significant,$
  ii. $t(x) \notin W$: $H_0\ is\ not\ rejected$

# P-values…

P-value is a measure of consistency between the null hypothesis and the observed data.

Hence, we will be more inclined to believe the alternative hypothesis when the P-value is small and less likely to believe the alternative hypothesis when the P-value is not so small.

We make a rule for statistical decisions as follows. We establish a cut-off value called the level of significance.

Then we reject the null hypothesis in favor of the alternative hypothesis when the P-value is less than the level of significance.

If the P-value is not less than the level of significance, then we fail to reject the null hypothesis in favor of the alternative hypothesis.

# Test Decision Using the p-value

It is possible to use the p-value instead of critical regions for making test decisions.

The p-value of the test statistic $T(\boldsymbol{X})$ is defined as follows:

Right-sided case: $P(T \geq t(x) \mid H_0 \; is \; true) = p - value$
Left-sided case: $P(T \leq t(x) \mid H_0 \; is \; true) = p - value$

Two sided case:

$$2\min\{P(T \leq t(x)|H_0 \; is \; true), P(T \geq t(x)|H_0 \; is \; true)\} = p \; value$$

- If $p - value < 0.01$ : very strong evidence against $H_0$, i.e., very strongly Reject $H_0$.
- If $0.01 < p - value < 0.05$ : strong evidence against $H_0$, i.e., strongly Reject $H_0$.
- If $0.05 < p - value < 0.10$ : some weak evidence against $H_0$, i.e., Not Reject $H_0$.
- If $p - value > 0.10$ : no evidence against $H_0$, i.e. Not Reject $H_0$.

# Parametric Tests

**Important Parametric Tests**

The widely used sampling distribution for parametric tests are

$$Z - test$$
$$t - test$$
$$\chi^2 - test$$
$$F - test$$

**Note:**

All these tests are based on the assumption of normality (i.e., the source of data is considered to be normally distributed).

# Parametric Tests : Applications

Usually assume certain properties of the population from which we draw samples.

- Observation come from a normal population
- Sample size is small
- Population parameters like mean, variance, etc. are hold good.
- Requires measurement equivalent to interval scaled data.

# Parametric Tests : Z-test

***Z – test*: This is most frequently test in statistical analysis.**

It is based on the normal probability distribution.

Used for judging the significance of several statistical measures particularly the mean.

It is used even when *binomial distribution* or $t -$ *distribution* is applicable with a condition that such a distribution tends to normal distribution when $n$ becomes large.

Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when **population variance** is known.

# Single Sample: Test Concerning a Single Mean

- Null Hypothesis: $H_0: \mu = \mu_0$ (Variance known)
- Value of Test Statistic: $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\mu < \mu_0$ | $z < -z_\alpha$ |
| $\mu > \mu_0$ | $z > z_\alpha$ |
| $\mu \neq \mu_0$ | $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ |

# Parametric Tests : t-test

***t – test*:** It is based on the t-distribution.

It is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of

small sample(s)

**population variance is not known** (in this case, we use the variance of the sample as an estimate of the population variance)

# Single Sample: Test Concerning a Single Mean

- Null Hypothesis: $H_0: \mu = \mu_0$ (Variance unknown)
- Value of Test Statistic: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\mu < \mu_0$ | $t < -t_{\alpha,n-1}$ |
| $\mu > \mu_0$ | $t > t_{\alpha,n-1}$ |
| $\mu \neq \mu_0$ | $t < -t_{\frac{\alpha}{2},n-1}$ or $t > t_{\frac{\alpha}{2},n-1}$ |

# Two Sample: Test on Two Mean

- Null Hypothesis: $H_0: \ \mu_1 - \mu_2 = d_0$ ($\sigma_1$ and $\sigma_2$ known)
- Value of Test Statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$
- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\mu_1 - \mu_2 < d_0$ | $z < -z_\alpha$ |
| $\mu_1 - \mu_2 > d_0$ | $z > z_\alpha$ |
| $\mu_1 - \mu_2 \neq d_0$ | $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ |

# Two Sample: Test on Two Mean

- Null Hypothesis: $H_0$: $\mu_1 - \mu_2 = d_0$ ($\sigma_1 = \sigma_2$ but unknown)
- Value of Test Statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\mu_1 - \mu_2 < d_0$ | $t < -t_{\alpha,n_1+n_2-2}$ |
| $\mu_1 - \mu_2 > d_0$ | $t > t_{\alpha,n_1+n_2-2}$ |
| $\mu_1 - \mu_2 \neq d_0$ | $t < -t_{\frac{\alpha}{2},n_1+n_2-2}$ or $t > t_{\frac{\alpha}{2},n_1+n_2-2}$ |

# Two Sample: Test on Two Mean

Null Hypothesis: $H_0$: $\mu_1 - \mu_2 = d_0$ ($\sigma_1 \neq \sigma_2$ but unknown)

- Value of Test Statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t_\nu$
- Degrees of Freedom: $\nu = \dfrac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$
- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\mu_1 - \mu_2 < d_0$ | $t < -t_{\alpha,\nu}$ |
| $\mu_1 - \mu_2 > d_0$ | $t > t_{\alpha,\nu}$ |
| $\mu_1 - \mu_2 \neq d_0$ | $t < -t_{\frac{\alpha}{2},\nu}$ or $t > t_{\frac{\alpha}{2},\nu}$ |

# Parametric Tests : $\chi^2$ -test

$\chi^2 - test$: It is based on Chi-squared distribution.

It is used for comparing a sample variance to a theoretical population variance.

# Parametric Tests : $F$ -test

$F - test$: It is based on F-distribution.

- It is used to compare the variance of two independent samples.
- This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means.

# Single Sample: Test Concerning Variance

- Null Hypothesis: $H_0: \sigma^2 = \sigma_0^2$
- Value of Test Statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$
- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\sigma^2 < \sigma_0^2$ | $\chi^2 < \chi^2_{1-\alpha,n-1}$ |
| $\sigma^2 > \sigma_0^2$ | $\chi^2 > \chi^2_{\alpha,n-1}$ |
| $\sigma^2 \neq \sigma_0^2$ | $\chi^2 < \chi^2_{1-\frac{\alpha}{2},n-1}$ $or$ $\chi^2 > \chi^2_{\frac{\alpha}{2},n-1}$ |

# Two Sample: Test Concerning Variance

- Null Hypothesis: $H_0: \ \sigma_1^2 = \sigma_2^2$
- Value of Test Statistic: $\mathrm{F} = \frac{s_1^2}{s_2^2} \sim F_{v_1, v_2}$, where $v_1 = n_1 - 1, \ v_2 = n_2 - 1$
- Significance Level: $\alpha$

| Alternative Hypothesis $H_1$ | Critical Region |
|---|---|
| $\sigma_1^2 < \sigma_2^2$ | $F < F_{1-\alpha}(v_1, v_2)$ |
| $\sigma_1^2 > \sigma_2^2$ | $F > F_{\alpha}(v_1, v_2)$ |
| $\sigma_1^2 \neq \sigma_2^2$ | $F < F_{1-\frac{\alpha}{2}}(v_1, v_2) \quad or \quad F > F_{\alpha/2}(v_1, v_2)$ |

# Case Study 1: **Coffee Sale**

A coffee vendor nearby Yesvantpur Junction Railway Station has been having average sales of 500 cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days, after the inauguration of the bus stand, the daily sales were as under:

550 570 490 615 505 580 570 460 600 580 530 526

On the basis of this sample information, can we conclude that the sales of coffee have increased?

Consider 5% level of confidence.

# Hypothesis Testing : 5 Steps

The following five steps are followed when testing hypothesis

1. Specify $H_0$ and $H_1$, the null and alternate hypothesis, and an **acceptable level of $\alpha$.**
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified $H_0$.
3. Collect the sample data and calculate the test statistic.
4. Make a decision to either reject or fail to reject $H_0$.
5. Interpret the result in common language suitable for practitioner.

# Case Study 1: **Step 1**

**Step 1: Specification of hypothesis and acceptable level of α**

Let us consider the hypotheses for the given problem as follows.

$H_0: \mu = 500$ cups per day

The null hypothesis that sales average 500 cups per day and they have not increased.

$H_1: \mu > 500$

The alternative hypothesis is that the sales have increased.

Given the acceptance level of $\alpha = 0.05$ $(i.e., 5\%\ level\ of\ significance)$

# Case Study 1: **Step 2**

**Step 2: Sample-based test statistics and the rejection region for specified $H_0$**

Given the sample as

550 570 490 615 505 580 570 460 580 530 526

Since the sample size is small and the population standard deviation is not known, we shall use $t-test$ assuming normal population. The test statistics $t$ is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

To find $\bar{X}$ and $S$, we make the following computations.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{6576}{12} = 548$$

# Case Study 1: **Step 2**

| *Sample #* | $X_i$ | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ |
|---|---|---|---|
| 1 | 550 | 2 | 4 |
| 2 | 570 | 22 | 484 |
| 3 | 490 | −58 | 3364 |
| 4 | 615 | 67 | 4489 |
| 5 | 505 | −43 | 1849 |
| 6 | 580 | 32 | 1024 |
| 7 | 570 | 22 | 484 |
| 8 | 460 | −88 | 7744 |
| 9 | 600 | 52 | 2704 |
| 10 | 580 | 32 | 1024 |
| 11 | 530 | −18 | 324 |
| 12 | 526 | −22 | 484 |
| $n = 12$ | $\sum X_i = 6576$ | | $\sum (X_i - \bar{X})^2 = 23978$ |

# Case Study 1: **Step 2**

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

Hence, $t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{48}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$

**Note:**

Statistical table for t-distributions gives a *t*-value given *n*, the degrees of freedom and α, the level of significance and vice-versa.

# Case Study 1: **Step 3**

**Step 3: Collect the sample data and calculate the test statistics**

$$Degree\ of\ freedom = n - 1 = 12 - 1 = 11$$

As $H_1$ is one-tailed, we shall determine the rejection region applying one-tailed in the right tail because $H_1$ is more than type ) at 5% level of significance.

Using table of $t - distribution$ for 11 degrees of freedom and with 5% level of significance,

$$R: t > 1.796$$

# Case Study 1: **Step 4**

**Step 4: Make a decision to either reject or fail to reject $H_0$**

The observed value of $t = 3.558$ which is in the rejection region and thus $H_0$ is rejected at 5% level of significance.

# Case Study 1: **Step 5**

**Step 5: Final comment and interpret the result**

We can conclude that the sample data indicate that coffee sales have increased.

# Hypothesis Testing : Non-Parametric Test

***Non-Parametric tests***

- ✓Does not under any assumption
- ✓Assumes only nominal or ordinal data

**Note:** Non-parametric tests need entire population (or very large sample size)

# Analysis of Variance (ANOVA)

- ANOVA is a statistical technique
  - It is similar in application to techniques such as t-test, Z-test and $\chi^2$-test in that it is used to compare means and the relative variance between them.
- Why not use t-test, Z-test and $\chi^2$-test ?
- Why analysis of variance for comparing means?

t-test is used to:

- To infer mean of a single population
- t-test can be used to compare two populations

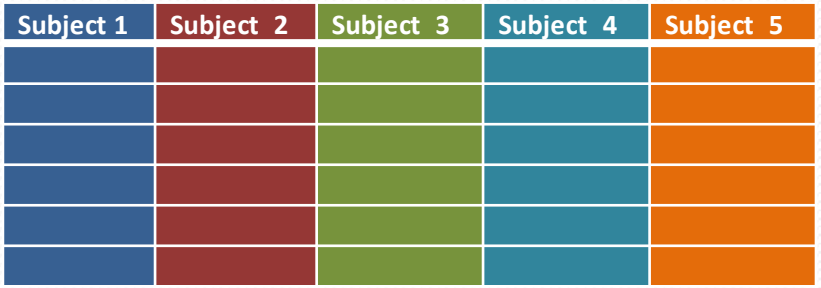
*However, t-test is not useful to compare mean of more than two populations.*

# Extending the two population procedure

- Construct pairwise comparison on all means.
- For 5 populations →10 possible pairs.
- Considering $\alpha = 0.05$, probability of correctly failing to reject the null hypothesis for all 10 tests is $(0.95)^{10}$, assuming that the tests are independent
- Thus the true value of α for this set of comparison is 0.4, instead of .05
- It inflates the Type I error.

# Extending the two population procedure

- Statistical Inference I
  - A teacher is interested in a comparison of the average percentage marks obtained in the examinations of five different subjects and has available the marks of eight students who all completed each examination.




  - What is the number of populations?
  - How many samples? What are there sizes? Are each samples independent to each other?

# Some Terminologies

Factor

A characteristic under consideration, thought to influence the measured observations

Level (also called treatment)

A value of the factor

Typical data for a **Single-Factor** Experiment

| Level | Observations | | | | Total | Mean |
|---|---|---|---|---|---|---|
| 1 | $y_{11}$ | $y_{12}$ | … | $y_{1n1}$ | | |
| 2 | $y_{21}$ | $y_{22}$ | … | $y_{2n2}$ | | |
| … | … | … | … | … | | |
| … | … | … | … | … | | |
| … | … | … | … | … | | |
| k | $y_{k1}$ | $y_{k2}$ | … | $y_{knk}$ | | |

# One-way ANOVA

The purpose of the procedure is to compare sample means of $k$ populations.
In general, One-way ANOVA technique can be used to study the effect of $k$ $(>\ 2)$ levels of a single factor.

To determine if different levels of the factor affect measured observations differently, the following hypotheses are tested.

$$H_0\colon\ \mu_i = \mu \quad \text{all} \quad i = 1,2,\dots,k$$
$$H_1\colon\ \mu_i \neq \mu \quad \text{some} \quad i = 1,2,\dots,k$$

That is, at least one equality is not satisfied

where $\mu_i$ is the population mean for a level $i$.

# Assumptions

When applying one-way analysis of variance, there are three key assumptions that should be satisfied as follows.

1. The observations are obtained independently and randomly from the populations defined by the factor levels.
2. The population at each factor level is (approximately) normally distributed.
3. These normal populations have a common variance, $\sigma^2$.

Thus, for factor level $i$, the population is assumed to have a distribution which is $N(\mu_i, \sigma^2)$.

# Heuristic Justification of ANOVA

If the null hypothesis is true, that is, each of the $\mu_i$ has the same value, say, $\mu$, then the distribution of each of the *k* sample means, $\overline{y_{i.}}$ will have mean $\mu$ and variance $\sigma^2/n$ .

It then follows that, if we calculate a variance using the sample means as observations,

$$\hat{\sigma}_B^2 = \sum(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)$$

Then the quantity is an estimate of $\sigma^2/n$ .

Hence, $n\hat{\sigma}_B^2$ is an estimate of $\sigma^2$.

This estimate has *k*-1 degree of freedom and is independent of the pooled estimate of $\sigma^2$.

# Heuristic Justification of ANOVA

Out of several sampling distributions, the F-distribution describes the ratio of two independent estimates of a common variance.

The parameters of the distribution are the degrees of freedom of the numerator and denominator variances, respectively.

If the null hypothesis of equal mean is true, then we can compute the two estimates of $\sigma^2$ namely

$\hat{\sigma}_B^2 = \sum(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)$ and $s_p^2$, the pooled variance.

Therefore, the ratio $\frac{n\hat{\sigma}_B^2}{s_p^2}$ has the F-distribution with degrees of freedom ($k$-1) and $n-k$.

# Heuristic Justification of ANOVA

Thus, the procedure for testing the hypothesis.

$$H_0: \mu_i = \mu \quad \text{all} \quad i = 1, 2, \ldots, k$$
$$H_1: \text{at least one equality is not satisfied}$$

We are to reject $H_0$, if the calculated value of $F = \frac{n\hat{\sigma}_B^2}{s_p^2}$ exceeds α (confidence level) of the F-distributions with ($k$-1) and $n - k$ degrees of freedom.

# One-way ANOVA

| Level | Observations | | | | Total | Average |
|---|---|---|---|---|---|---|
| 1 | $y_{11}$ | $y_{12}$ | …… | $y_{1n}$ | $y_{1.}$ | $\overline{y_{1.}}$ |
| 2 | $y_{21}$ | $y_{22}$ | …… | $y_{2n}$ | $y_{2.}$ | $\overline{y_{2.}}$ |
| . | . | . | …… | . | . | …… |
| . | . | . | …… | . | . | …… |
| . | . | . | …… | . | . | …… |
| $k$ | $y_{k1}$ | $y_{k2}$ | | $y_{kn}$ | $y_{k.}$ | $\overline{y_{k.}}$ |
| | | | | | $y_{..}$ | $\overline{y}_{..}$ |

An entry in the table (e.g., $y_{ij}$) represents the $j^{th}$ observation taken under the factor at level $i$.

- There will be, in general, $n$ observations under the $i^{th}$ level.
- $y_{i.}$ represents the total of the observations under the $i^{th}$ level.
- $\overline{y_{i.}}$ represent the average of the observation under the $i^{th}$ level.
- $y_{..}$ represent the grand total of all the observation under the factor.
- $\overline{y}_{..}$ represent the average grand total of all the observation under the factor.

# One-way ANOVA

Expressed symbolically,

$y_{i.} = \sum_{j=1}^{n_i} y_{ij}\ ,\ \bar{y}_{i..} = \frac{y_{i.}}{n_i}, y_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\ , \bar{y}_{..} = {}^{y_{..}}/_{N};\ N = \sum_{i=1}^{k} n_i$

The correlated sum of squares for each factor level $SS_i = \sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\ ;\ i = 1(1)k.$

We then calculate a pooled sum of squares $SS_p = \sum_{i=1}^{k} SS_i$

Finally, the pooled sample of variance is $s_p = \frac{SS_p}{pooled\ degree\ of\ fredom} = \frac{SS_p}{\sum n_i - k}$

Note that if the individual variances are available, the same can be computed as

$$s_p = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum n_i - k}$$

where $s_i^2$ are the variances for each sample. This is also called **variance within samples** and also popularly be denoted as

# Realationship Analysis

# Relationship Analysis

## Example: Wage Data

A large data regarding the wages for a group of employees from the eastern region of India is given.

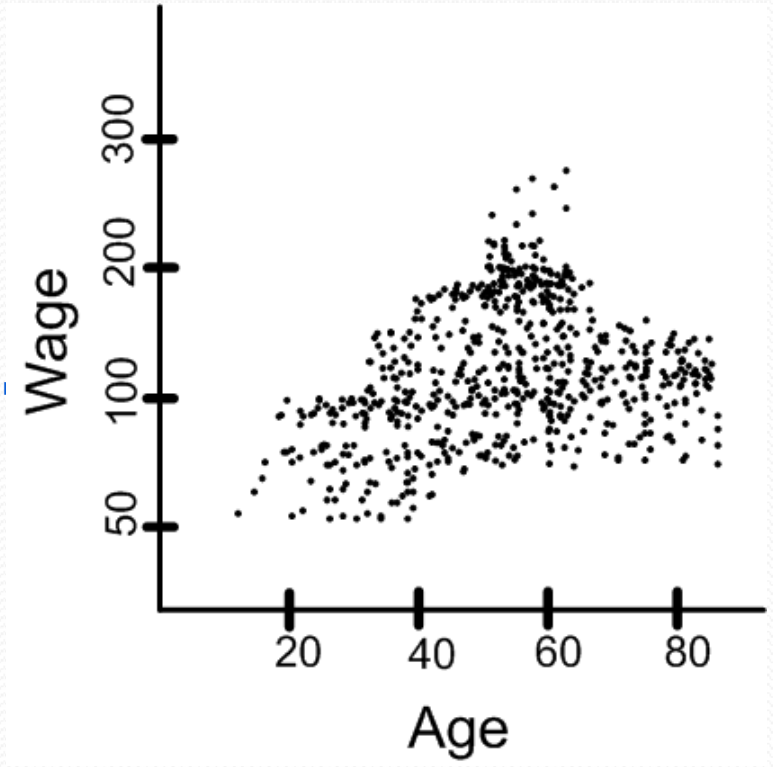
In particular, we wish to understand the following relationships:

- *Employee's age and wage:* How wages vary with ages?
- *Calendar year and wage:* How wages vary with time?
- *Employee's age and education:* Whether wages are anyway related with employees' education levels?

# Relationship Analysis

## Example: Wage Data

- Case I. Wage versus Age
  - From the data set, we have a graphical representations, which is as follows:



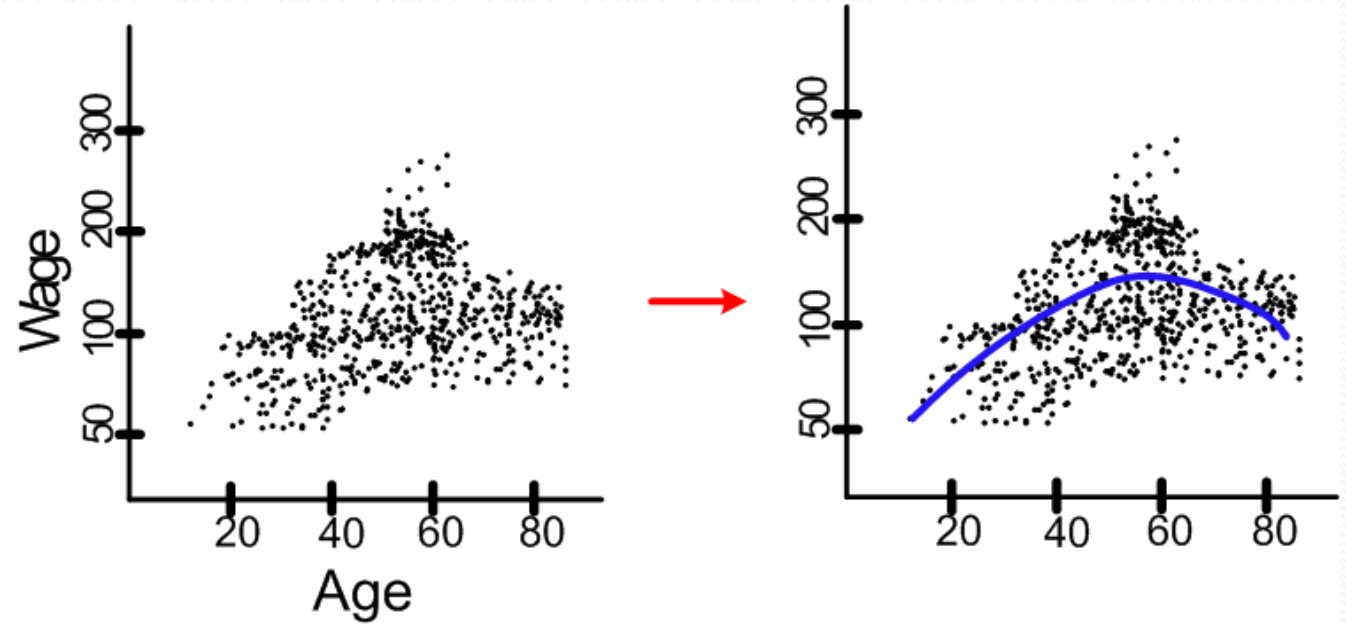

?

How wages vary with ages?

# Relationship Analysis

Example: Wage Data

- *Employee's age and wage:* How wages vary with ages?

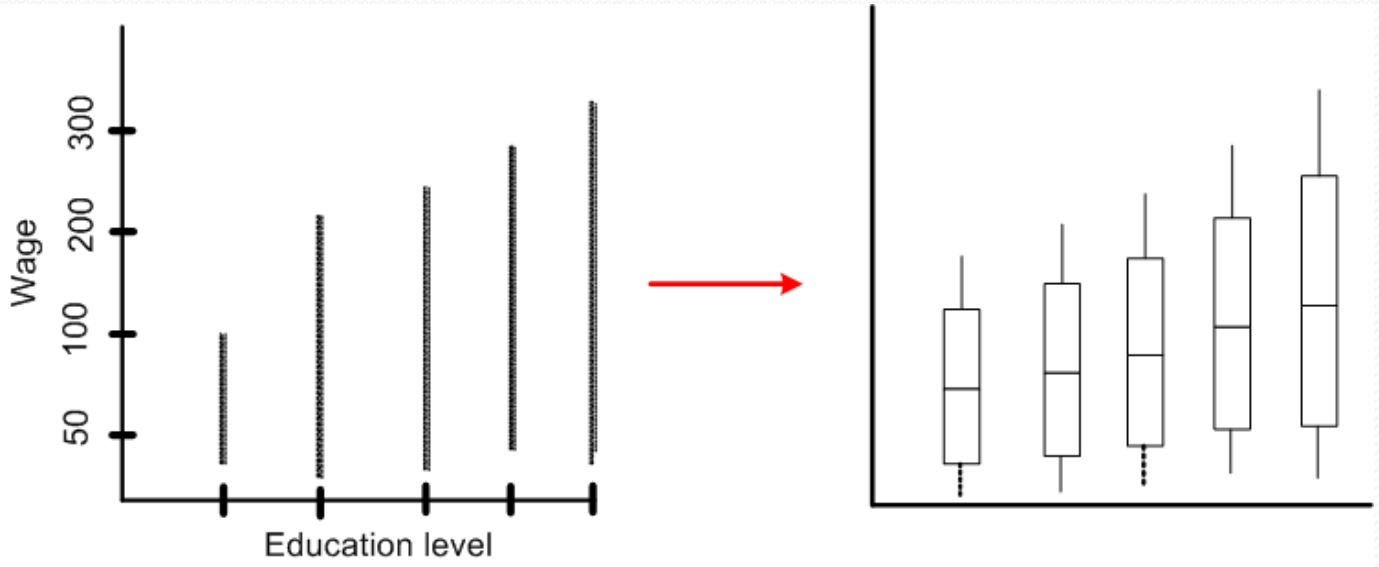



Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

# Relationship Analysis

Example: Wage Data

- *Wage and education level:* Whether wages vary with employees' education levels?




Interpretation: On the average, wage increases with the level of education.

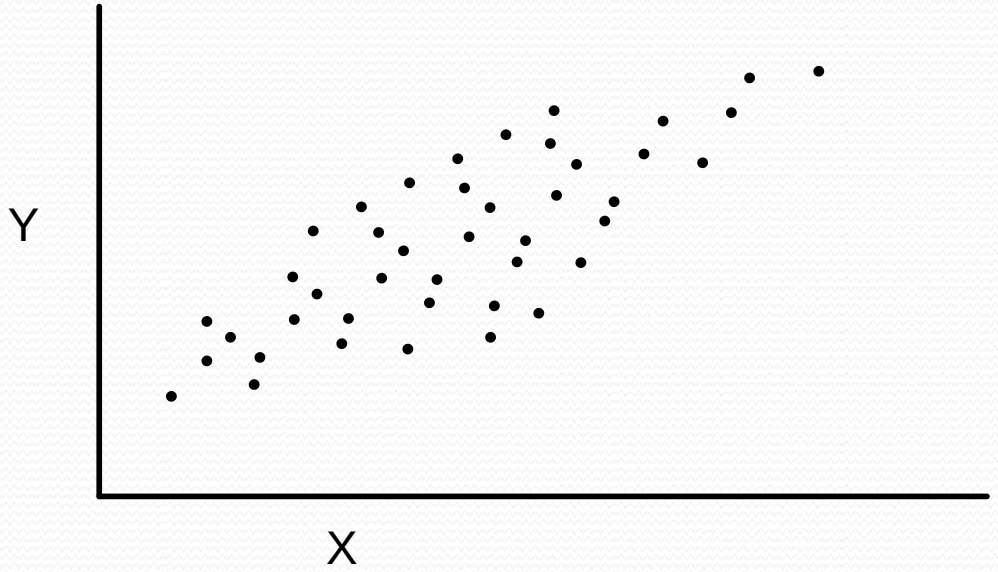
# Relationship Analysis

Given an employee's wage can we predict his age?

Whether wage has any association with both year and education level?

etc….

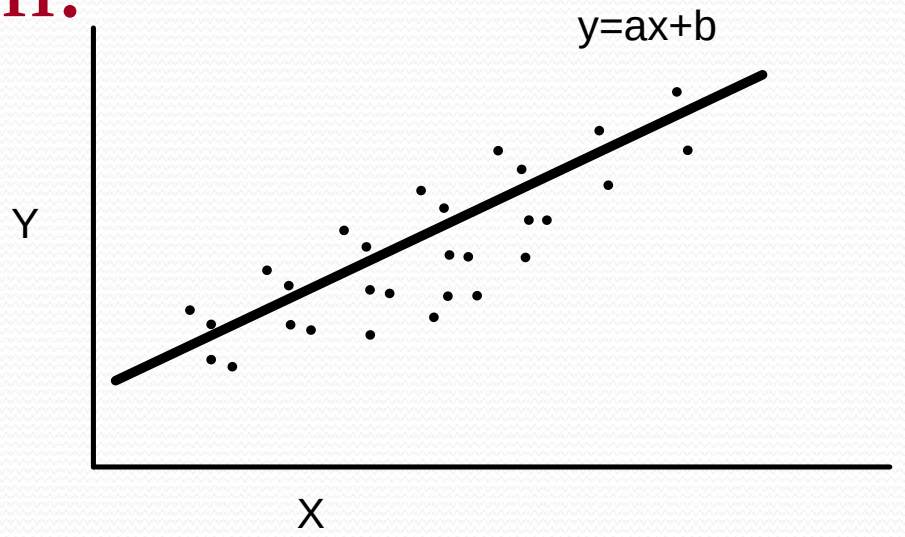
# Question for You!




Suppose there are countably infinite points in the *XY plane*. We need a huge memory to store all such points.

**Is there any way out to store this information with a least amount of memory?**

# Solution:




Just decide the values of **a** and **b**
(as if storing one point's data only!)

Note: Here, the trick was to find a relationship among all the points.

# Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?
If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?
If yes, of what degree and in which direction?

To find solutions to the above questions, two approaches are known.
**Correlation Analysis**
**Regression Analysis**

# Correlation Analysis

In statistics, the word correlation is used to denote some form of association between two variables.

Example: **Weight** is correlated with **height**

| $A$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | … … |
|---|---|---|---|---|---|---|---|
| $B$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ | … … |

The correlation may be positive, negative or zero.

**Positive correlation:** If the value of the attribute $A$ increases with the increase in the value of the attribute $B$ and vice-versa.

**Negative correlation:** If the value of the attribute $A$ decreases with the increase in the value of the attribute $B$ and vice-versa.

**Zero correlation:** When the values of attribute $A$ varies at random with $B$ and vice-versa.

# Correlation Coefficient

**Correlation coefficient is used to measure the degree of association.**

It is usually denoted by *r*.

The value of *r* lies between +1 and -1.

Positive values of *r* indicates positive correlation between two variables, whereas, negative values of *r* indicate negative correlation.

$r = +1$ implies perfect positive correlation, and otherwise.

The value of *r* nearer to +1 or -1 indicates high degree of correlation between the two variables.

*r* = 0 implies, there is no correlation

# Correlation

Correlation coefficient between two R.V.s X and Y, usually denoted by $r(X,Y)$ $or$ $r_{XY}$ is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$r_{XY}$ provided a measure of linear relationship between X and Y.
It is a measure of degree of relationship.

# Heart Disease and Cigarettes

Data on coronary heart disease and cigarette smoking in 21 developed countries (Landwehr and Watkins, 1987)

Data have been rounded for computational convenience.

# Data

*Surprisingly, the U.S. is the first country on the list--the country with the highest consumption and highest mortality.*

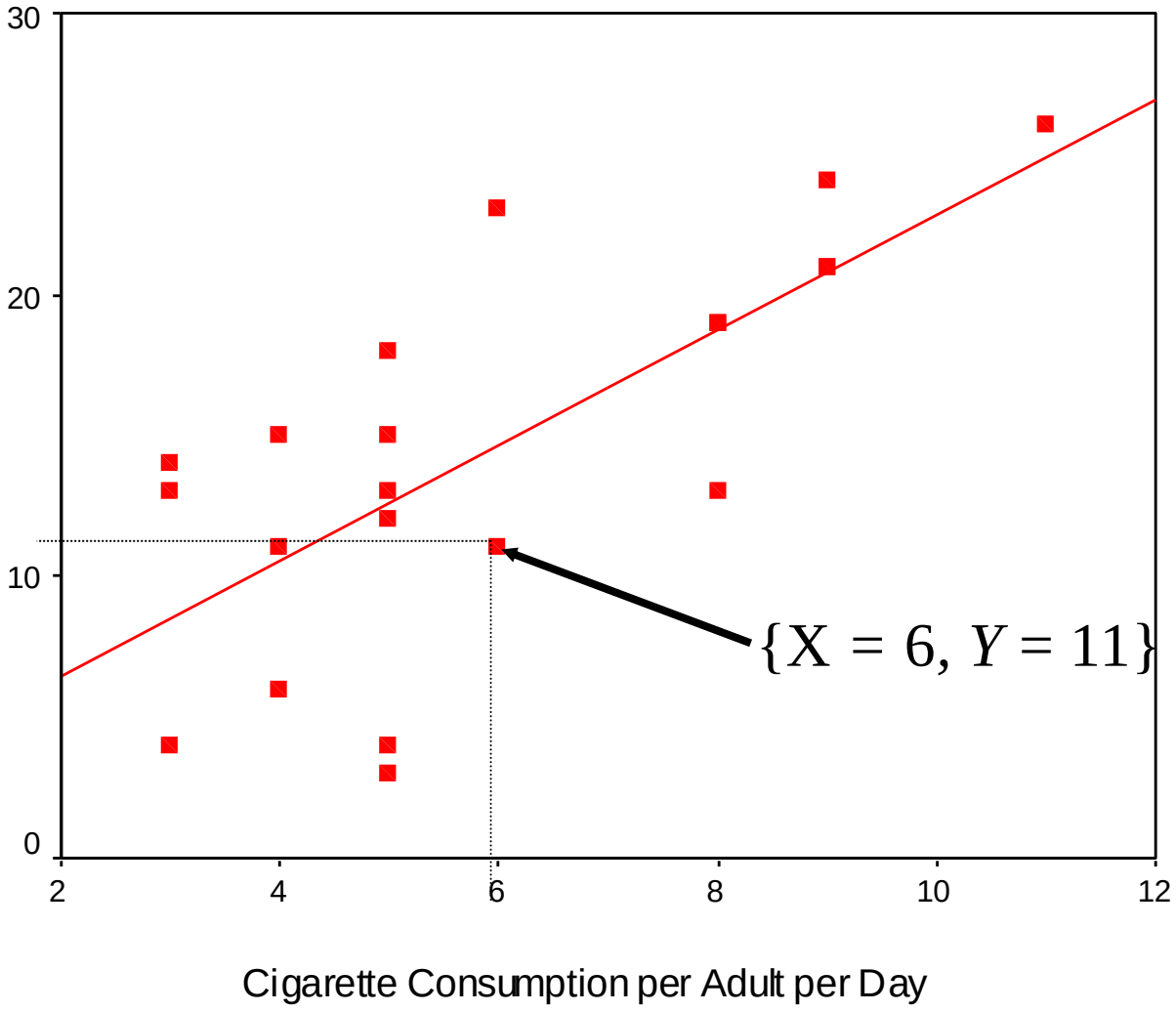
## Scatterplot of Heart Disease

- CHD Mortality goes on ordinate (Y axis)
- Cigarette consumption on abscissa (X axis)
- Best fitting line included for clarity

| Country | Cigarettes | CHD |
|---|---|---|
| 1 | 11 | 26 |
| 2 | 9 | 21 |
| 3 | 9 | 24 |
| 4 | 9 | 21 |
| 5 | 8 | 19 |
| 6 | 8 | 13 |
| 7 | 8 | 19 |
| 8 | 6 | 11 |
| 9 | 6 | 23 |
| 10 | 5 | 15 |
| 11 | 5 | 13 |
| 12 | 5 | 4 |
| 13 | 5 | 18 |
| 14 | 5 | 12 |
| 15 | 5 | 3 |
| 16 | 4 | 11 |
| 17 | 4 | 15 |
| 18 | 4 | 6 |
| 19 | 3 | 13 |
| 20 | 3 | 4 |
| 21 | 3 | 14 |

# What Does the Scatterplot Show?

- As smoking increases, so does coronary heart disease mortality.
- Relationship looks strong
- Not all data points on line.
- This gives us residuals" or "errors of prediction"



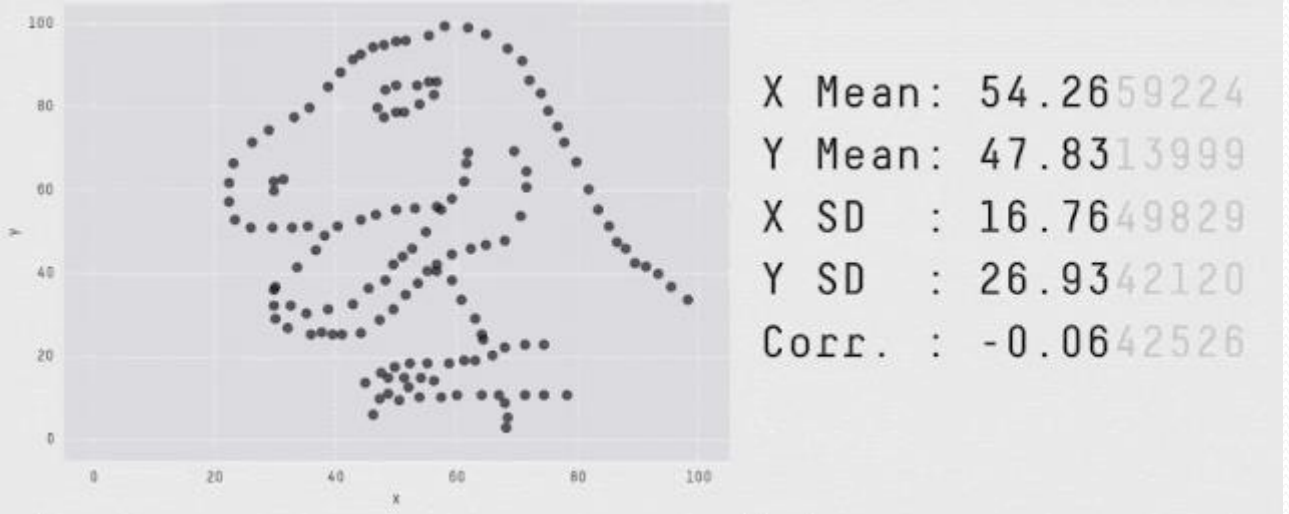
# Example: Heart Disease and Cigarettes

| Country | X (Cig.) | Y (CHD) | $(X-\bar{X})$ | $(Y-\bar{Y})$ | $(X-\bar{X})*(Y-\bar{Y})$ |
|---|---|---|---|---|---|
| 1 | 11 | 26 | 5.05 | 11.48 | 57.97 |
| 2 | 9 | 21 | 3.05 | 6.48 | 19.76 |
| 3 | 9 | 24 | 3.05 | 9.48 | 28.91 |
| 4 | 9 | 21 | 3.05 | 6.48 | 19.76 |
| 5 | 8 | 19 | 2.05 | 4.48 | 9.18 |
| 6 | 8 | 13 | 2.05 | -1.52 | -3.12 |
| 7 | 8 | 19 | 2.05 | 4.48 | 9.18 |
| 8 | 6 | 11 | 0.05 | -3.52 | -0.18 |
| 9 | 6 | 23 | 0.05 | 8.48 | 0.42 |
| 10 | 5 | 15 | -0.95 | 0.48 | -0.46 |
| 11 | 5 | 13 | -0.95 | -1.52 | 1.44 |
| 12 | 5 | 4 | -0.95 | -10.52 | 9.99 |
| 13 | 5 | 18 | -0.95 | 3.48 | -3.31 |
| 14 | 5 | 12 | -0.95 | -2.52 | 2.39 |
| 15 | 5 | 3 | -0.95 | -11.52 | 10.94 |
| 16 | 4 | 11 | -1.95 | -3.52 | 6.86 |
| 17 | 4 | 15 | -1.95 | 0.48 | -0.94 |
| 18 | 4 | 6 | -1.95 | -8.52 | 16.61 |
| 19 | 3 | 13 | -2.95 | -1.52 | 4.48 |
| 20 | 3 | 4 | -2.95 | -10.52 | 31.03 |
| 21 | 3 | 14 | -2.95 | -0.52 | 1.53 |
| Mean | 5.95 | 14.52 | | | |
| SD | 2.33 | 6.69 | | | |
| Sum | | | | | 222.44 |

$$Cov_{cig.\&CHD} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1} = \frac{222.44}{21-1} = 11.12$$

$$r = \frac{cov_{XY}}{s_X s_Y} = \frac{11.12}{(2.33)(6.69)} = \frac{11.12}{15.59} = .713$$

# What stories can scatter plot tell?



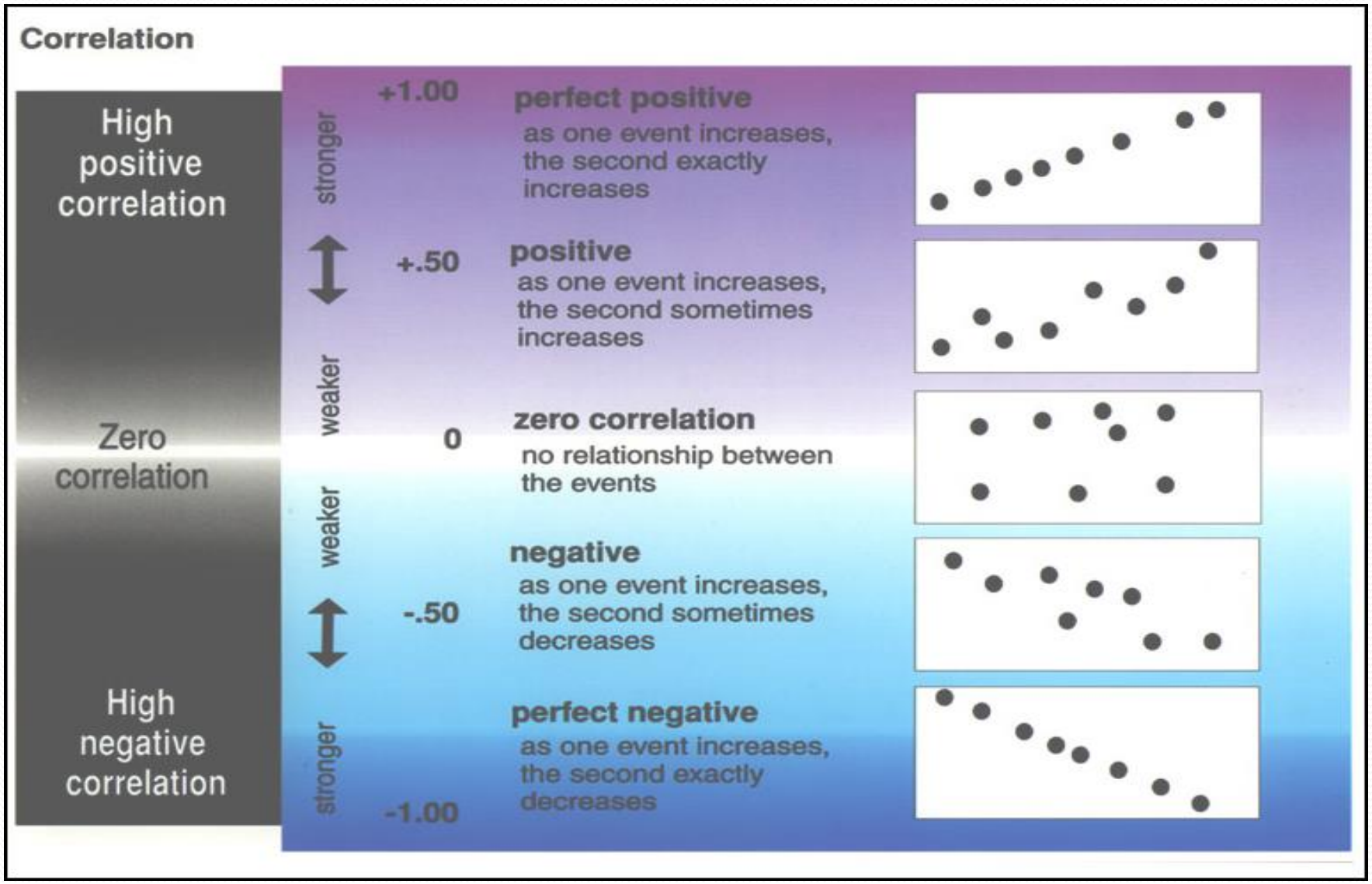

In 1973, a famous statistician, Francis Anscombe, demonstrated how important it is to visualize the data. The concept got extended later to create Datasaurus Dozen.

It is a collection of 12 scatterplots with the same means, standard deviations, and correlation coefficient for X and Y (up to 2 decimal places).

However, the shape of the data is very different from each other. Therefore, the scatterplots tell very different stories about the behavior and interrelationships of X and Y.

Data available at https://cran.r-project.org/web/packages/datasauRus/vignettes/Datasaurus.html

# Remarks from Scatter Plots & Correlation Coefficient



# Regression Analysis

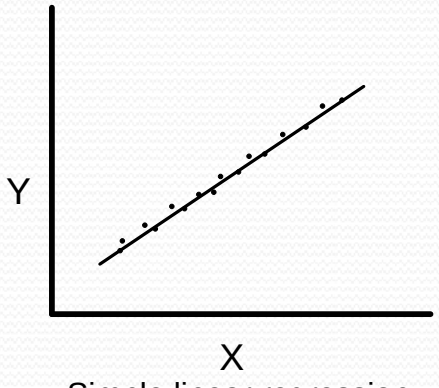
The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

**Classification of Regression Analysis Models**

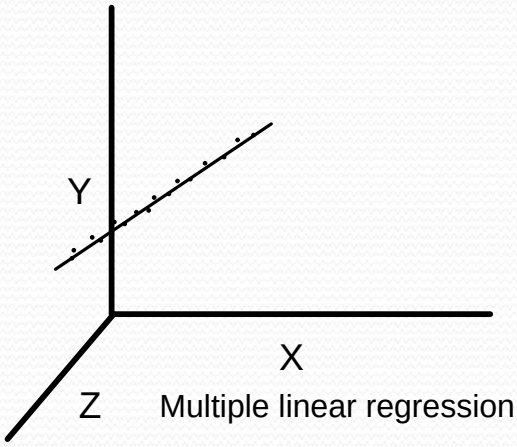
Linear regression models

1. Simple linear regression
2. Multiple linear regression

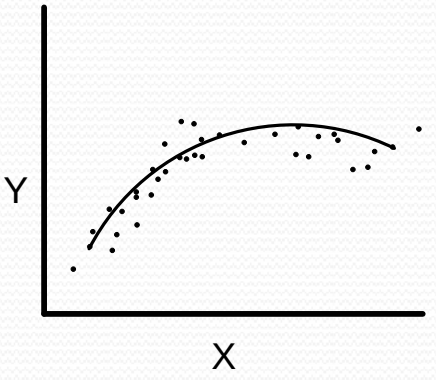
Non-linear regression models



Simple linear regression



Multiple linear regression



Non-linear regression

# Galton Board

- Sir Francis Galton, Charles Darwin's half-cousin, invented the 'Galton Board' in 1874 to demonstrate that the normal distribution is a natural phenomenon.

- It specifically shows that the binomial distribution approximates a normal distribution with a large enough sample size.

# Linear Regression

- The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable.
- That means we want to understand the relationship.
- The line of regression is the line of "best fit" and is obtained by **the principle of least squares**.

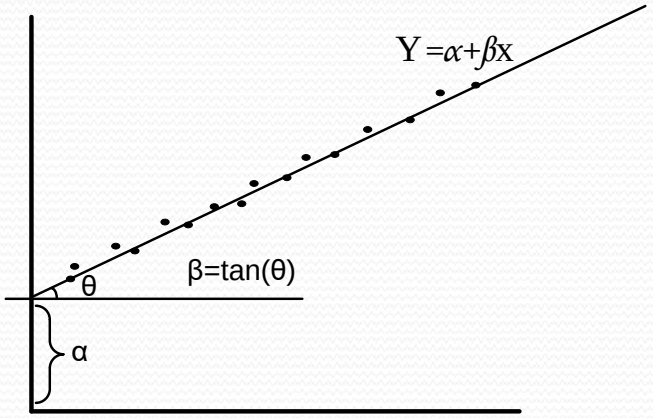
*Y* - the variables you are predicting
i.e. dependent variable

*X* - the variables you are using to predict
i.e. independent variable

$\hat{Y}$- your predictions (also known as Y')

# Simple Linear Regression Model

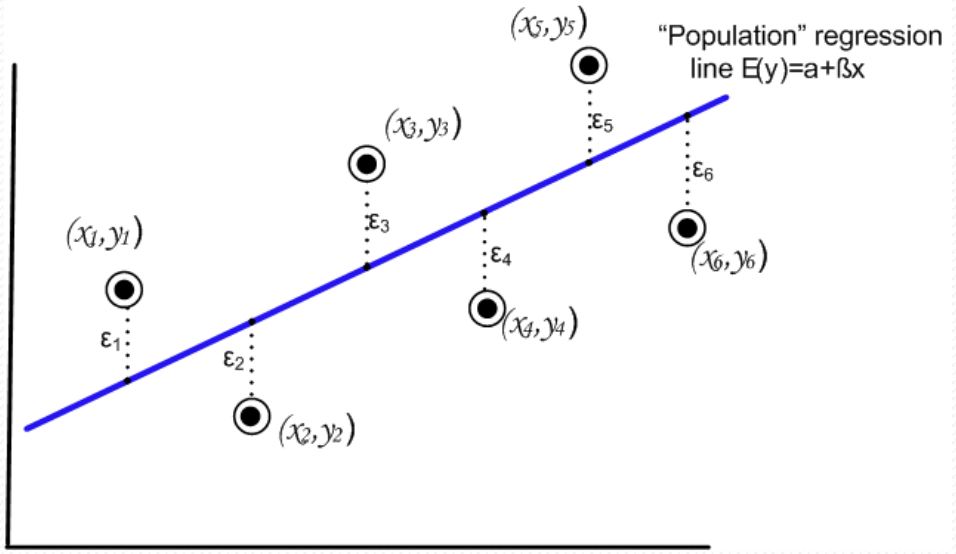
In simple linear regression, we have only two variables:
Dependent variable (also called Response), usually denoted as $Y$.
Independent variable (alternatively called Regressor), usually denoted as $x$.
A reasonable form of a relationship between the Response $Y$ and the Regressor $x$ is the linear relationship, that is in the form $Y = \alpha + \beta x$




**Note:**
There are infinite number of lines (and hence $\alpha_s$ *and* $\beta_s$)

The concept of regression analysis deal with finding the best relationship between $Y$ and $x$ (and hence best fitted values of $\alpha$ and $\beta$) quantifying the strength of that relationship.

# Regression Analysis




Given the set $[(x_i, y_i), i = 1,2, \ldots \ldots, n]$ of data involving $n$ pairs of $(x, y)$ values, our objective is to find "true" or population regression line such that $Y = \alpha + \beta x + \in$

Here, $\in$ is a random variable with $E(\in) = 0$ and $var(\in) = \sigma^2$. The quantity $\sigma^2$ is often called the **error variance**.

**Note:**

$E(\in) = 0$ implies that at a specific $x$, the $y$ values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).

$\alpha$ and $\beta$ are called **regression coefficients**.

$\alpha$ and $\beta$ values are to be estimated from the data.

# Assumption on the model

The Linear Regression Model for $i\ th$ observation

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Assumptions:

$E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$

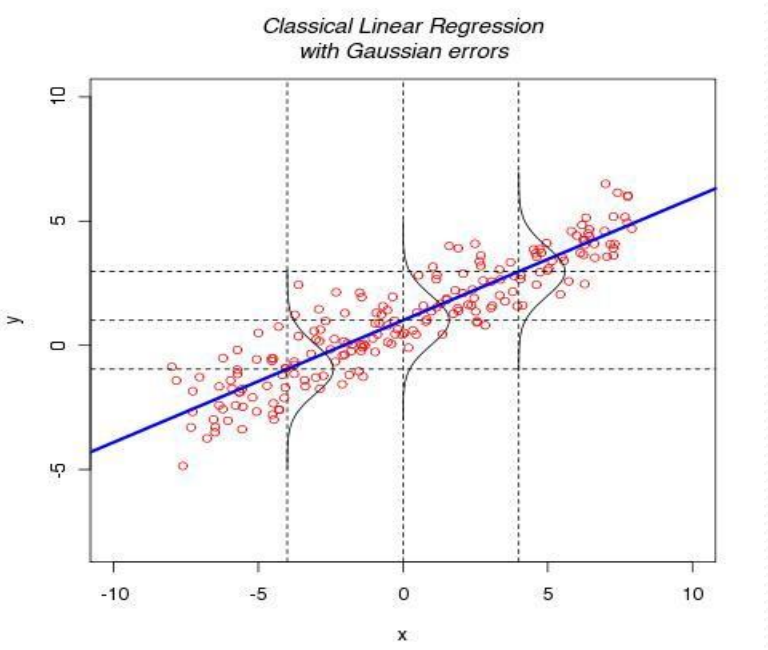
$Cov(\epsilon_i, \epsilon_j) = 0, i \neq j;\ \epsilon_i\ (iid) \sim N(0, \sigma^2)$

Consequent assumptions on Y:

$E(Y_i) = \alpha + \beta x_i$

$Var(Y_i) = \sigma^2$

$Y_i's\ are\ independent\ and\ normally\ distributed.$



# Basic Assumptions of the Linear Regression Model

The slope $\beta$ of the population regression line is the *average* change in $Y$ associated with a 1-unit increase in $x$.

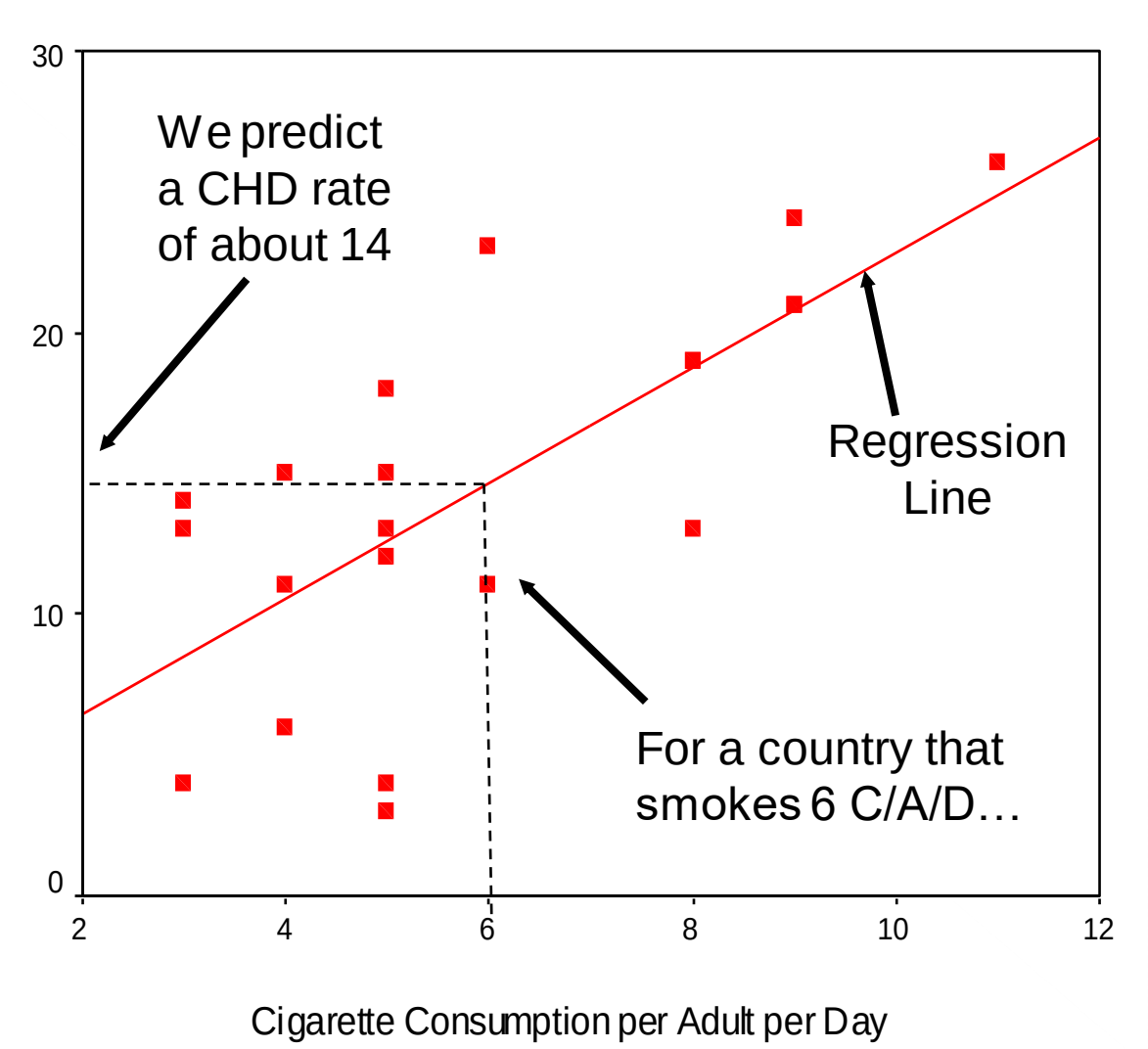
The $Y$ intercept $\alpha$ is the height of the population line when $x = 0$.

The value of $\sigma$ determines the extent to which $(x, y)$ observations deviate from the regression line.

When $\sigma$ is small, most observations will be quite close to the line, but when $\sigma$ is large, there are likely to be some large deviations.

# Example: Heart Disease and Cigarettes

| Country | Cigarettes | CHD |
|---|---|---|
| 1 | 11 | 26 |
| 2 | 9 | 21 |
| 3 | 9 | 24 |
| 4 | 9 | 21 |
| 5 | 8 | 19 |
| 6 | 8 | 13 |
| 7 | 8 | 19 |
| 8 | 6 | 11 |
| 9 | 6 | 23 |
| 10 | 5 | 15 |
| 11 | 5 | 13 |
| 12 | 5 | 4 |
| 13 | 5 | 18 |
| 14 | 5 | 12 |
| 15 | 5 | 3 |
| 16 | 4 | 11 |
| 17 | 4 | 15 |
| 18 | 4 | 6 |
| 19 | 3 | 13 |
| 20 | 3 | 4 |
| 21 | 3 | 14 |

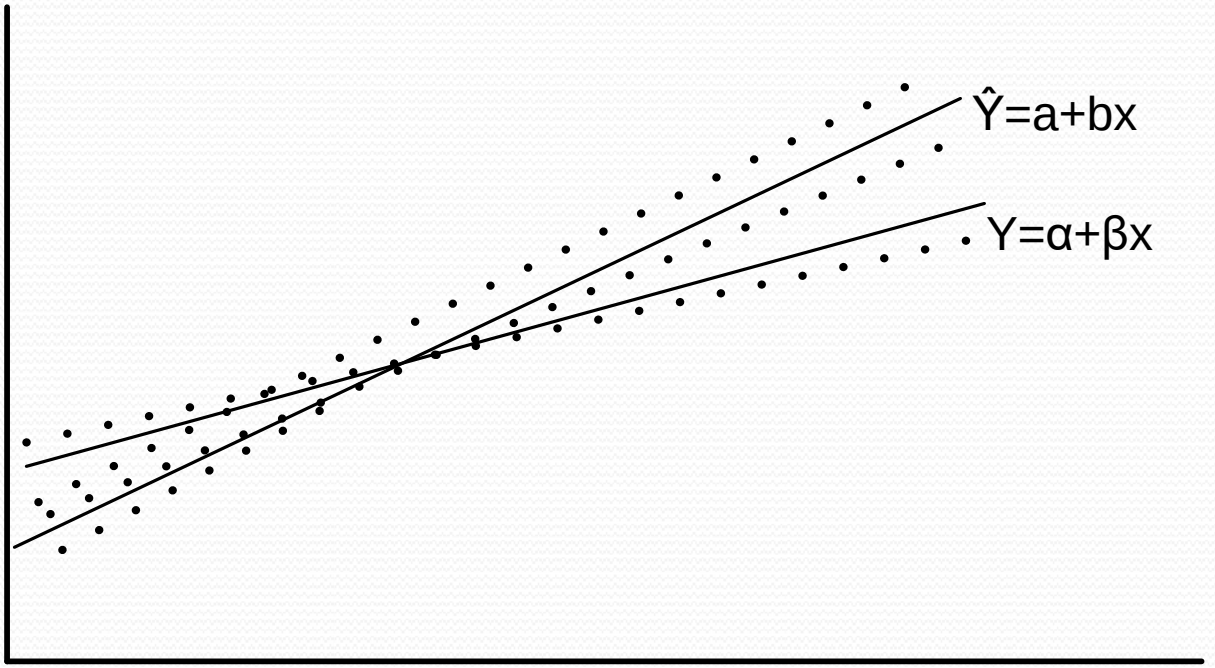


# True versus Fitted Regression Line

The task in regression analysis is to estimate the regression coefficients $\alpha$ and $\beta$. Suppose, we denote the estimates $a$ for $\alpha$ and $b$ for $\beta$. Then the fitted regression line is

$$\hat{Y} = a + bx$$

where $\hat{Y}$ is the predicted or fitted value.



# Regression line

Formula:

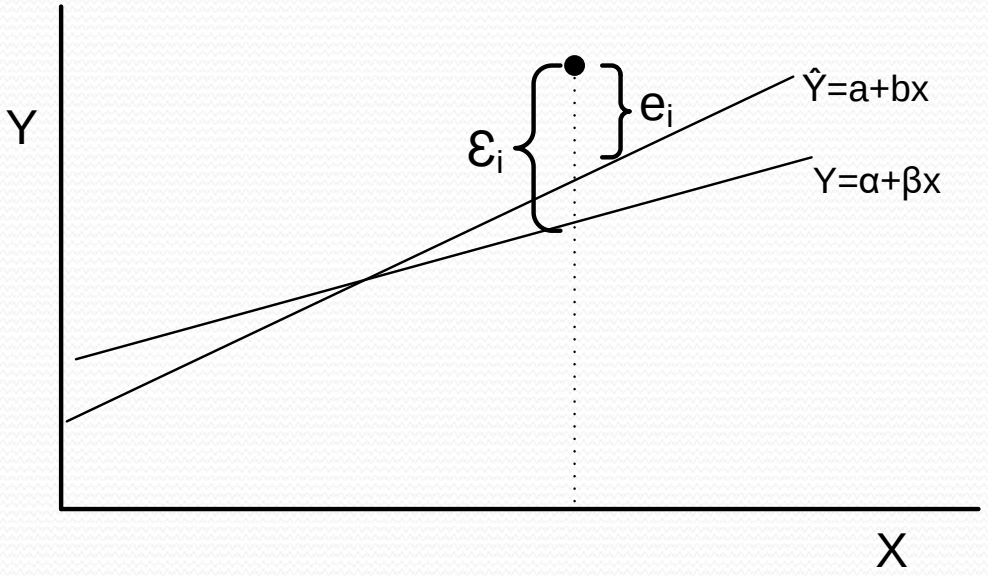
$$\hat{Y} = a + bX$$

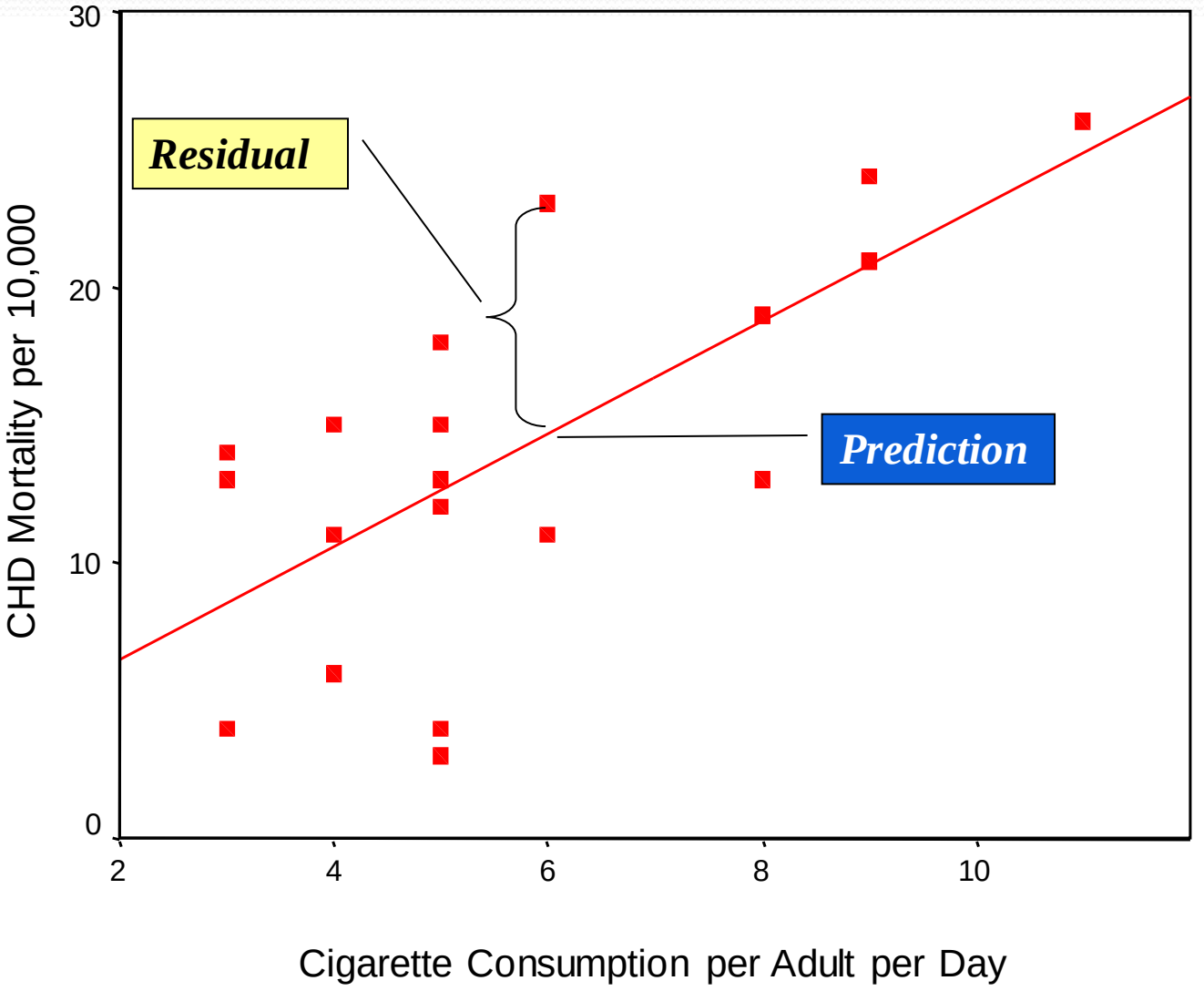
- $\hat{Y}$ = the predicted value of $Y$ (e.g. CHD mortality)
- $\boldsymbol{X}$ = the predictor variable (e.g. average cig./adult/country)
- $a$ and $b$ are "Coefficients"
- $b$ = slope ,i.e. change in predicted $Y$ for one unit change in $X$
- $a$ = intercept , i.e. value of when $X = 0$

# Least Square Method to estimate $\alpha$ and $\beta$

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\hat{Y} = a + bx$. Thus, $i^{th}$ residual is

$$e_i = Y_i - \hat{Y}_i, \; i = 1,2,3, \ldots \ldots, n$$




Residual
Prediction
CHD Mortality per 10,000
30
20
10
0
2
4
6
8
10
Cigarette Consumption per Adult per Day

# Least Square method

The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as SSE

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

We are to minimize the value of SSE and hence to determine the parameters of *a* and *b*.

Differentiating SSE with respect to *a* and *b*, we have

$$\frac{\partial(SSE)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\frac{\partial(SSE)}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) . x_i$$

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$

$$\frac{\partial(SSE)}{\partial b} = 0$$

# Least Square method to estimate $\alpha$ and $\beta$

Thus, we set

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

These two equations can be solved to determine the values of $a$ and $b$, and it can be calculated that

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{\sigma_X^2}$$

$$a = \bar{y} - b\bar{x}$$

- Line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$:

$$Y - \bar{y} = \frac{\sigma_Y}{\sigma_X} r_{XY}(X - \bar{x})$$

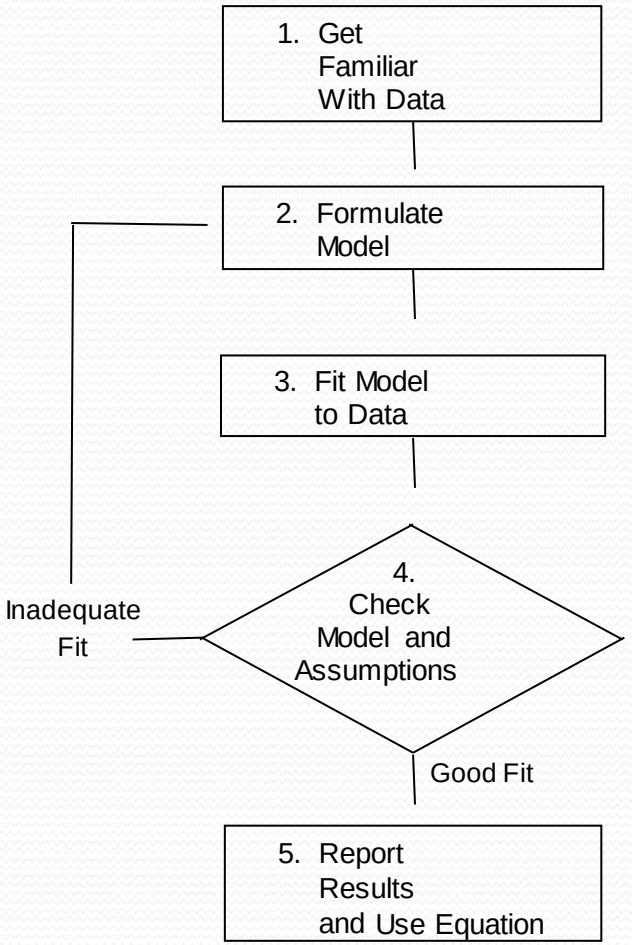
# For the Data: Heart Disease and Cigarettes

$$Cov(X,Y) = 11.12619$$
$$\sigma^2{}_X = 2.334014^2 = 5.447619$$
$$b = 11.12619/5.447619 = 2.042395$$
$$a = 14.52381 - 2.042395 * 5.952381 = 2.366696$$

the equation of the least-squares line:

$$\hat{y} = 2.367 + 2.042\,x$$

# Five Step Regression Procedure: Overview




- Look at plots
- Look at descriptive statistics

- Linear or curvilinear?
- One X or more Xs?
- Transform?
- Discrete X, discrete Y?

- Do the regression

- Look at residuals plots
- Look at unusual observations
- Look at R-Sq
- Look at P-values for b

- Make predictions for X-values of interest

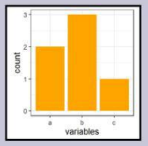
# Cheatsheet

## What type of **DATA VISUALIZATION** to choose?

## What do you want to show?

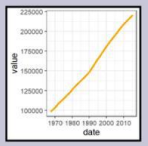
## COMPARISON

COMPARISON CHARTS SHOW THE DIFFERENCES BETWEEN VALUES SO YOU CAN QUICKLY COMPARE CATEGORIES AS WELL AS SEE HOW VALUES CHANGE OVER TIME.




### BAR PLOT

COMPARING CATEGORIES WITHIN THE SAME MEASURE OR THE SAME MEASURES.

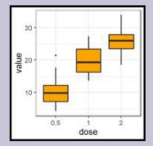



### LINE PLOT

COMPARING TRENDS OVER TIME.

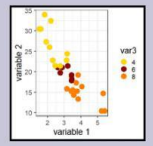
## RELATIONSHIP

RELATIONSHIP CHARTS ARE USED TO EXPLORE RELATIONSHIPS BETWEEN VALUES. THEY ALLOW YOU TO FIND CORRELATIONS, OUTLIERS AND CLUSTERS OF DATA.




### BOXPLOT

DISPLAYING OUTLIERS AND DATA CLUSTERS.

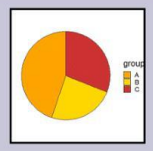



### SCATTER PLOT

DISPLAYING THE RELATIONSHIP BETWEEN TWO OR THREE MEASURES FOR A DIMENISON.

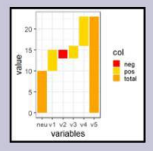
## COMPOSITION

COMPOSITION CHARTS ARE USED TO ANALYZE HOW EACH COMPONENT VALUE AFFECTS TO TOTAL.




### PIE CHART

DISPLAYING A STATIC COMPOSITION OF VALUES.

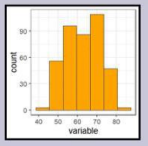



### WATERFALL CHART

DISPLAYING THE STATIC COMPOSITION OF A VALUE WITH ACCUMULATION OR SUBTRACTION FROM THE TOTAL.

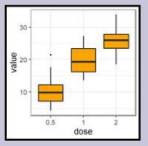
## DISTRIBUTION

DISTRIBUTION CHARTS ARE USED TO EXPLORE HOW VALUES ARE GROUPED IN YOUR DATA.




### HISTOGRAM

DISPLAYING THE DISTRIBUTION OF DATA INTERVALS.

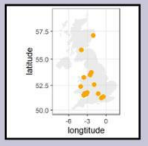



### BOXPLOT

DISPLAYING RANGES AND DISTIRIBUTION OF NUMERIC DATA.

## GEOGRAPHICAL DATA

GEORGAPHIC CHARTS PRESENT DATA BY GEOGRAPHIC LOCATION ON A MAP AS POINTS OR AREAS.



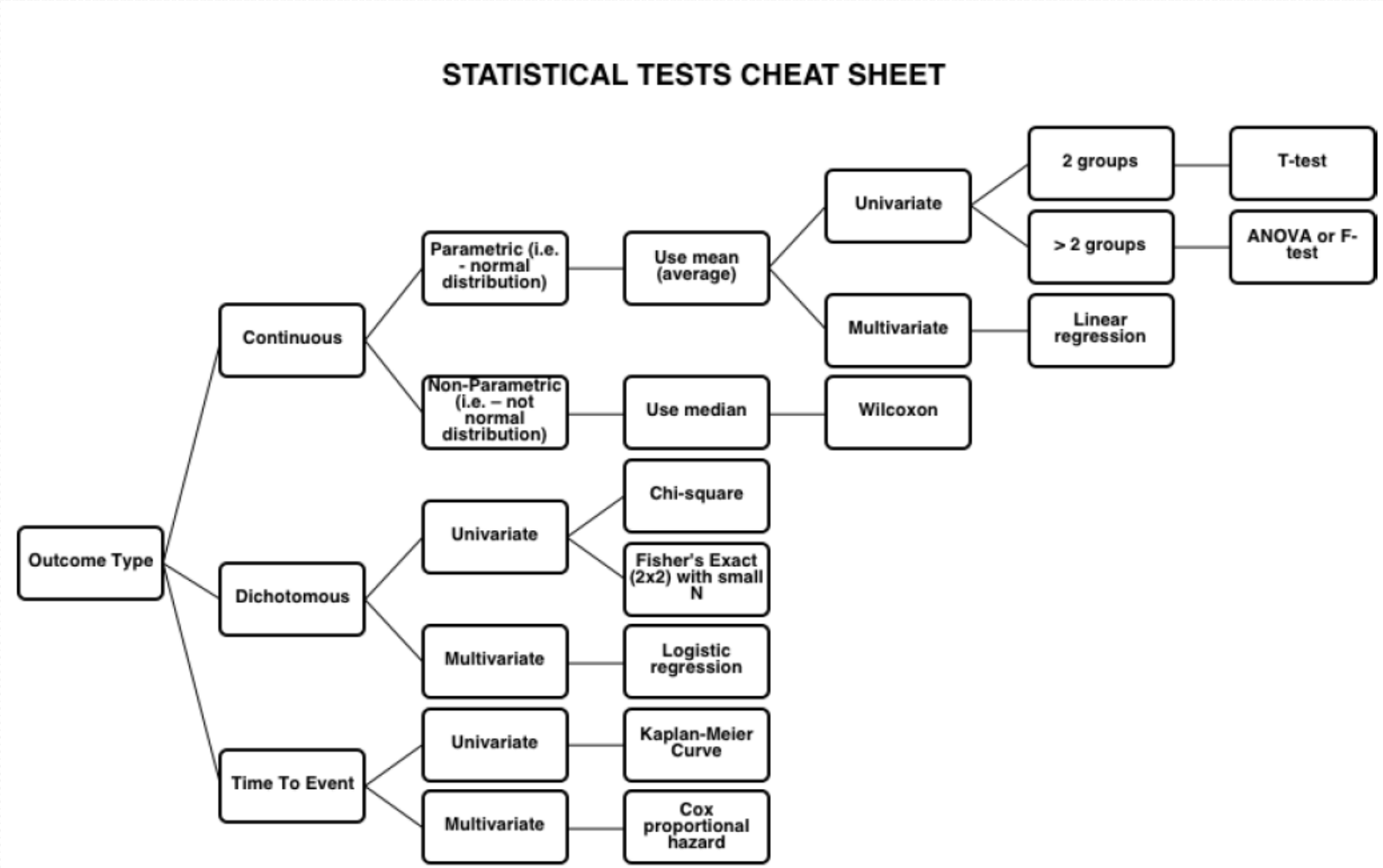

### MAP

DISPLAYING DATA REPRESENTED GEOGRAPHICALLY BY A POINT OR AREA.

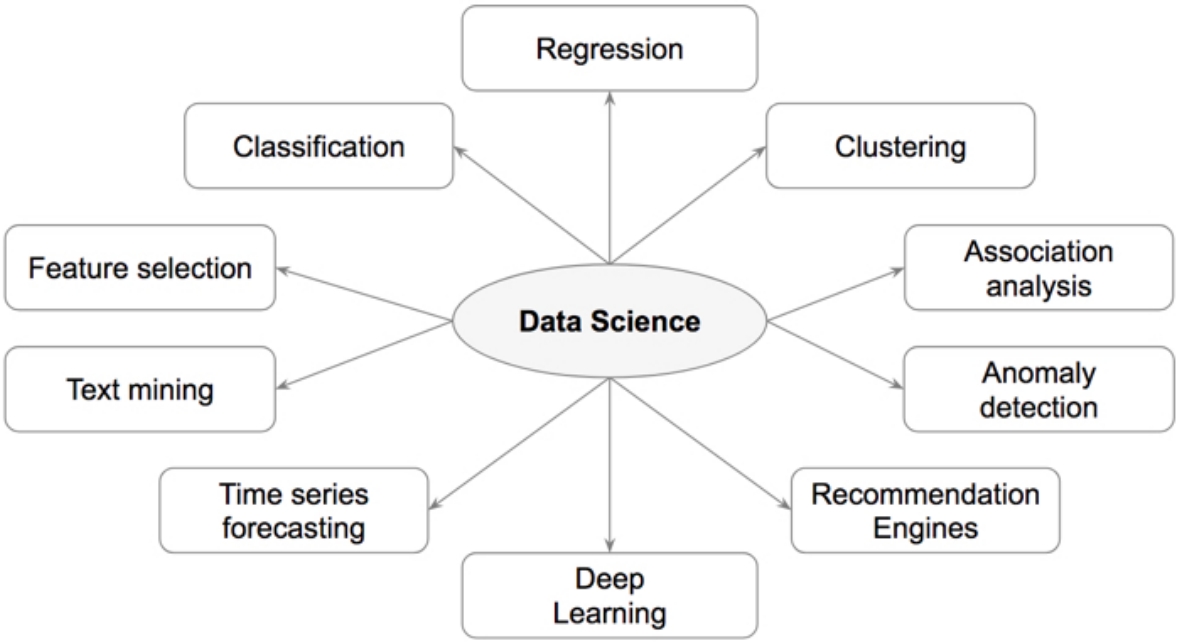
**REMEMBER!** THESE CHARTS ARE JUST AN EXAMPLE. ALWAYS USE A CHART THAT REPRESENTS YOUR DATA MOST TRANSPARENT AND WITHOUT MISUNDERSTANDING.

PAULINA JASZCZUK | PAWEŁ FIJAŁKOWSKI | ADAM FREJ | MICHAŁ PIASECKI

# Cheatsheet


STATISTICAL TESTS CHEAT SHEET
Outcome Type
Continuous
Parametric (i.e. - normal distribution)
Use mean (average)
Univariate
2 groups
T-test
> 2 groups
ANOVA or F-test
Multivariate
Linear regression
Non-Parametric (i.e. – not normal distribution)
Use median
Wilcoxon
Dichotomous
Univariate
Chi-square
Fisher's Exact (2x2) with small N
Multivariate
Logistic regression
Time To Event
Univariate
Kaplan-Meier Curve
Multivariate
Cox proportional hazard

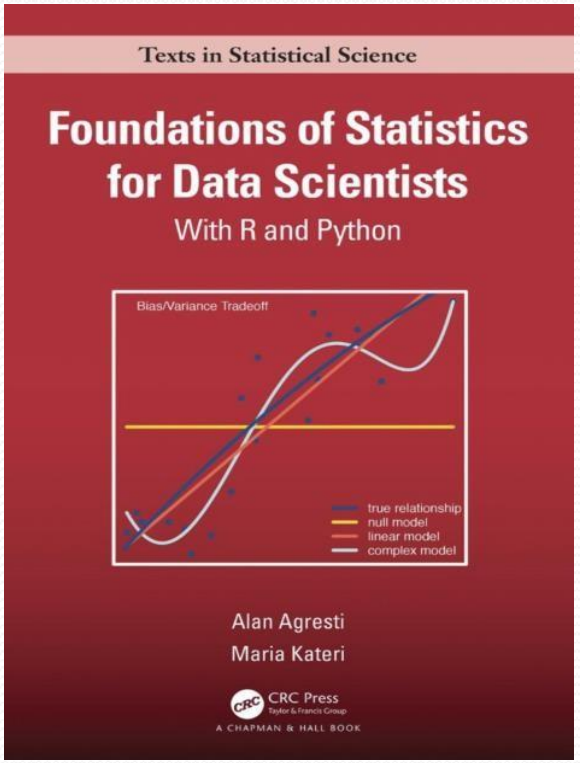
# Cheatsheet


Regression
Classification
Clustering
Feature selection
Data Science
Association analysis
Text mining
Anomaly detection
Time series forecasting
Recommendation Engines
Deep Learning

# Cheatsheet

| Dependent Variable Type (Ys) | Independent Variable Type (Xs) | Modelling Technique |
|---|---|---|
| Numerical | Numerical | 1. Linear Regression or Best Subset Regression<br>2. Non-linear Regression or Regression Splines<br>3. Regression Trees, Neural Nets, etc. |
| Numerical | Categorical + Numerical | 1. Linear Regression with Dummy Variables<br>2. Polynomial Regression with Dummy Variables<br>3. Regression Trees, Neural Nets, etc. |
| Categorical | Numerical | 1. Logistics Regression<br>2. Classification Trees<br>3. Support Vector Machines, Neural Nets, etc. |
| Categorical | Categorical + Numerical | 1. Logistic Regression with Dummy Variables<br>2. Classification Trees<br>3. Advanced Neural Nets, etc. |
| Numerical (Time dependent) | Numerical Exogenous Variables | 1. ARIMA, ETS, Naïve Model<br>2. Autoregressive Neural Network<br>3. RNN, LSTM, etc. |

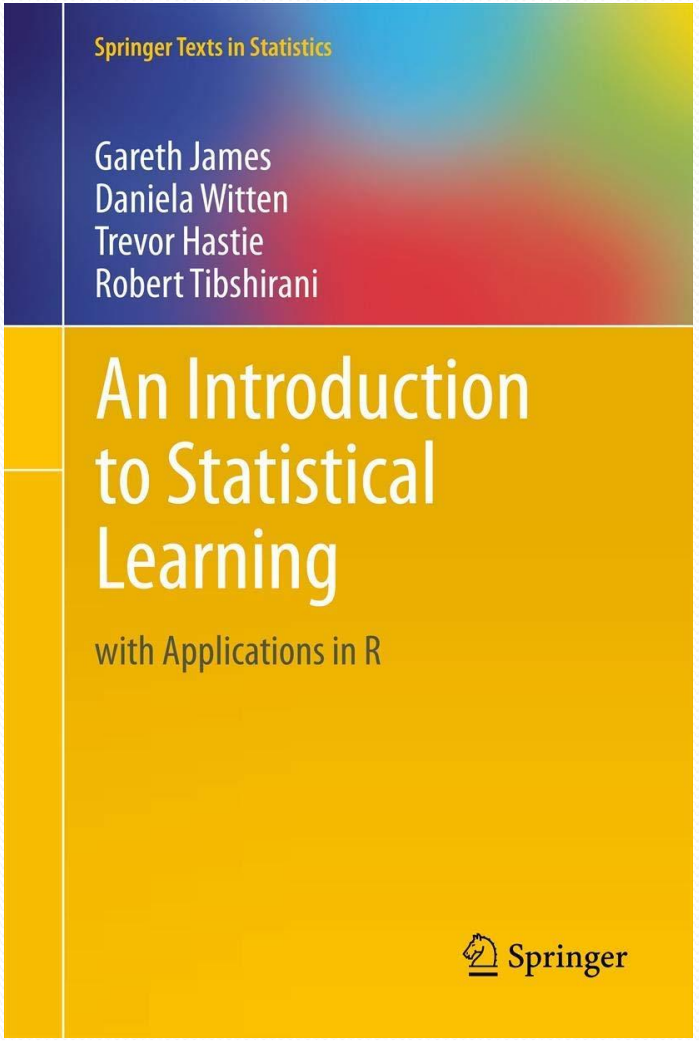
# Reference

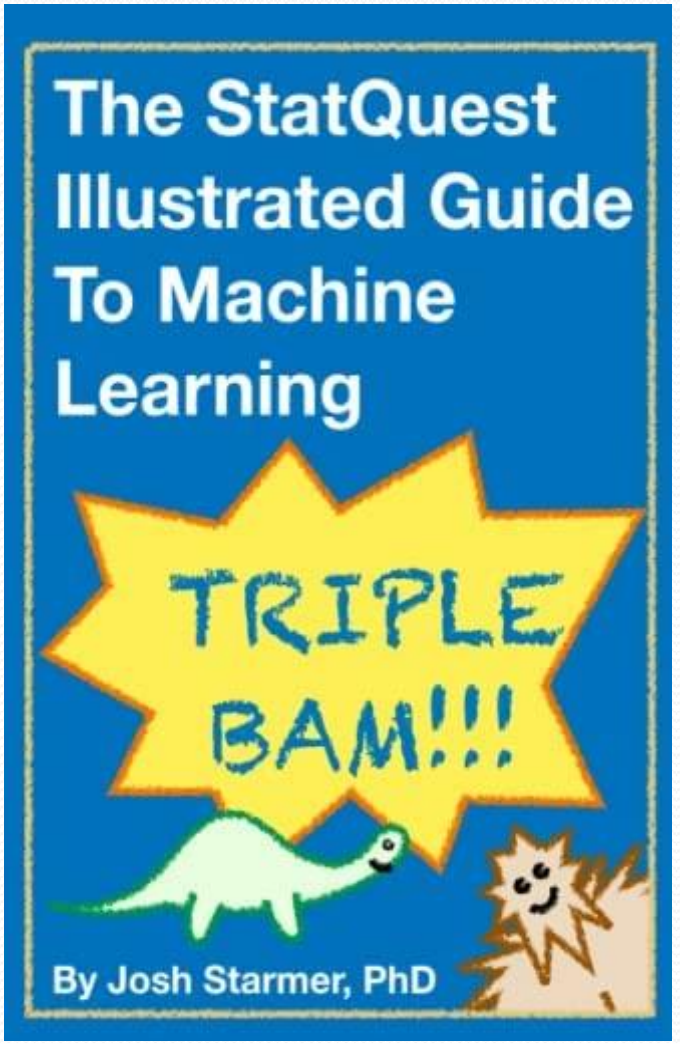



Foundations of Statistics for Data Scientists With R and Python By Alan Agresti, Maria Kateri (2022)

https://artofstat.com/web-apps

# Data Science & ML Reference



# THANK YOU!

For any query drop an email at tanujitisi@gmail.com