

Statistic, Sampling Techniques & Optimization

BY :- SHOBHIT TYAGI

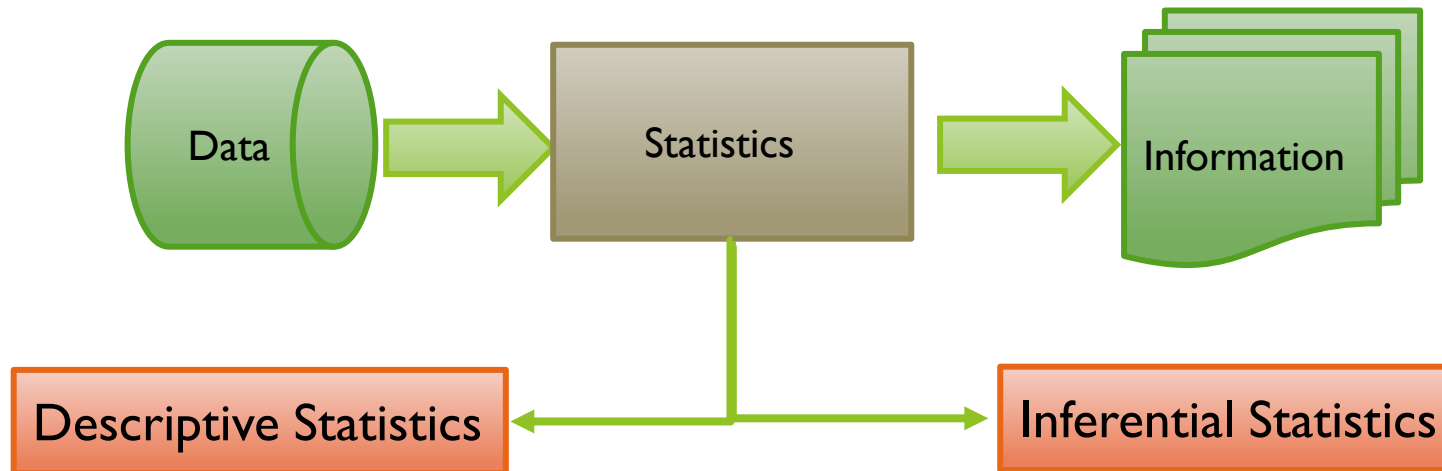


Contents

- Statistics
 - Descriptive Statistics
 - Inferential Statistics
- Sampling Techniques
- Optimization

Statistics

- It is the science that deals with the collection, description and analysis of data.
- It can be used to describe a particular data set as well as to draw conclusion.

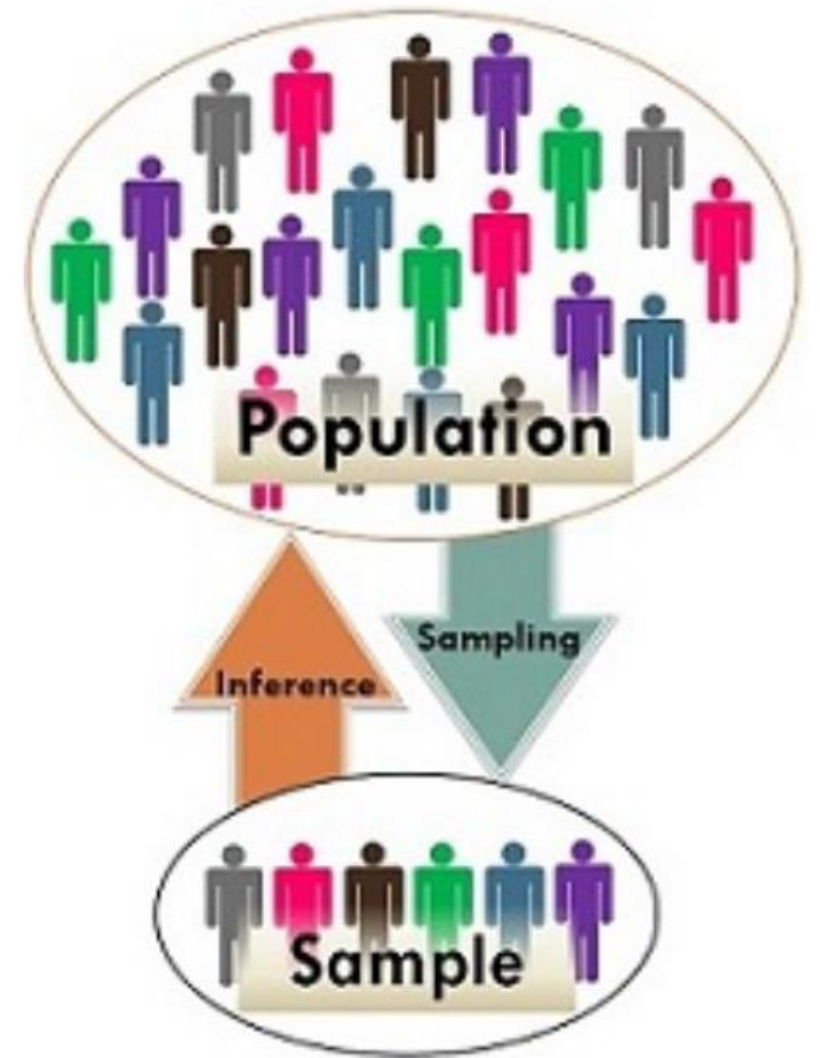


Descriptive Statistic

- Summary statistic that quantitatively describes or summarizes features of a collection of information

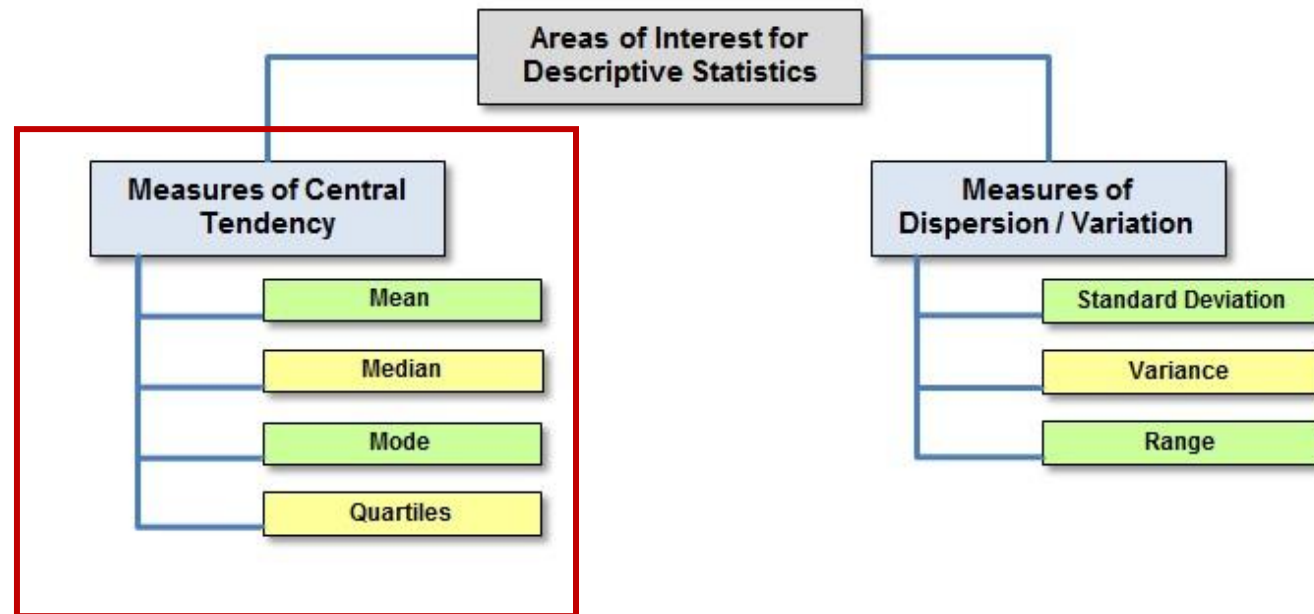
Sample vs Population

- Population denotes a large group consisting of elements
 - ▶ having something in common.
- In contrast, a sample is nothing but a part of population which is selected such that it represents the entire population.



Descriptive Statistics

- Descriptive statistics are set of coefficients that summarize a given data set.
- It can be further divided into two parts, mainly.
 - Measure of central tendency
 - Measure of variability



Measure of Central Tendency

- These measures indicate where most values in a distribution fall and is also referred as the **central location of a distribution**.

Mathematically

3, 7, 10, 8, 31, 10, 2

$$\text{Med}(X) = \begin{cases} X\left[\frac{n}{2}\right] & \text{if } n \text{ is even} \\ \frac{(X\left[\frac{n-1}{2}\right] + X\left[\frac{n+1}{2}\right])}{2} & \text{if } n \text{ is odd} \end{cases}$$

$$\text{Mean (avg)} = \frac{3 + 7 + 10 + 8 + 31 + 10 + 2}{7} = \frac{71}{7}$$

↓
10.14

7 numbers

$$\text{Median} = 2, 3, 7, 8, 10, 10, 31$$

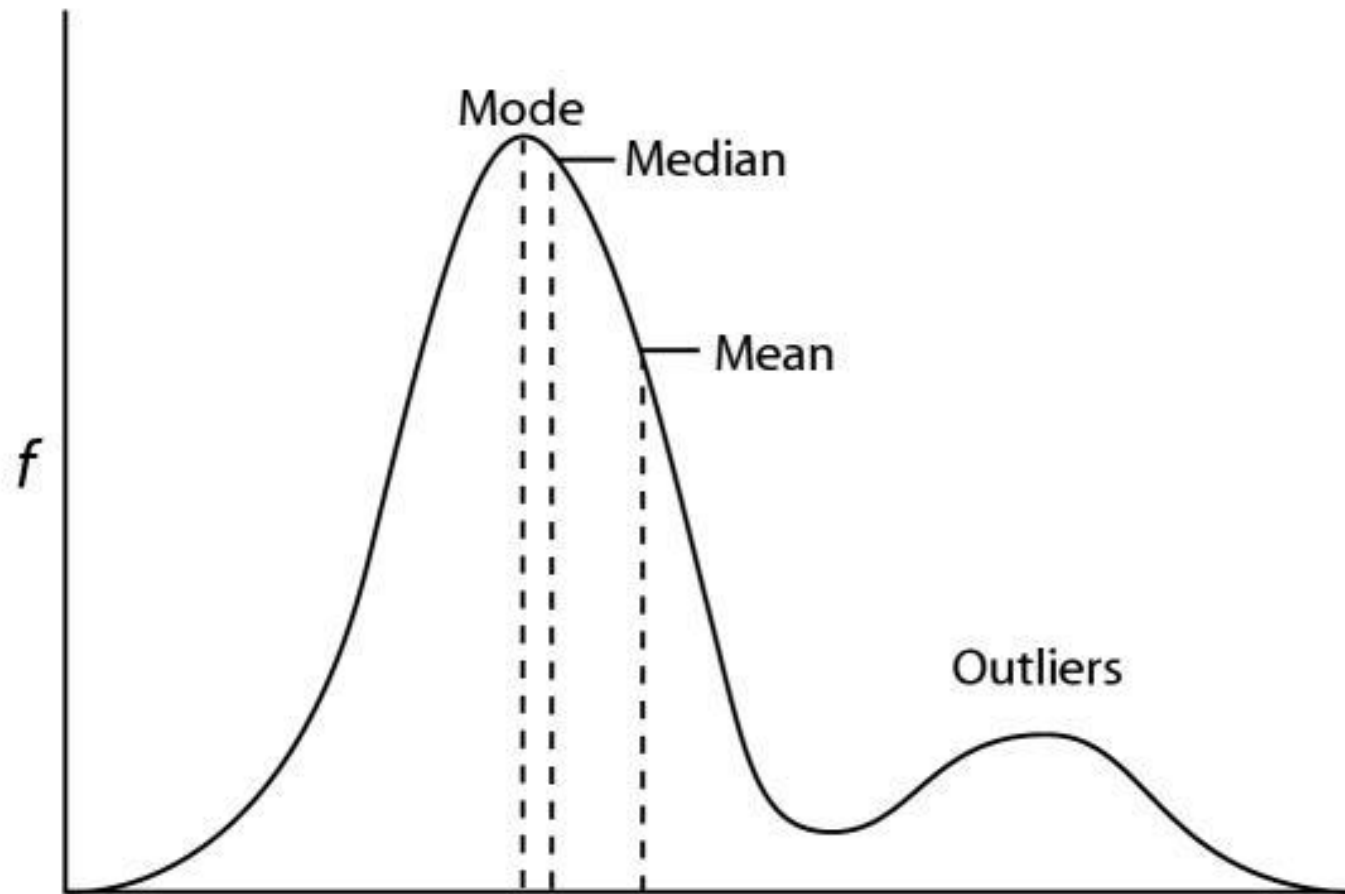
↓
8

↑
middle

Mode

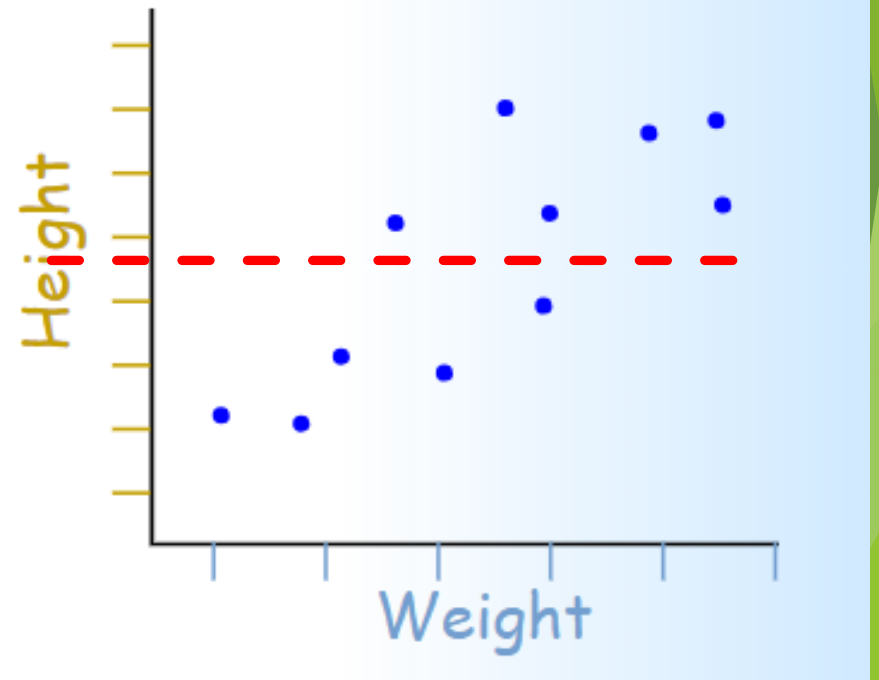
3, 7, 10, 8, 31, 10, 2

↓
10



Measures of Variability

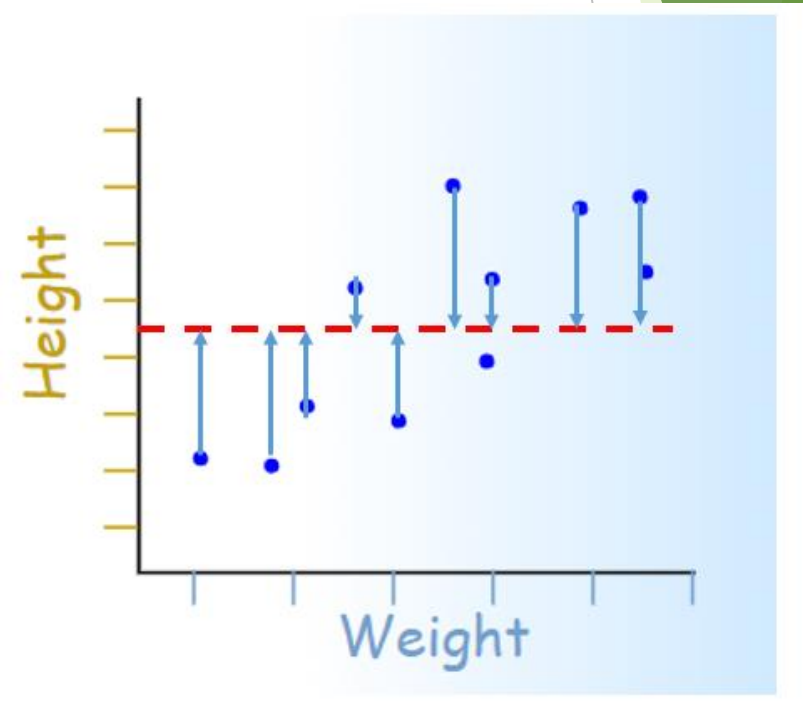
- They measure the **spread** of the data around the **mean**.
- The most commonly used measures of spread are:
 - Standard deviation (S)
 - Variance (S^2)



Measures of Spread (Standard deviation)

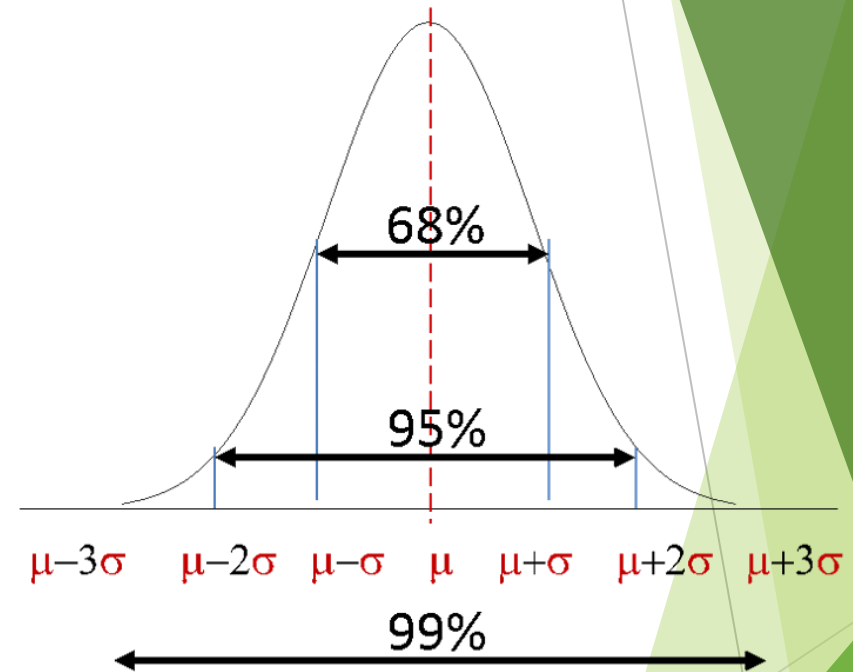
- ▶ **Standard deviation** is the square root of the average distance of the data points from the mean.
- ▶ Standard deviation measures the spread of a data distribution. The more spread out a data distribution is, the greater its standard deviation.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$



The Standard Normal Distribution


- Approximately 68% of values in the distribution are within 1 SD of the mean,
 - i.e., above or below.
 - $P(\mu - \sigma < X < \mu + \sigma) = 0.68$
- Approximately 95% of values in the distribution are within 2 SD of the mean.
 - $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$
- Approximately 99% of values in the distribution are within 3 SD of the mean.
 - $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.99$



Measures of Spread - Variance

- **Variance** is simply the square of standard deviation.
- Mathematically

$$s^2 = \frac{\sum (x - \bar{x})^2}{n}$$



54 77 67 68 46 64 62 56 38 Height (inches)

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} = \frac{(54 - 59.11)^2 + (77 - 59.11)^2 + \dots + (56 - 59.11)^2 + (38 - 59.11)^2}{9}$$

Variance = σ^2 = 127.43

Inferential Statistic

- ▶ Random sample of data taken from population to describe and make inferences about the population

Wake Up Call



What is the difference between Descriptive and Inferential Statistics

Inferential Statistics

- It gives the amount of uncertainty associated with the sample estimate.

Example:

- Suppose I want to know that what percentage of the total population in the US like football.

Ask each and every person about his/her liking

Approx. 325 million people



Take a sample of say 1000 people

Calc. how many out of 1000 like football

Generalize it for the entire population

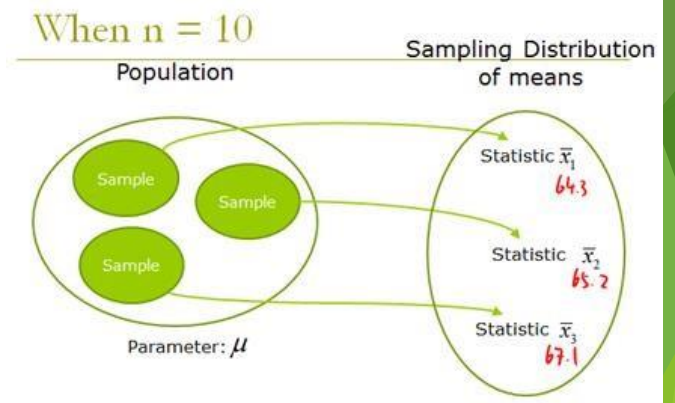
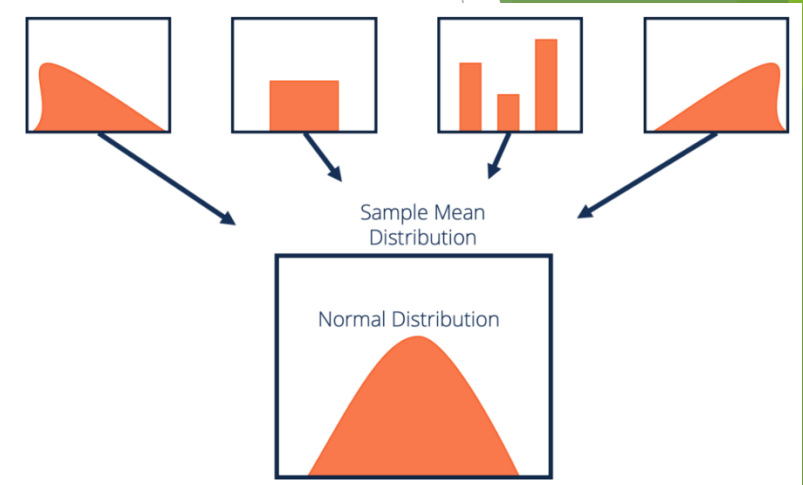
Not a good idea !

Inferential Statistics (Contd.)

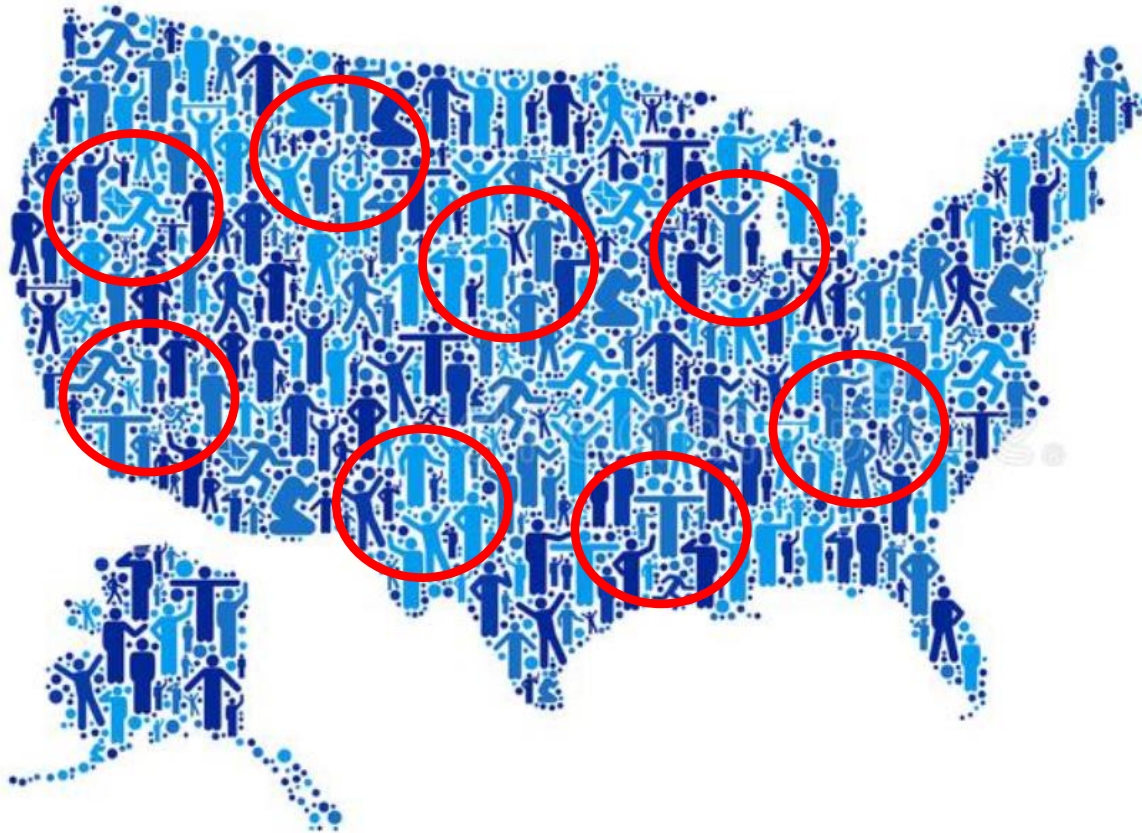


Central Limit Theorem

Take a large number of samples each of size 1000



Inferential Statistics (Contd.)



Take a large number of samples each of size 1000

Determine the percentage of people who like football

Plot a histogram

Confidence Interval

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$

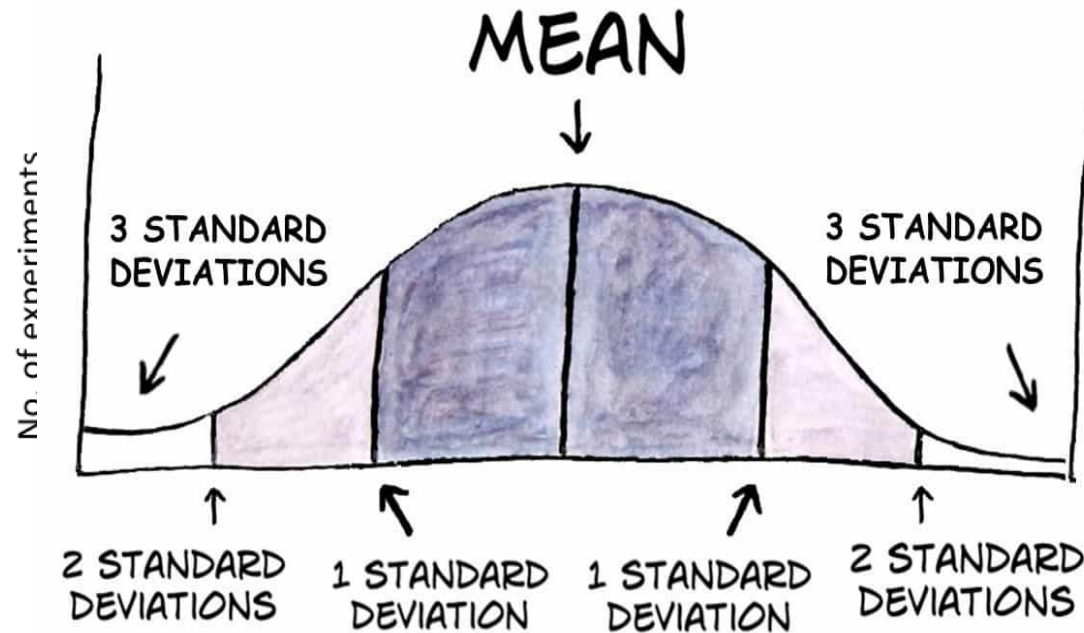
The histogram seem be follow a normal distribution

It can also be said with confidence that 95 % of the data points lie in the range of 62 (i.e. 65 - 3) and 68 (i.e. 65 + 3)

Thus it can be said with 95 % confidence that 62% – 68% of the **people in US like football**

The peak can be observed at 65%

Out of all the experiments that we conducted, maximum number of times the outcome was 65 %



Empirical Rule: 68-95-99.7

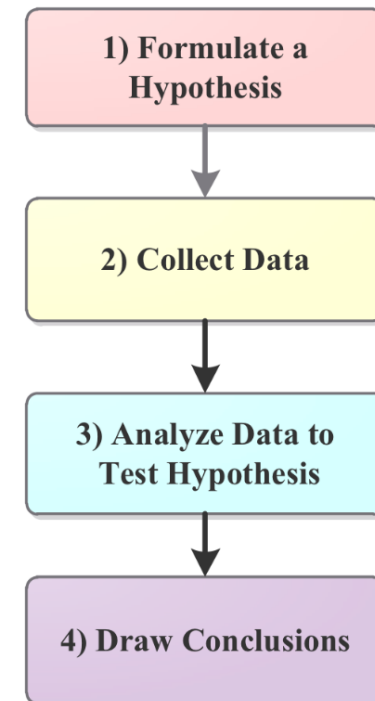
Outcome of each experiment

Hypothesis Testing

- It is used in scenarios where, analyzing the entire population is not possible due large size.
- It helps us to draw conclusions about the entire population by analyzing samples of relative smaller size.

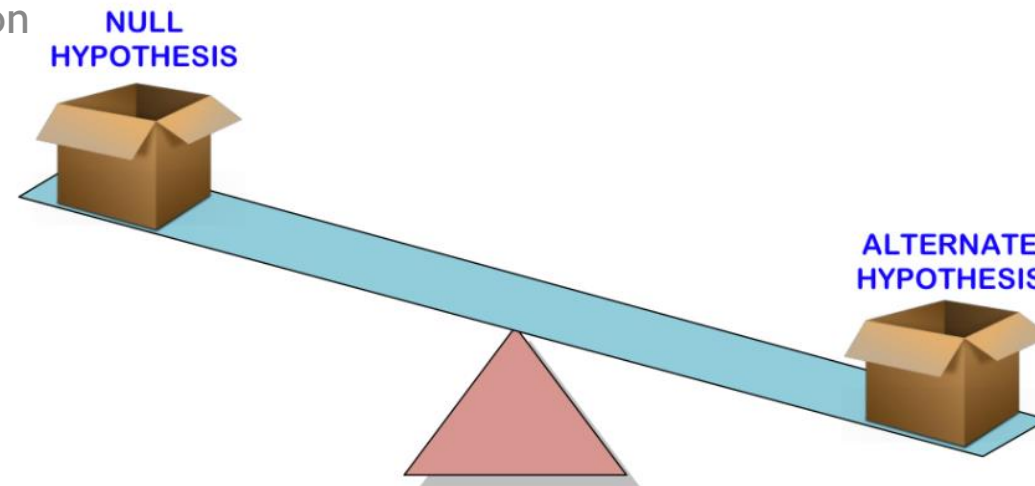
Hypothesis:

Changing my landing page color from black to blue will have a statistically significant impact on conversions



Terminologies

Null hypothesis changing my landing page color will have no impact on conversions.



Level of significance degree of significance in which we accept or reject the null-hypothesis.

Alternate hypothesis changing my landing page color will have significant impact on conversions.
{ what you are really aiming to prove through testing }

You want to disprove your null hypothesis in order to prove the alternate

Statistical Tests

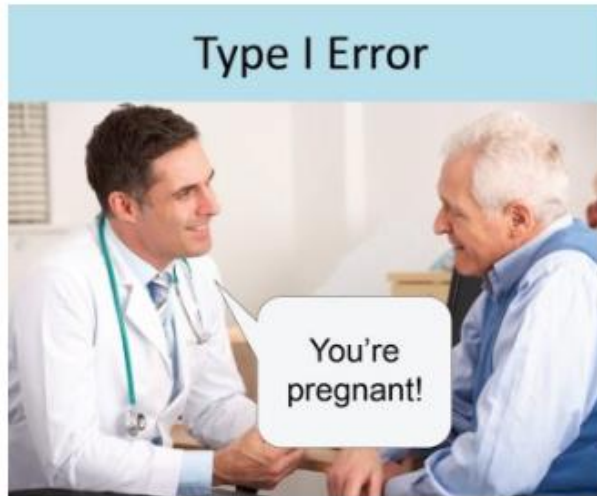
- T Test (Student T test)
 - one sampled t-test
 - two-sampled t-test.

Some Other Statistical Tests

- ▶ To except or reject the null hypothesis we have the following tests:
 - ▶ Correlation
 - ▶ Z-test
 - ▶ t-test
 - ▶ Chi-square test
 - ▶ ANOVA

Type I & Type II Errors

- ▶ Null Hypothesis : Person is **not** pregnant
- ▶ Alternate Hypothesis : Person is pregnant



Type I: Person is not Pregnant and we predict that the person is pregnant. We reject our Null Hypothesis when it is True



Type II: Person is pregnant and we predict its not Pregnant. We are failing to reject (accept) null hypothesis when it is False

Type I & Type II Errors

- In Statistics we define it by matrix

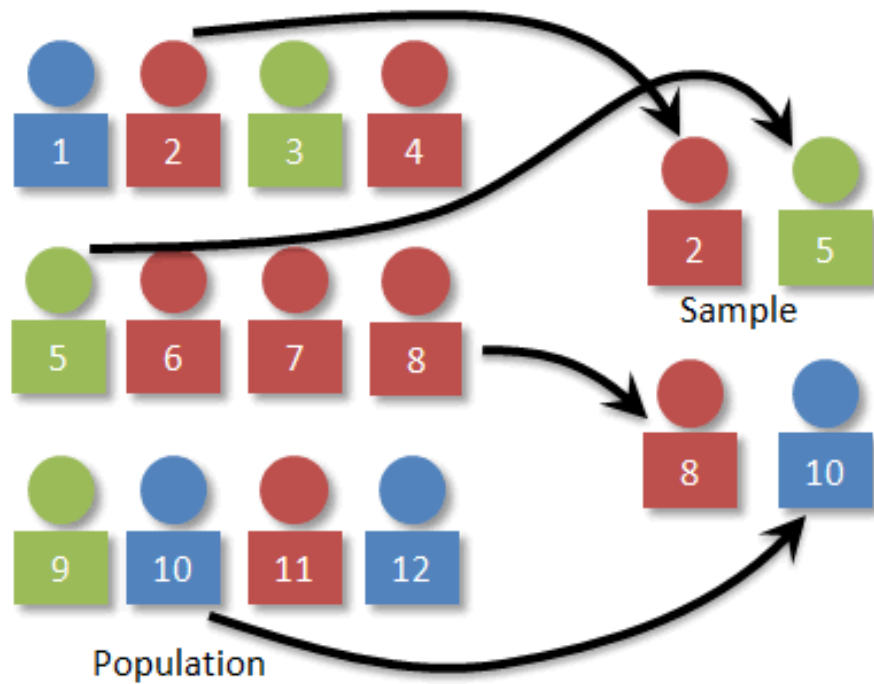
		Given the Null Hypothesis Is	
		True	False
Your Decision Based On a Random Sample	Reject	Type I Error	Correct Decision
	Do Not Reject	Correct Decision	Type II Error

Two Types of Errors in Decision Making

Sampling Techniques

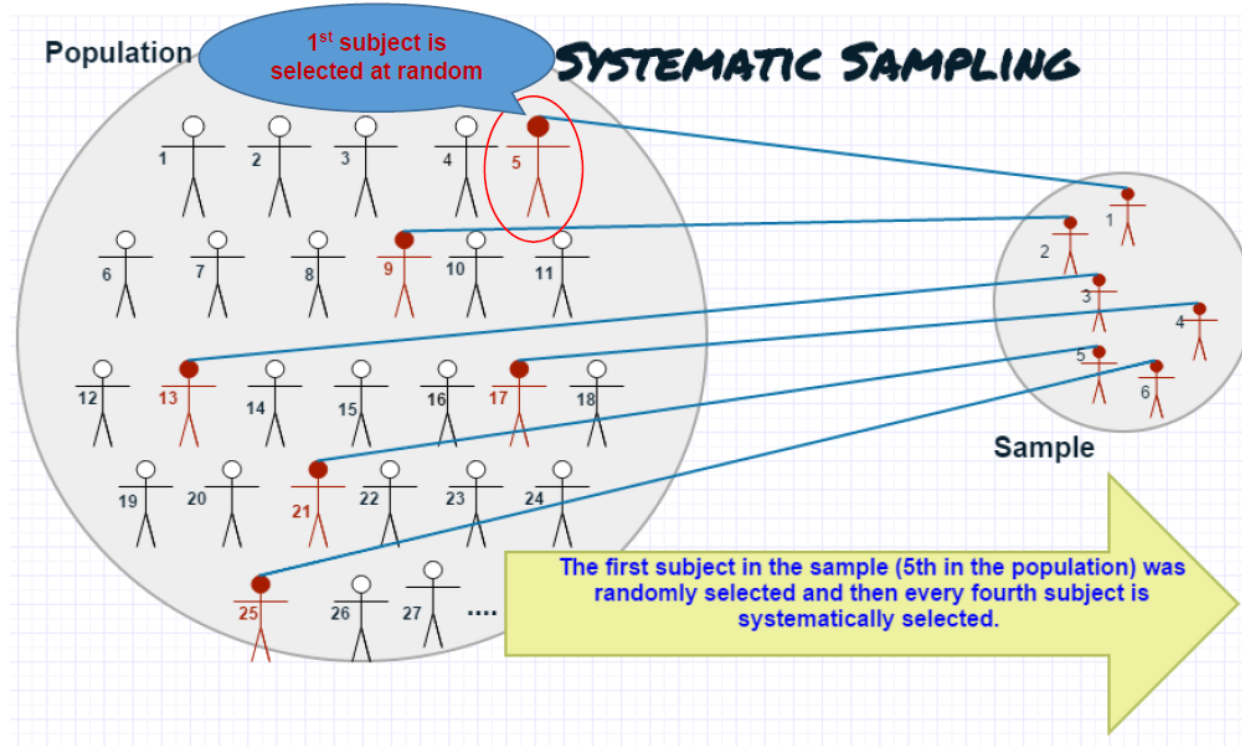
- ▶ Random sample of data taken from population to describe and make inferences about the population

Simple Random Sampling



When the only task is to select a much smaller subset of a large population

Systematic Sampling



sampling frame

ID	Name
1	Anthony Armadillo
2	Cathy Day
3	Ernesto Smith
4	Dina Tofaz
5	Mike Cox
6	Jim Hanano
7	Jerry Johnson
8	Ron Fowl
9	Wayne Cooper
10	Melissa White
11	Nathan Fox
12	Matt Lee
13	Juan Maxi
14	Rob Black
15	Mike Smith
16	Sodhi Mecheal

$N=16$
 $n=4$
 $K=N/n=4$

r =random number
between 1 and
4. In this case
 $r=3$.

The 3rd subject in
the sampling frame
is the first subject
in the systematic
sample.

systematic sample

ID	Name
3	Ernesto Smith
7	Jerry Johnson
11	Nathan Fox
15	Mike Smith

The sample is uniformly distributed over the sample frame

Stratified Sampling

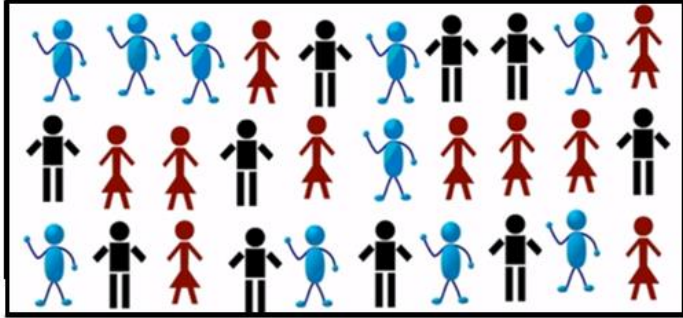


Fig. 1 Mixed population

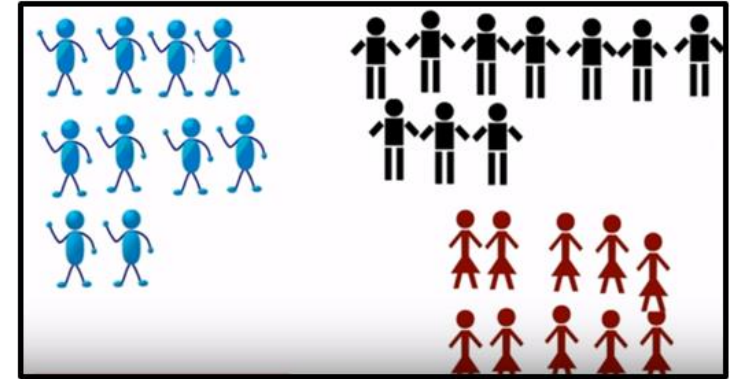


Fig. 2 Population divided into three sub-groups based on the color

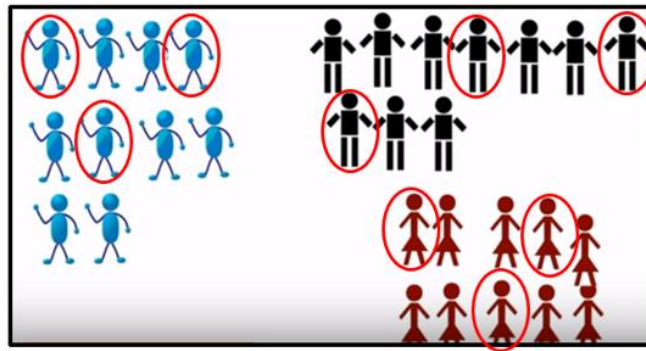
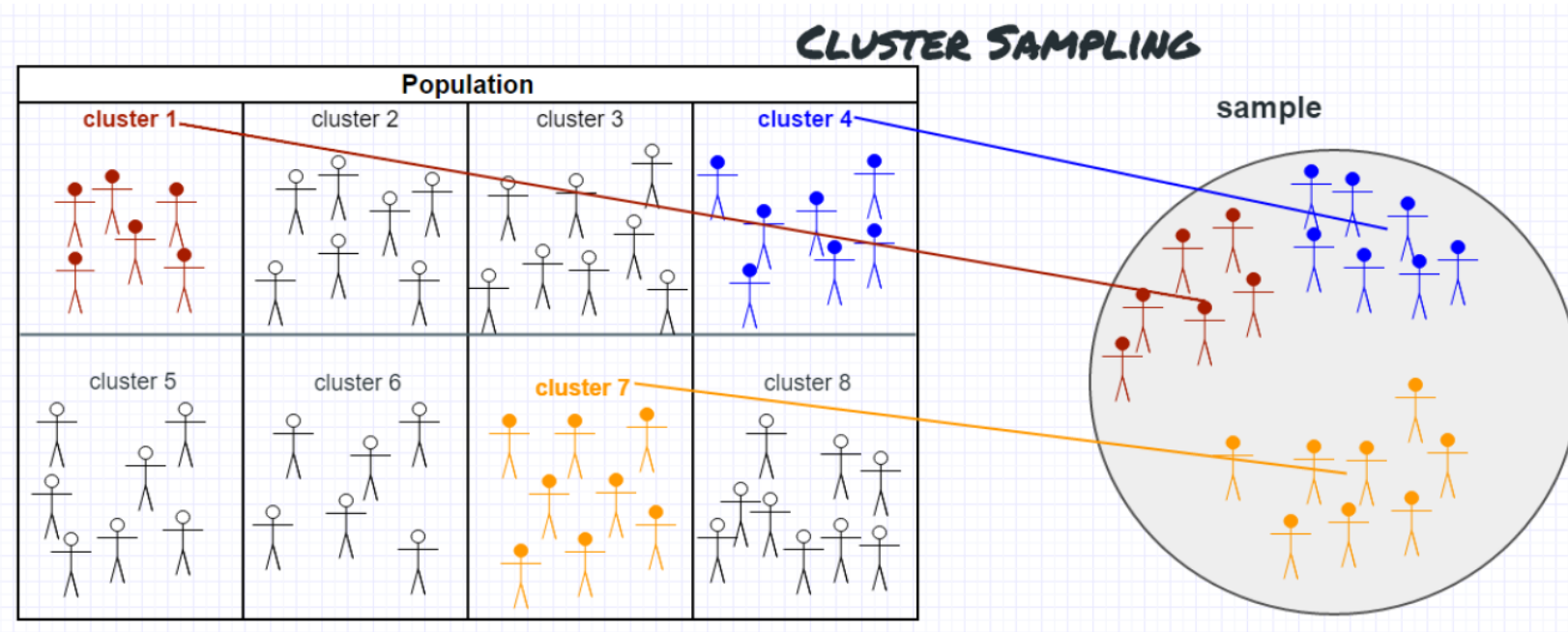


Fig. 3 Selecting datapoint from each subgroup

It is used in scenarios where it is very important to select samples from all sub group

Cluster Sampling



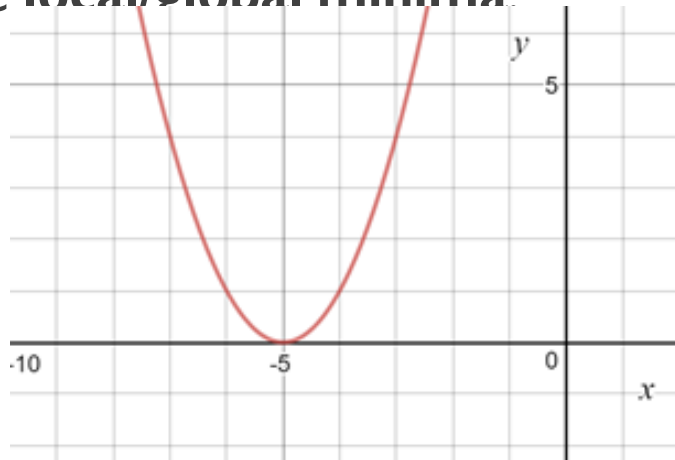
When stratified sampling cannot be applied due to lack of data understanding

Optimization

- ▶ Getting the optimal output for your problem

Optimization

- Optimization algorithm helps to find the minimum of a function.
- We move in the **negative direction** of the **gradient of the function** to reach the **local/global minima**.



Point $(x=-5, y=0)$

Hence $x=-5$ is the local and global minima of the function.

Gradient Descent

Goal: Local minima of the function $y=(x+5)^2$ starting from the point $x=3$

Gradient Descent

- The equation of a line is as follows:

$$\hat{y} = mx_i + b$$

slope of the line

bias of intercept

- For best fitting line we need to adjust 'm' and 'b' in such as way that the resulting should do **justice** with all the points in the sample space.
- It is used to find the best combination of m and b.



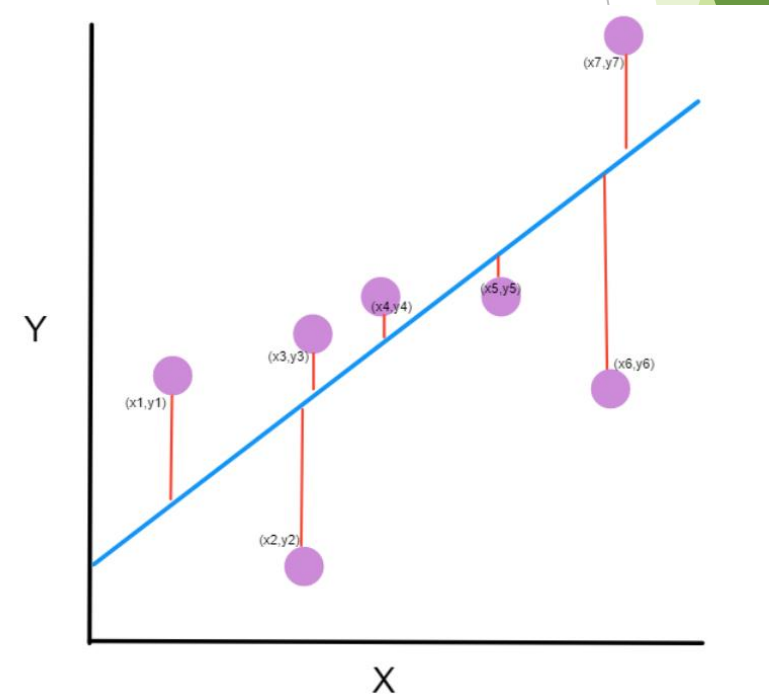
Cost function

- ▶ Gradient descent is an optimization algorithm that works on cost function.
- ▶

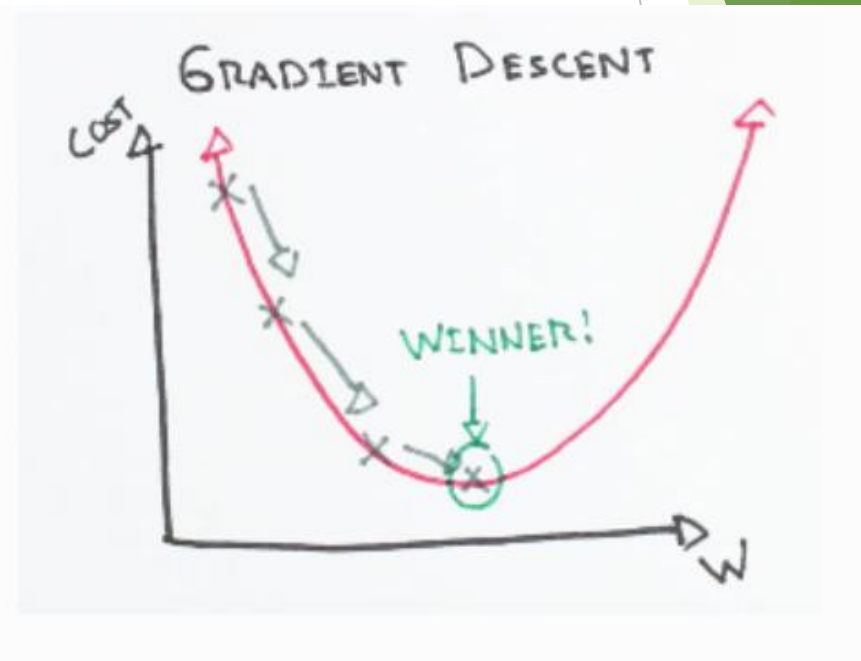
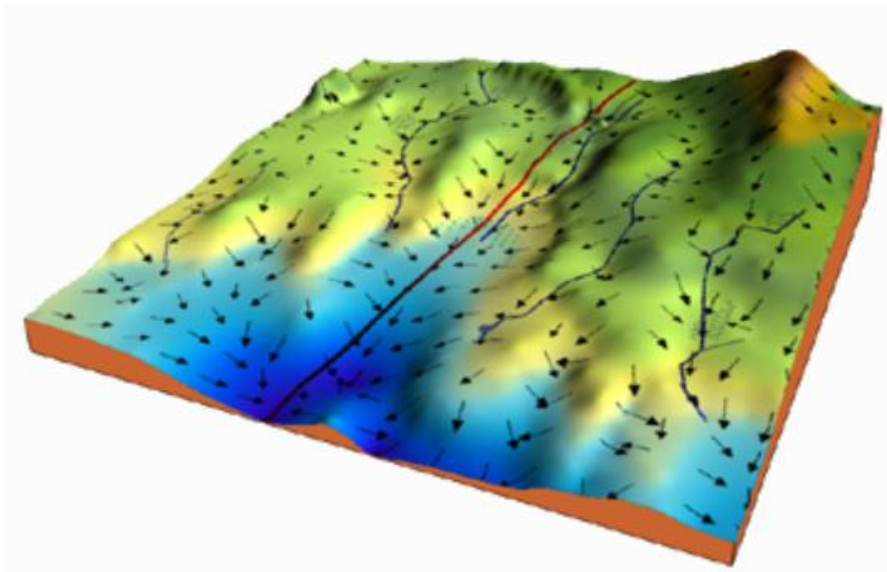
$$MSE = \frac{1}{N} \sum_{i=1}^n y_i - \hat{y}_i$$

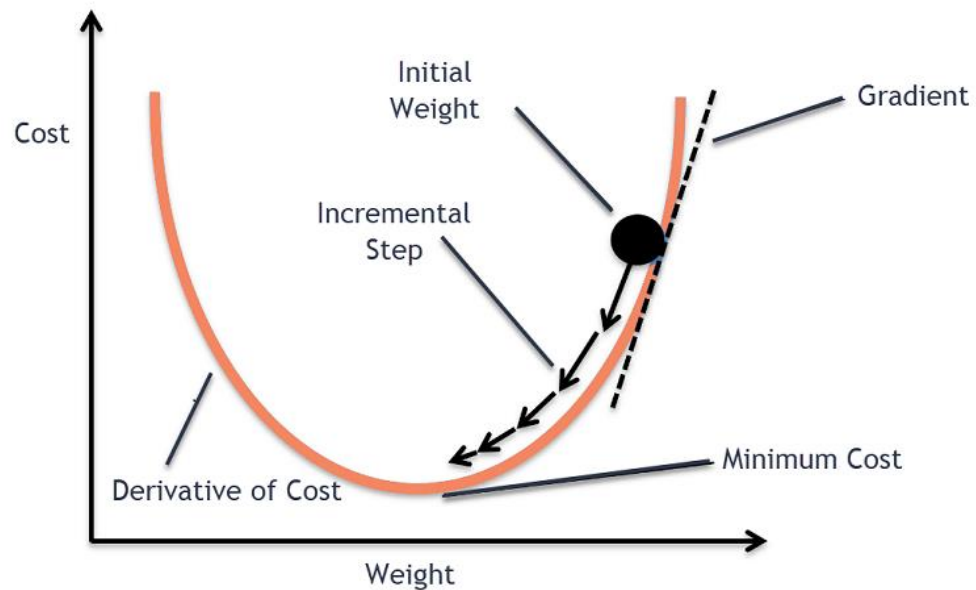
$$MSE = \frac{1}{N} \sum_{i=1}^n y_i - (mx_i + b)$$

$$Objective = \min(MSE)$$



Example





Hypothesis: $h_{\theta}(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}} J(\theta_0, \theta_1)$

Algorithm

Repeat {

$$m_{new} \rightarrow m_{old} \pm \alpha \frac{\partial (MSE)}{\partial m}$$

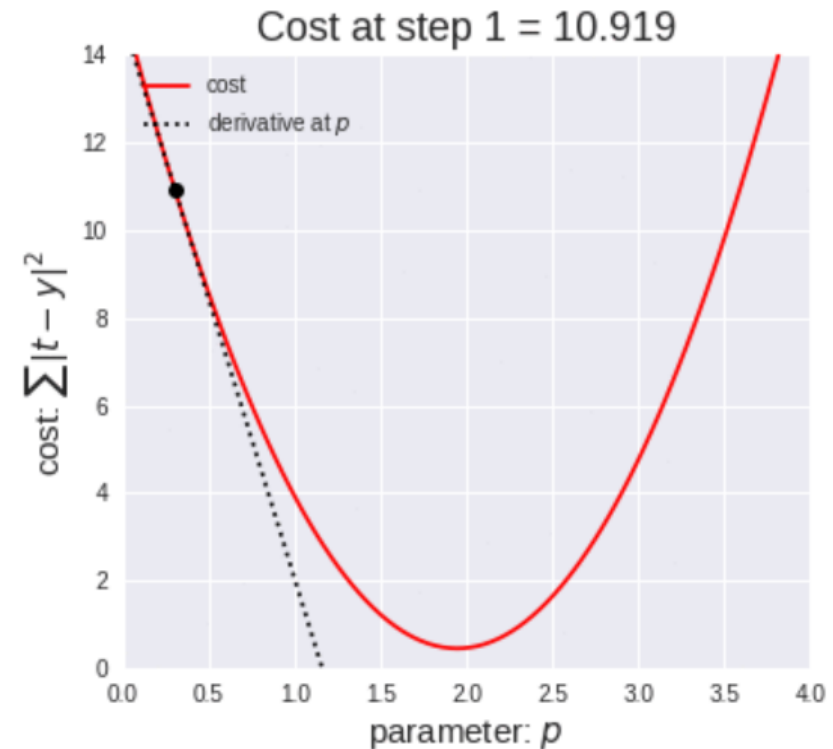
$$b_{new} \rightarrow b_{old} \pm \alpha \frac{\partial (MSE)}{\partial b}$$

}

Where α is the learning rate |

$\frac{\partial (MSE)}{\partial m}$, $\frac{\partial (MSE)}{\partial b}$ are the slope of the tangents to the function

So in every iteration we arrive at that value of m and b that take us towards the bottom



Learning Rate

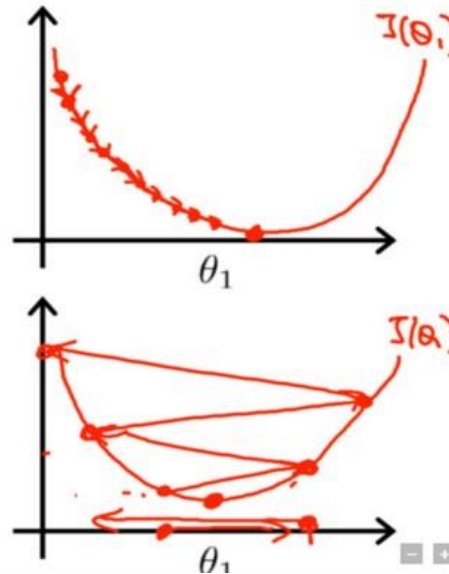
Learning rate is a hyper-parameter that controls how much we are adjusting the weights of our network with respect the loss gradient.

$\text{new_weight} = \text{existing_weight} - \text{learning_rate} * \text{gradient}$

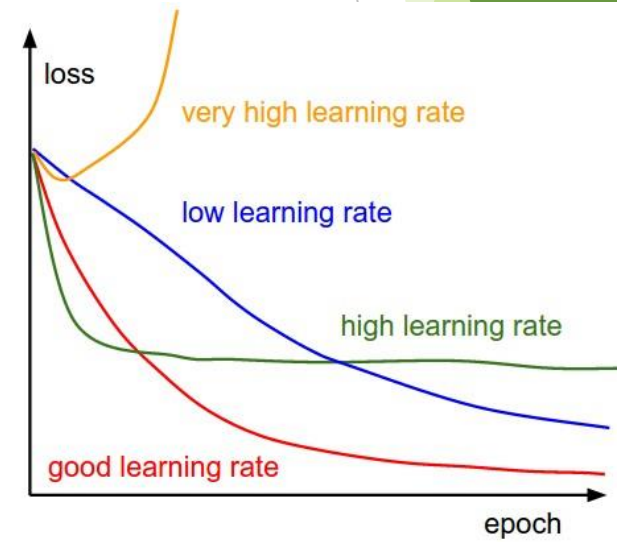
$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If α is too small, gradient descent can be slow.

If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.



The learning rate affects how quickly our model can converge to a local minima



Gradient descent with small (top) and large (bottom) learning rates.
Source: Andrew Ng's Machine Learning course on Coursera

Effect of various learning rates on convergence (Img Credit: cs231n)

Thank you

References

<https://www.math.arizona.edu/~jwatkins/statbook.pdf>

<https://towardsdatascience.com/confidence-intervals-explained-simply-for-data-scientists-8354a6e2266b>

<https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population>

<https://www.analyticsvidhya.com/blog/2020/10/how-does-the-gradient-descent-algorithm-work-in-machine-learning/>