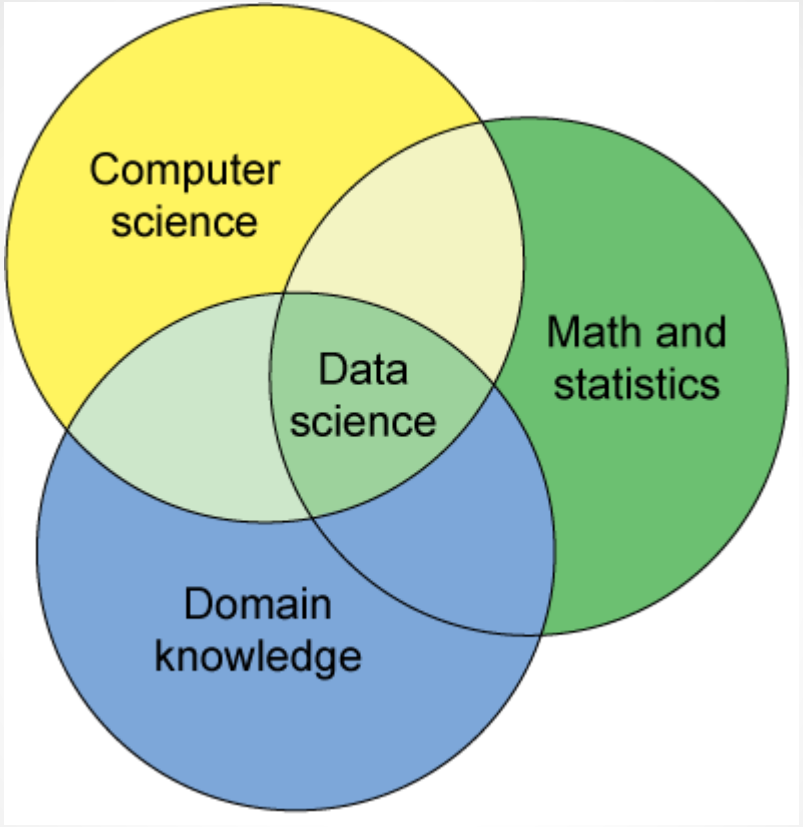
Statistics

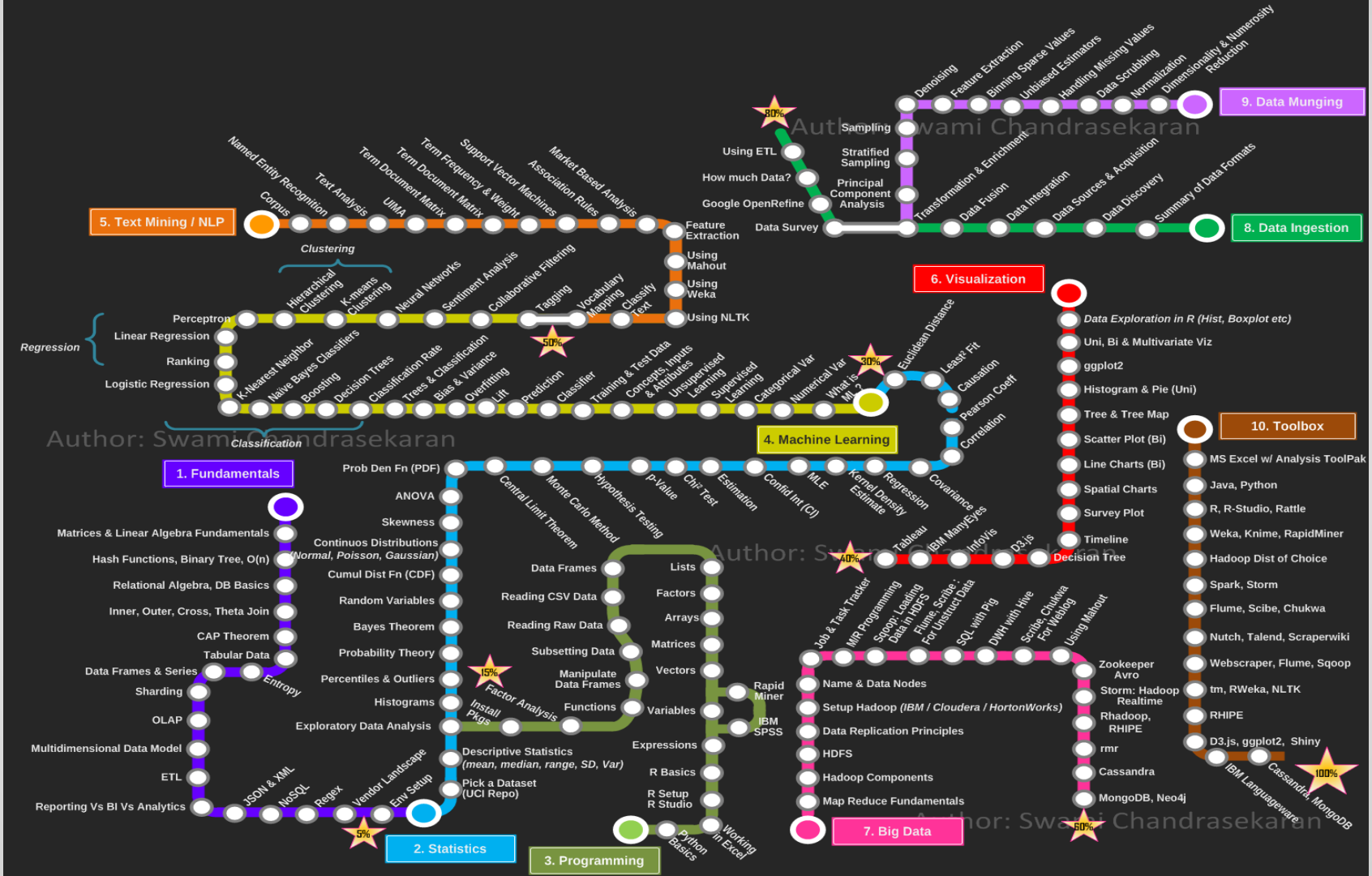
Gateway to Data Science



What is Data Science ?

How to Become a Data Scientist

Be Specific !!


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
Using NLTK
Using Weka
Using Mahout
Feature Extraction
Corpus
Named Entity Recognition
Text Analysis
Term Document Matrix
UIMA
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
5. Text Mining / NLP
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe: Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper
Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
80%
Using ETL
How much Data?
Google OpenRefine
Data Survey
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
9. Data Munging
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran

Why is Statistics Useful ?

Lets hear it from a Data Scientist at Twitter.

Objectives

1. Organize and summarize data
2. Reach inferences (sample --> population)

Statistics:

Descriptive statistics (1)

Inferential statistics (2)

Descriptive statistics

- Grouped data - The frequency distribution
- Measures of central tendency
- Measures of dispersion (variation, spread, scatter)
- Measures of position
- Measures of shape of distribution: graphs, skewness, kurtosis

Inferential statistics

- Estimation
- Hypothesis testing reaching a decision
  - Parametric statistics (z-test)
  - Non-parametric statistics << Distribution-free statistics
- Modeling, Predicting

Descriptive statistics

Boring stuff !

Frequency Distribution

Age:

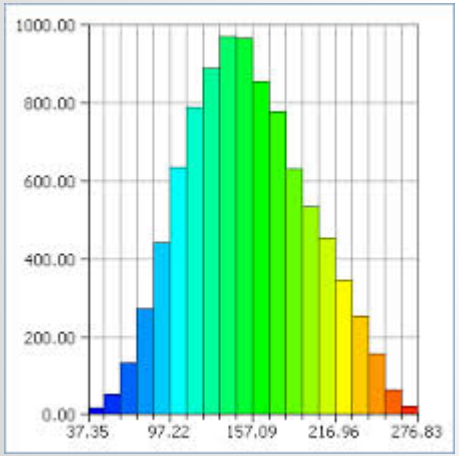
22, 20, 18, 23, 20, 25, 22, 20, 18, 20

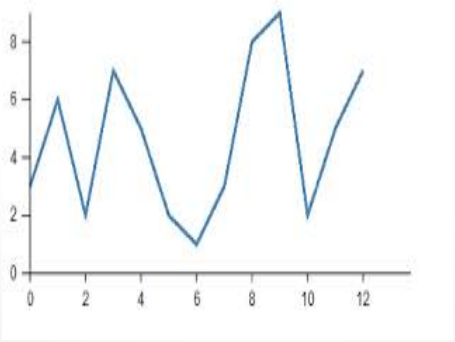
Age	Frequency	Relative Freq
18	2	0.2
20	4	0.4
22	2	0.2
23	1	0.1
25	1	0.1

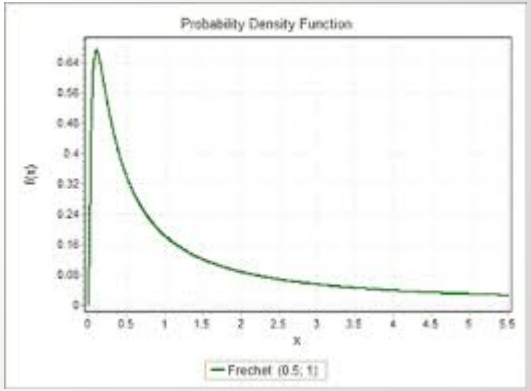
Distribution Visualization

- Histogram
- Density Plots

helps to understand Descriptive Statistics of data:







Effect of Bin Size

http://www.shodor.org/interactivate/activities/Histogram/

- Big Bin size : Loses details
- Small Bin Size: Too much detail

What is a Normal Distribution

Is This Normal ?

Lets Create our Own Distribution

Enter marks in 5 subjects in any of your exams.

Setup environment for Exercises in R

Visualizing Data In R

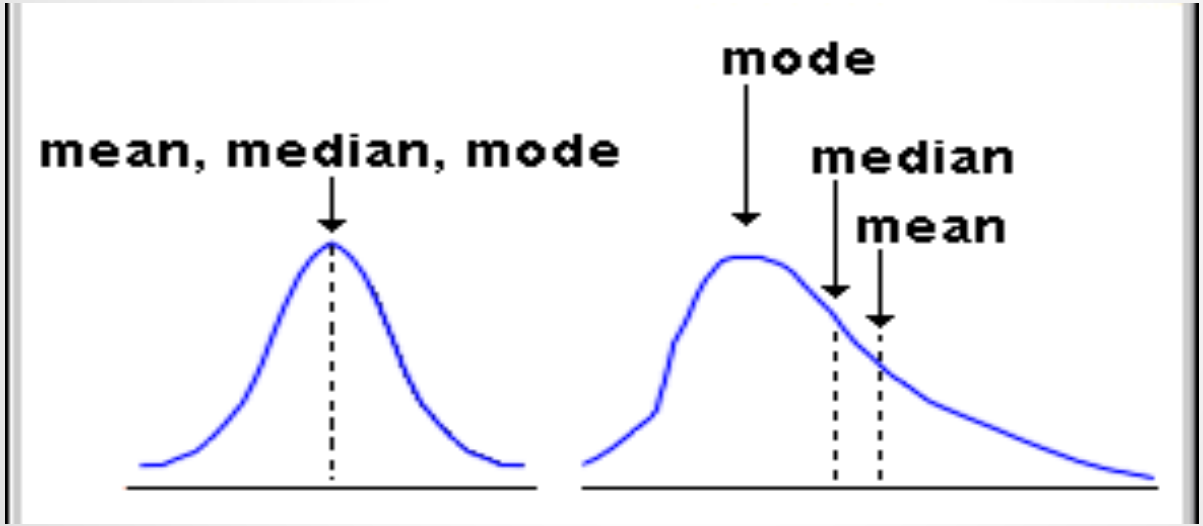
Exercise-1

Measures of central tendency

- The Mean
- The Median (Md)
- The Midrange (Mr)
- Mode (Mo)

1. What should be used as measure of central tendency ?
2. How does outliers affect Mean, median, mode ?
3. Relation between mean, median, mode in case of Normal Distribution.

Effect of Skewness on Measures of Central Tendency



Measures of Dispersion

- Range
- Variance
- Standard Deviation

$$\text{variance} = \sigma^2 = \frac{\sum (x_r - \mu)^2}{n}$$

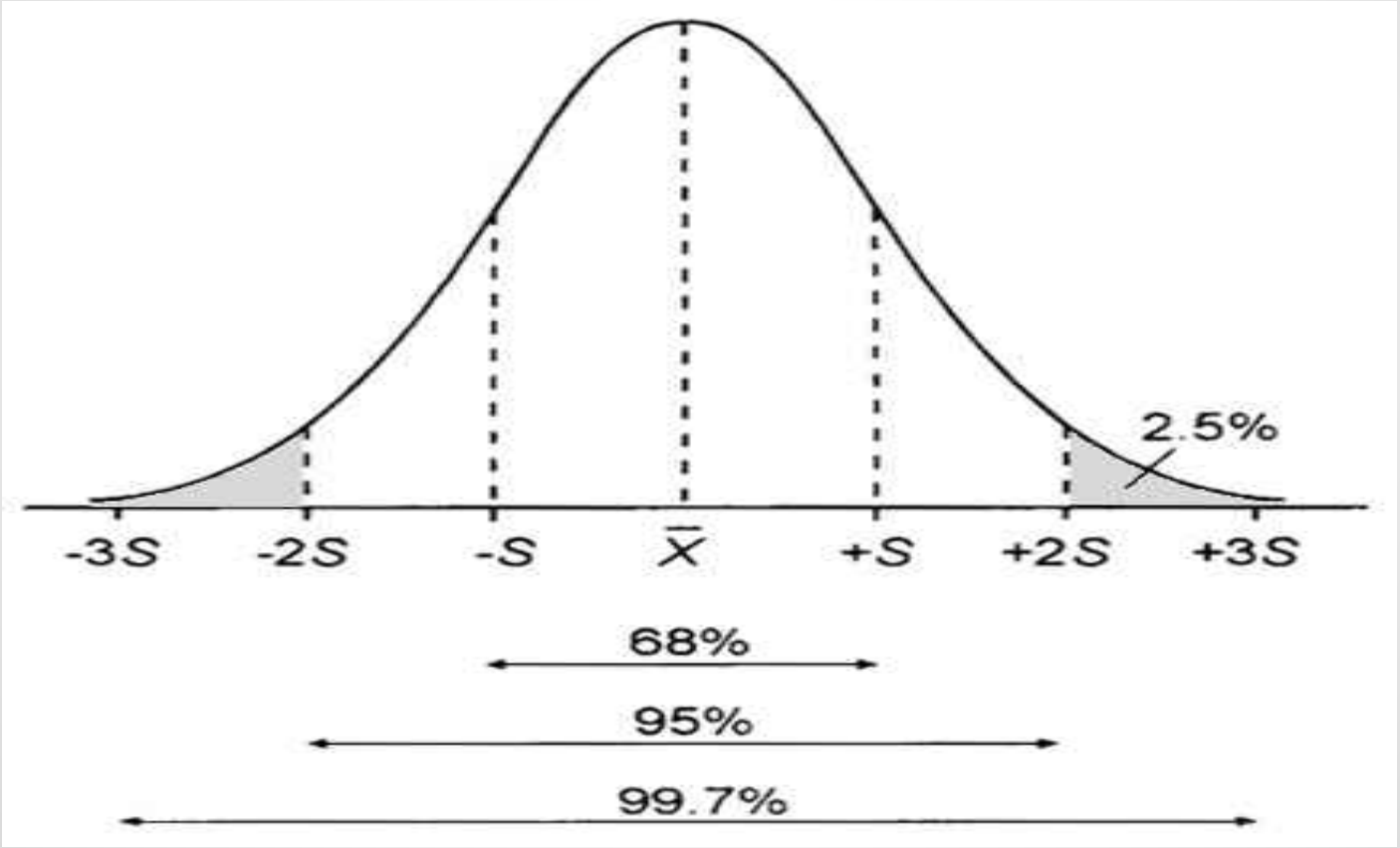
$$\text{standard deviation} \quad \sigma = \sqrt{\frac{\sum (x_r - \mu)^2}{n}}$$

μ = mean

What is so great about standard deviation ?

Let us see through Exercise-2.

Probability Density Function



Standardizing Distribution

Who is More Popular ?

Popularity Check

Person A:

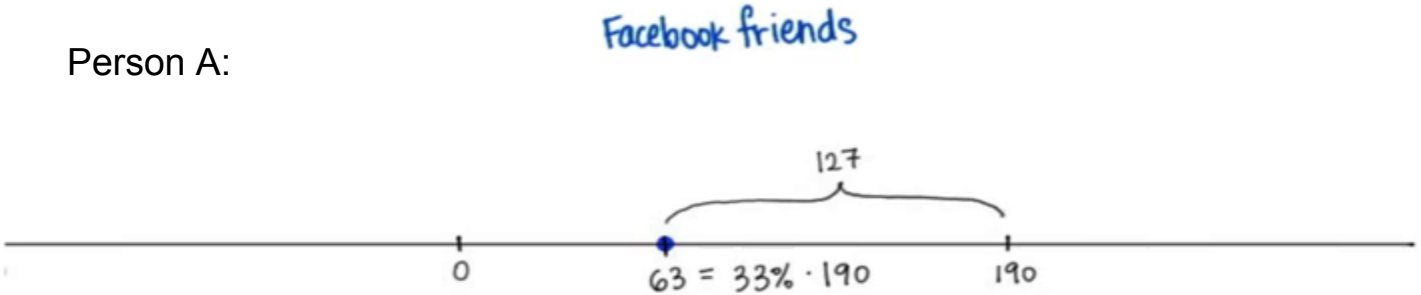
- uses only Facebook
- has 63 Friends

Person B:

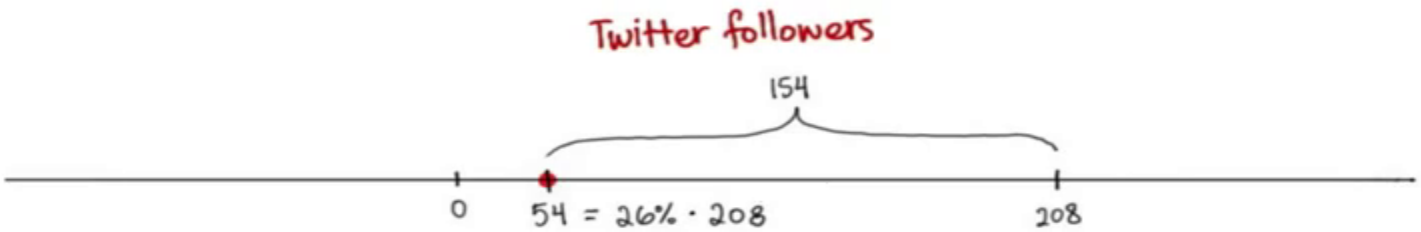
- uses only Twitter
- has 54 followers

	Mean	Standard Deviation
Facebook	190	36
Twitter	208	60

Person A:



Person B:

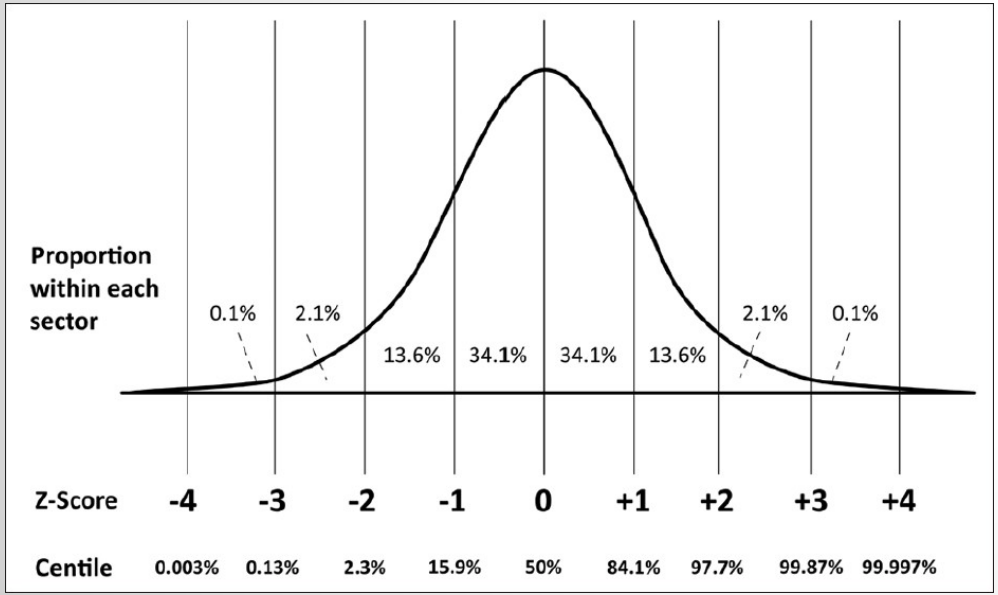


Compare Populations

Run Exercise-3

z-score

$$z = \frac{X - \overline{X}}{S}$$



- z-score is the number of standard deviations a value of sample x is from the mean.
- if mean = 0 and standard deviation = 1, then z-score = x.
- Can be used to compare two samples from complete different populations.
- Replacing every value of X with its z-score is called Standardizing the distribution.
- This new Standard Normal Distribution has mean 0 and standard deviation 1

So Who is More Popular ?

Sampling Distribution

Suppose that we draw all possible samples of size n from a given population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a sampling distribution.

Central Limit Theorem

The central limit theorem states that the sampling distribution of any statistic will be normal or nearly normal, if the sample size is large enough.

If n is large then the mean and variance of the sample mean $(\bar{X})$ are :

Mean $= \mu$

Variance $= \frac{\sigma^2}{n}$

n = Sample size

μ = Mean of the population

σ = Standard deviation of the population

Lets see CLT working

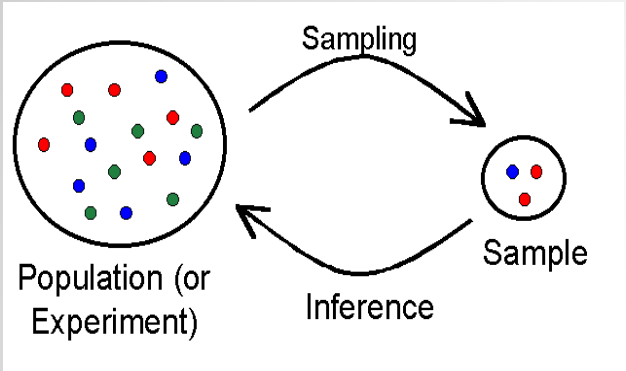
http://www.socr.ucla.edu/applets.dir/samplingdistributionapplet.html

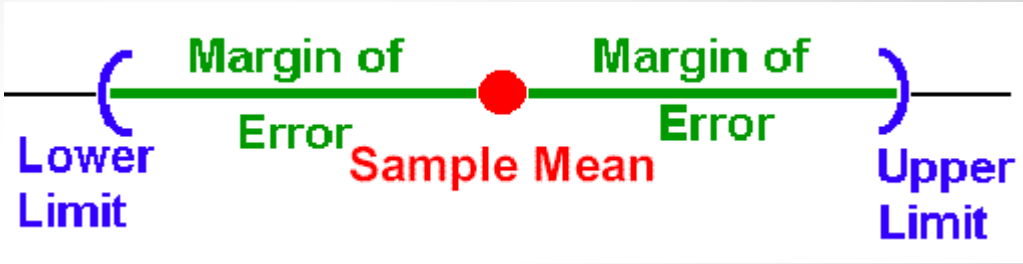
Inferential Statistics

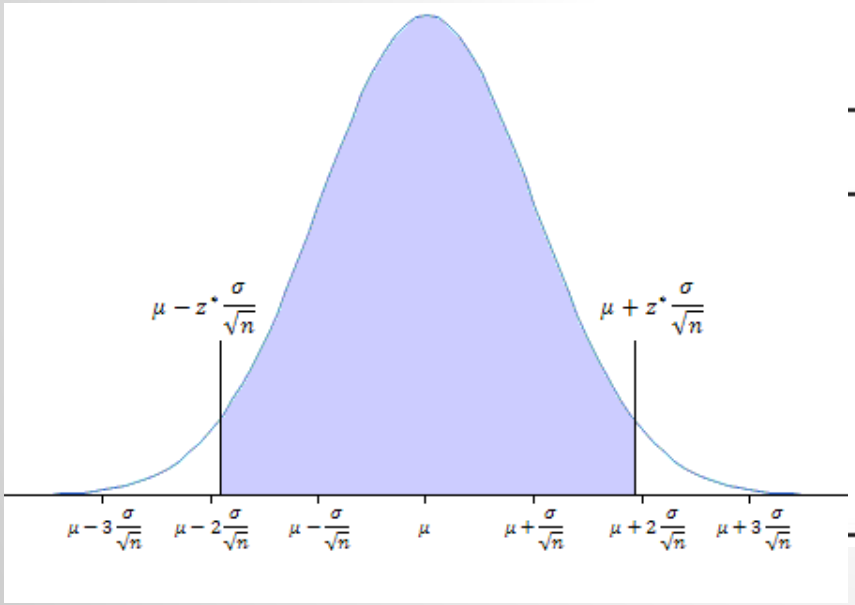
Estimation

- The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic.

- Point Estimate
- Confidence Interval
- Margin of Error
- Alpha level
- p-Value







Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291

Hypothesis Testing

- Hypothesis testing is designed to detect significant differences: differences that did not occur by random chance.

- In the “one sample” case: we compare a random sample (from a large group) to a population.

- We compare a sample statistic to a population parameter to see if there is a significant difference.

The Null and Alternative Hypotheses:

- **Null Hypothesis (H_0)**
  The difference is caused by random chance.
  The H_0 always states there is “no significant difference.”

- **Alternative hypothesis (H_1)**
  “The difference is real”.
  (H_1) always contradicts the H_0.

One (and only one) of these explanations must be true. Which one?

Test the Explanations

- We always test the Null Hypothesis.
- Assuming that the H_0 is true:

What is the probability (p-value) of getting the difference between the means due to random chance.

If the probability associated with this difference is less than an alpha level, reject the null hypothesis.

Test the Hypotheses

- Use the .05 value as a guideline to identify differences that would be rare or extremely unlikely if H_0 is true. This “alpha” value delineates the “region of rejection.”

- Use the *Z*-score formula for single samples and determine the probability of getting the observed difference just by chance.

- If the probability is less than .05, the calculated or “observed” *Z* score will be beyond ±1.96 (the “critical” *Z* score).

Lets get REAL !

Exercise-4

t-Test

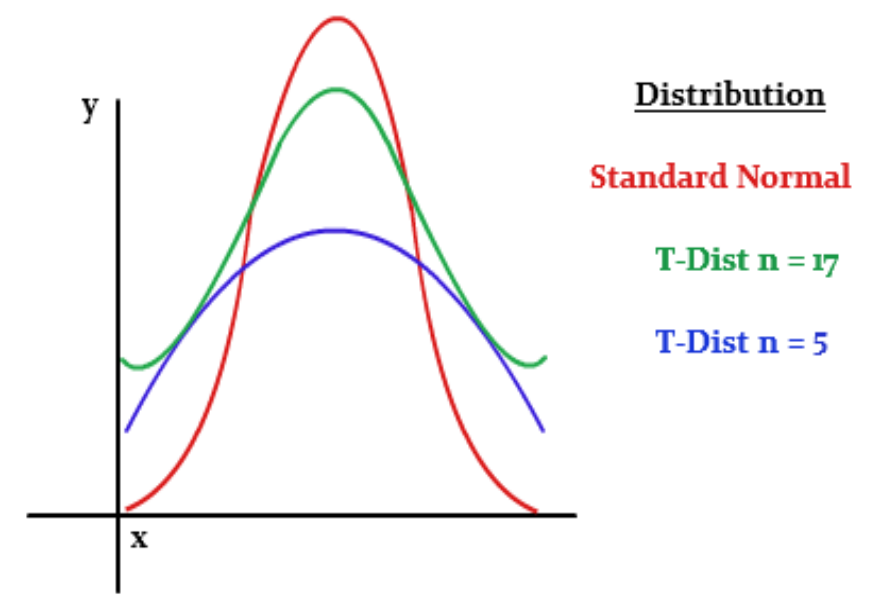
I do not know about the Population Variance

- We may not know the mean and variance of some populations, which means we cannot do a Z-Test. In this case, we use a T-test, Student's T to be specific, for use with a single group or sample of data.

- The standard error of the sampling distribution of the sample means is estimated.

- A **'t-distribution'** (not normal curve) is used to create confidence intervals.

t-Distribution



- Very similar to the Z distribution by assuming normality.
- Normality is obtained after about 100 data observations.
- Basic rule of parameter estimation: the higher the obs (N) of sample the more reflective of overall population.

One Sample t-Test

- Formula:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

where $S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$

- Using R:

```
t.test(y,mu=3) # H0: mu=3
```

Demo

Exercise-5

Two Sample t-Tests

Paired

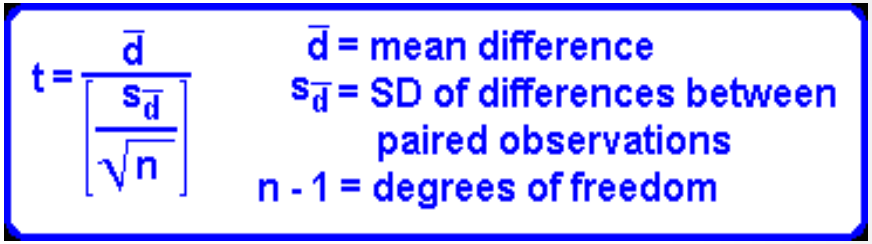
- to examine a single sample subjects/units under two conditions, such as pretest - posttest experiment.
- **repeated measures design**
- **Ex- effect of treatment.**

Independent

- used to compare groups of participants that are not related in any way.
- **between subject designs**
- **Ex- Comparison based on gender, Age group etc.**

Two Sample t-Test (Paired)

- Formula:



- Using R:

```
t.test(y1,y2,Paired=TRUE)
```

Two Sample t-Test (Independent)

- Formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- Using R:

```
t.test(y1,y2)
```

Demo

Exercise-6

ANOVA

More Than Two Samples to Compare

The basic ANOVA situation

Two variables: 1. Categorical, 1. Quantitative

Main Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

If categorical variable has only 2 values: 2-sample t-test

ANOVA allows for 3 or more groups

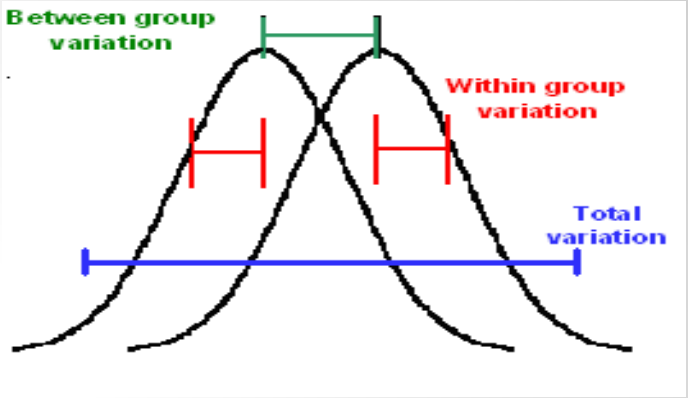
What does ANOVA do?

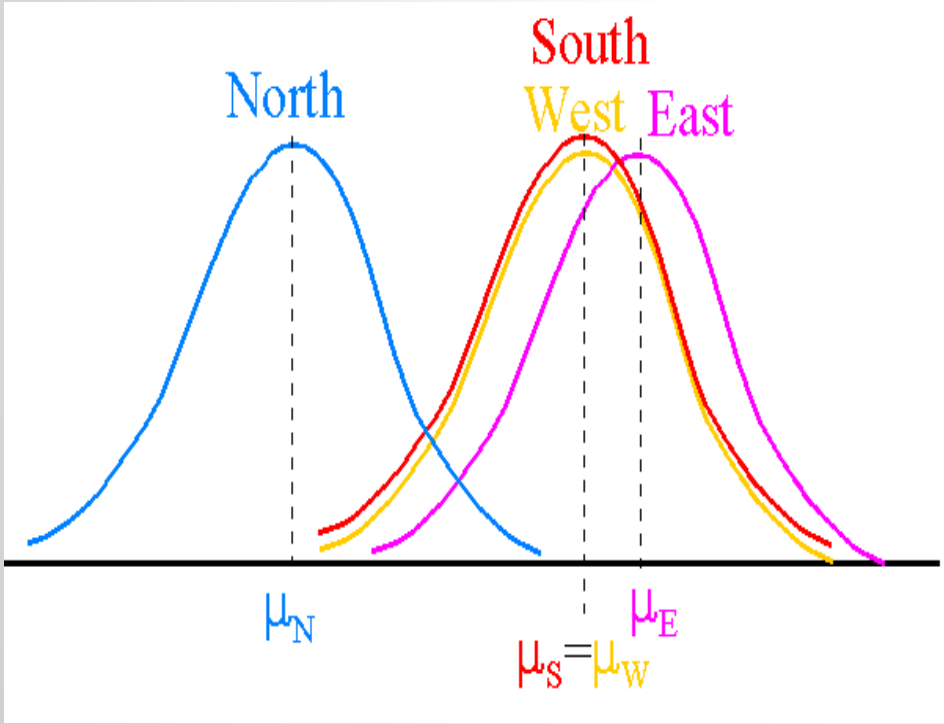
At its simplest (there are extensions) ANOVA tests the following hypotheses:

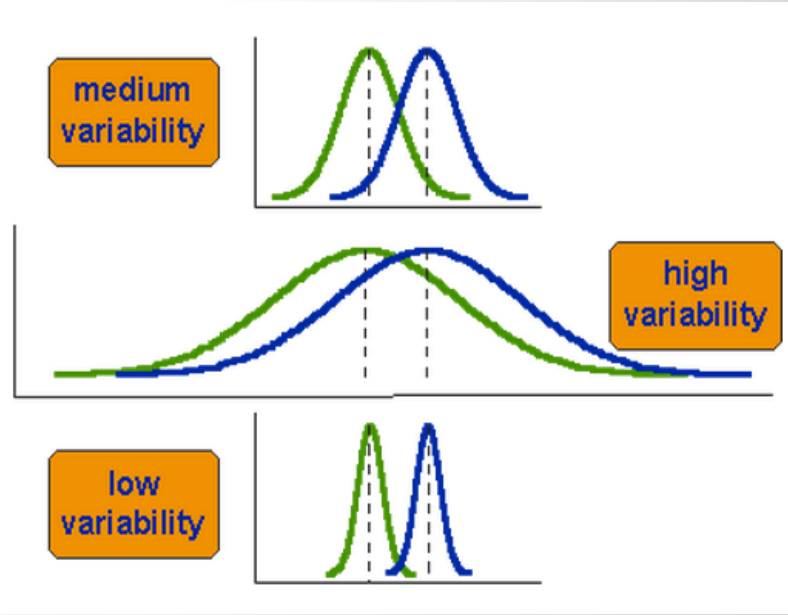
- H_0: The means of all the groups are equal.
- H_a: Not all the means are equal doesn't say how or which ones differ.
  Can follow up with "multiple comparisons"

Note: we usually refer to the sub-populations as "groups" when doing ANOVA.

Effect of Variabilities







How ANOVA works

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

$$SS_{between} = \sum_{j=1}^{p} n_j(\overline{x}_j - \overline{x})^2$$

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

$$SS_{within} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

F-statistic

$$F_s = \frac{MS_{(between)}}{MS_{(within)}}$$

A large F is evidence against H_0, since it indicates that there is more difference between groups than within groups.

Tukey's HSD for Multiple Comparison

HSD means honestly significant difference.

$$HSD = q\sqrt{\frac{MS_{within}}{n}}$$

q	->	Is a value from a table of the studentized range statistic based on alpha, df_{within} and k, the number of groups.
MS_{within}	->	Is the mean square within groups.
n	->	sample size

Multiple Comparison

To see which means are significantly different, we compare the observed differences among our means to the critical value of the Tukey test (HSD).

If mean of two samples is greater than HSD, then they are significantly different.

In R :

```
> fit <- aov(x ~ A, data)
> summary(fit)
> TukeyHSD(fit)
```

Demo

Exercise-7

Correlation

Data Science is Here

Correlation

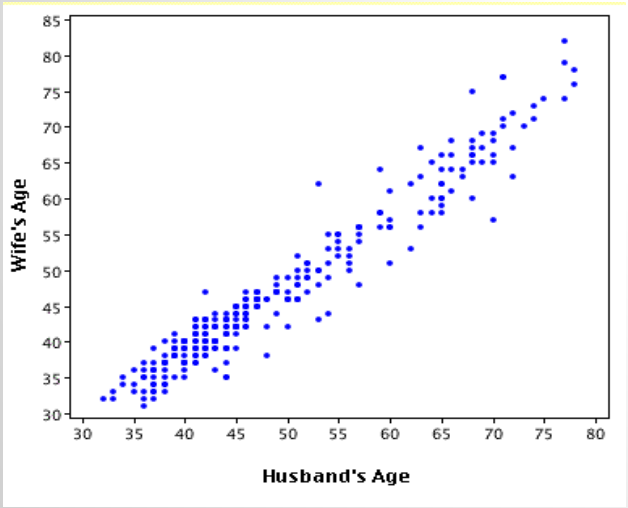
Correlation is a statistical technique used to determine the degree to which two variables are related.

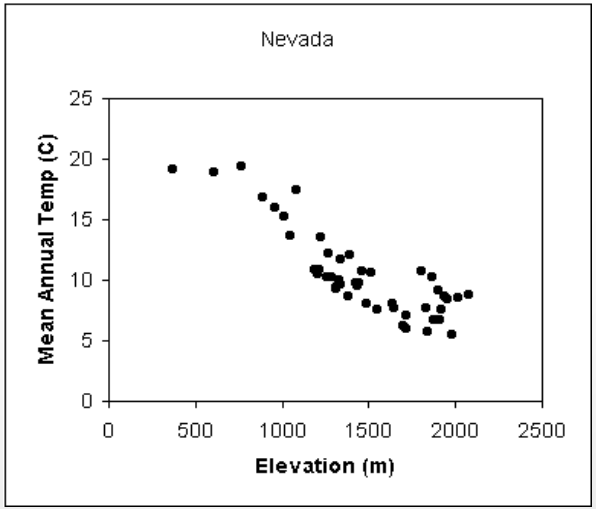
Usage:

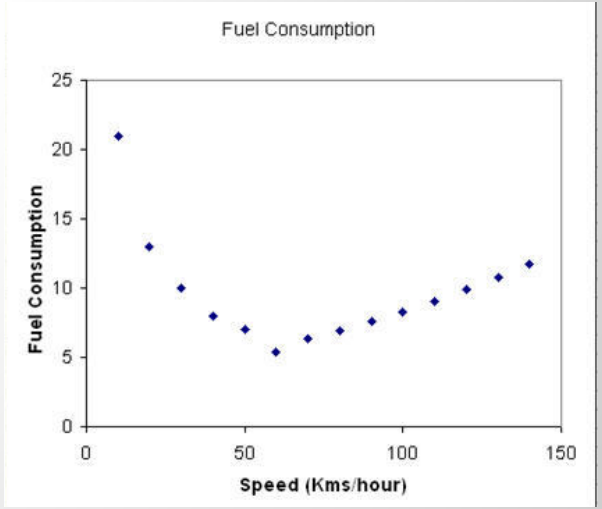
- is there a relationship
- if so, what is the equation
- use the equation for prediction

Scatter Plot

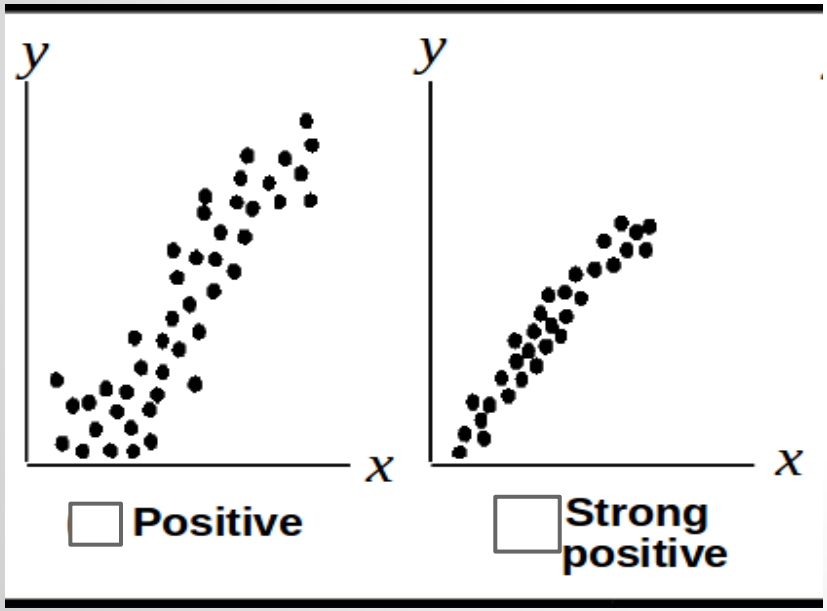
A graph in which the paired (x,y) sample data are plotted with a horizontal x axis and a vertical y axis. Each individual (x,y) pair is plotted as a single point.

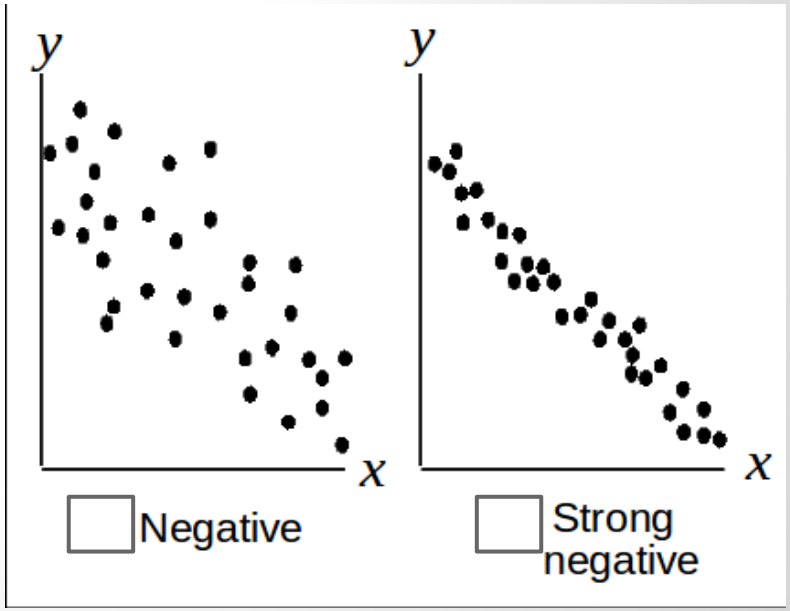




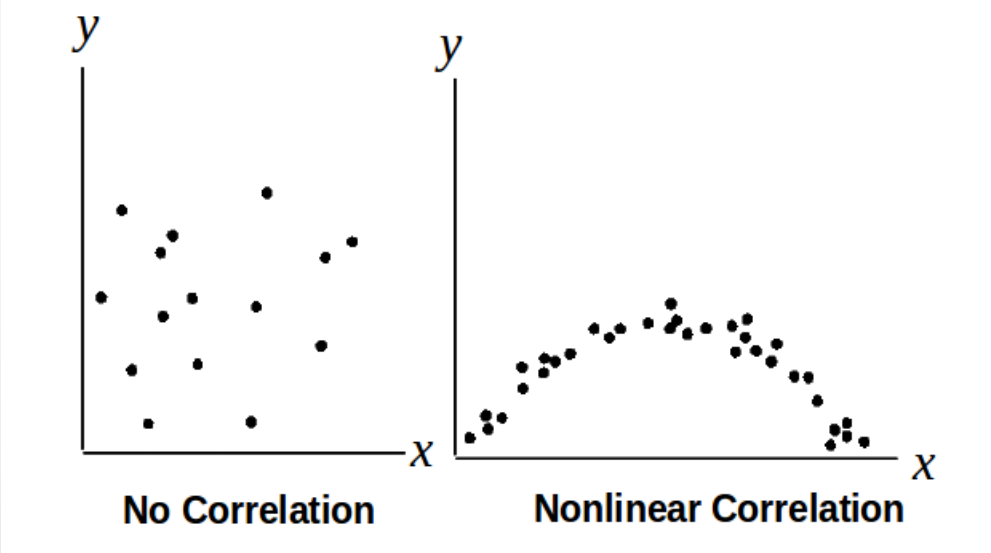


Types of Correlations





Types of Correlations



Correlation Coefficient (r)

- It is also called Pearson's correlation coefficient.
- It measures the nature and strength between two variables of the quantitative type.

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

- The sign of r denotes the nature of association .
- while the value of r denotes the strength of association.
- The value of r ranges between (-1) and (+1)
- r^2 - the proportion of response variation "explained" by the regressors in the model

Hypothesis Testing for Correlation

To determine whether there is a significant linear correlation between two variables:

Let ρ be the population's correlation coefficient.

The null hypothesis is:

$$H_0 : \rho = 0$$

The test statistic is:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

If the null hypothesis is true the test statistic follows the t distribution with n-2 degrees of freedom or $T \sim t(n-2)$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \rho < 0$	$T < -t_{1-\alpha}$
$H_1 : \rho > 0$	$T > t_{1-\alpha}$
$H_1 : \rho \neq 0$	$T < -t_{1-\alpha/2}$ or $T > t_{\alpha/2}$

Common Errors Involving Correlation

- Causation: It is incorrect to conclude that correlation implies causation.

- Averages: Averages suppress individual variation and may inflate the correlation coefficient.

- Linearity: There may be some relationship between x and y even when there is no significant linear correlation.

Demo

Is the correlation significant ?
Exercise-8

Regression

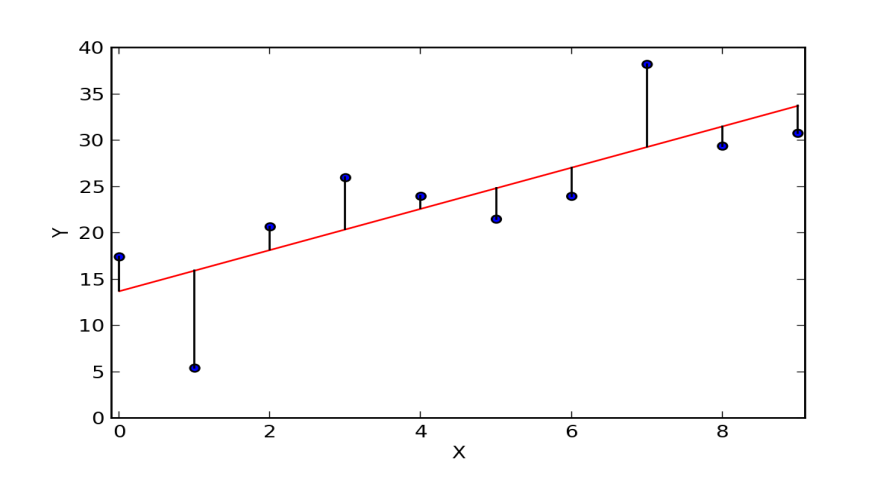
Lets Predict !

Linear Regression

- Uses a variable (x) to predict some outcome variable (y)
- Tells you how values in y change as a function of changes in values of x
- Correlation describes the strength of a linear relationship between two variables
- Regression tells us how to draw the straight line described by the correlation

Regression

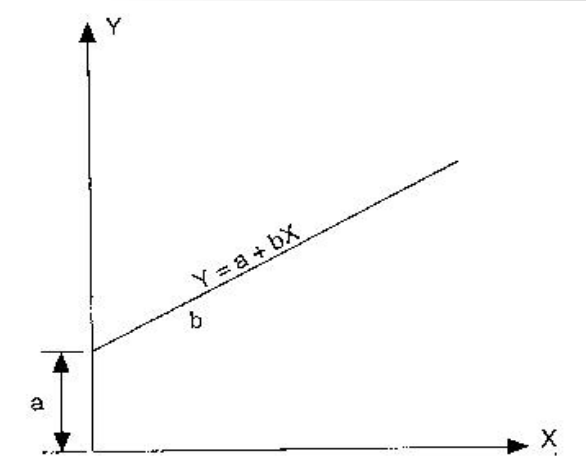
- Calculates the “best-fit” line for a certain set of data
- The regression line makes the sum of the squares of the residuals smaller than for any other line



Line of Regression

By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

- Y= a + bX
  - b, the slope (rate of change in Y per unit change in X)
  - a, intercept (value of Y when X is 0)



Hypothesis Testing for Regression

- Similar to hypothesis testing of correlation coefficient r.
- Tests for slope and intercept.
- Since slope is dependent on r (correlation coefficient), the result (accept or reject Null) of both test would always be same.

Demo

Exercise-8 Cont…

Chi-Square Test

Non-Parametric

Parametric and Nonparametric Tests

- The term "non-parametric" refers to the fact that the chi-square tests do not require assumptions about population parameters nor do they test hypotheses about population parameters.
- Previous examples of hypothesis tests, such as the t tests and analysis of variance, are parametric tests and they do include assumptions about parameters and hypotheses about parameters.
- The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (t and ANOVA) is the nature of the data.
- For chi-square, the data are frequencies rather than numerical scores.

The Chi-Square Test for Goodness-of-Fit

- The chi-square test for goodness-of-fit uses frequency data from a sample to test hypotheses about the shape or proportions of a population.
- Each individual in the sample is classified into one category on the scale of measurement.
- The data, called observed frequencies, simply count how many individuals from the sample are in each category.
- The null hypothesis specifies the proportion of the population that should be in each category.
- The proportions from the null hypothesis are used to compute expected frequencies that describe how the sample would appear if it were in perfect agreement with the null hypothesis.

Formula:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

In R:

```
chisq.test(frequencyVector)
```

Demo

Exercise-9