Applied Statistics
Week - 1

POST GRADUATE PROGRAM

AIML

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Agenda (2 hrs 45 mins)

1. Statistics and Descriptive Statistics (30 mins)
2. Probability Concepts (30 mins)
3. Probability Distributions (30 mins)
4. Case Studies (75 mins)
5. Q&A covered within each session
6. Summary (5 mins)

Classical Definitions

greatlearning

- "**Statistics** is the aggregate of facts affected to a marked extent by the multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other- *Prof. Horace Secrist*"

The underlying **Science** that involves

Quantitative basis

Data!

Data Collection & Organization
Data Analysis (EDA)
Interpretation of Analysis
Presentation of findings

Purpose
(Applied Statistics?)

Data Science** is a field which encompasses ***Statistical methods, tools, processes and systems

- **Business Analytics** (BA) can be defined as the broad use of data and quantitative analysis for **decision making** within organizations.

Statistics

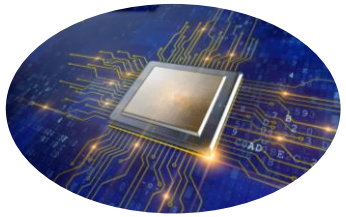
Business Analytics

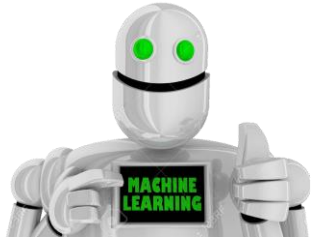
Relevance of Statistics Today?

greatlearning

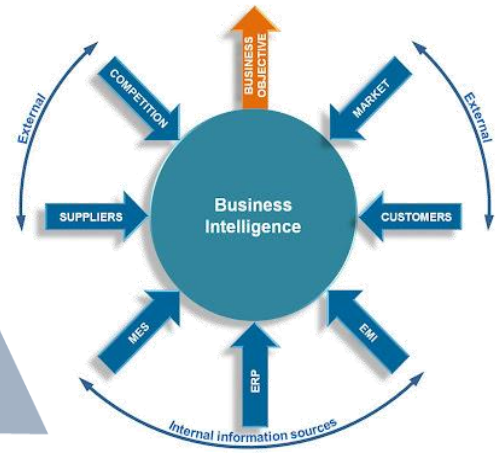
Statistics?

- Technology Development
- Data Explosion
  - ✓ Internet Revolution & Social Networking
  - ✓ Smart /mobile phones
- Data driven Organizations
  - ✓ Data insights drive
    - ✓ Customer understanding
    - ✓ Competitive Advantage & Profitability in Market

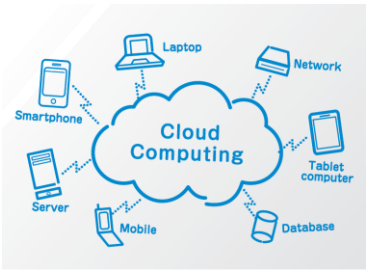




- Computing Power to: Process & analyze “large” data
- Faster & Sophisticated Algorithms for problem solving
- Data driven Organizations
  - ✓ Data visualizations for BI (Business Intelligence) & AI (Artificial Intelligence)



- Large Data Storage Capacity
- Parallel & Cloud Computing



Statistics and Big Data Today?

greatlearning

Big Data

- A set of data that cannot be managed, processed, or analysed with traditional software/algorithms within a reasonable amount of time.
- Big data revolves around the 5Vs:

  -Volume, Velocity, Variety, Value, Veracity .

Amazon: 20 orders Per sec, Twitter: 500M tweets / day, FB: 900M photos upload / day, 4M Google searches /day

Structured data – DBs, Unstructured data – Weather, Geo-spatial, Media data, NLP Semi-structured – HTML code

Volume

→ = *1000

KB → MB → GB → TB → PetaB → ExaB → ZetaB → YottaB

Small ascii data file, thumb size photo (jpg)

This Ppt file, mp3 audio, medium industry datasets (csv)

HD Movie file, 50 hrs of browsing

Desktop drive , Personal HDD, 8 smartphones

Autonomous car data in 1 year,

Data storage capacity of FB, All words spoken by Humans ever

Annual Internet traffic

Statistical Methods

greatlearning

Classifications

- Segmentation of customers
- Brand/Product positioning
- Event Success vs Failure
- Demographics & Psychographics
- Classifying for Social Networking

Pattern Recognition

- Unearth hidden patterns in Data
- Visual analytics: Histograms, Box-plots, Scatter/Line Plots
- Feature extraction, clustering, discrimination
- Speech (acoustic patterns) recognition, character recognition (image patterns)

Associations

- Associations amongst variables
- Measures: Pearson's correlation, chi-square test, Relative risk & odds ratio etc
- Association Rule: Market basket analysis for association amongst items
- Popular usage in epidemiology, psychology studies

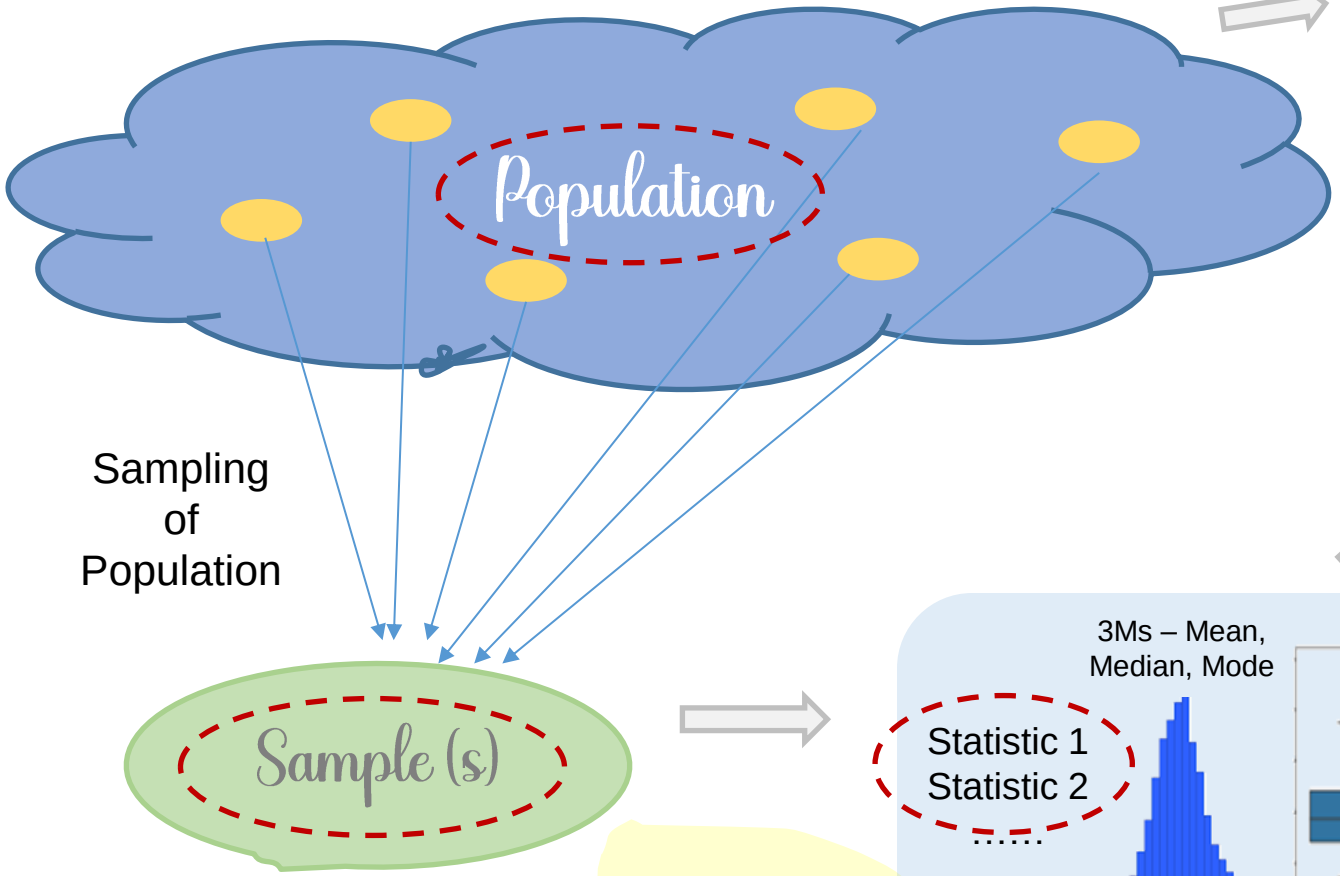
Predictive Modeling

- Supervised and Unsupervised ML models
- Classification, Regression, Clustering models
- Algorithms such as Linear Regression, Naïve Bayes, Logistical Regression, K-Nearest-Neighbors, Decision Trees, Ensembles, Neural-Networks etc

Applied Stats Terms

Examples: India electoral population, Current Cars in US, Worldwide traffic scenarios, Covid in Maharashtra

"Universe" of data related to problem



A "finite" and manageable sampling And representative of Population

Parameter 1
Parameter 2
….
Hypothesis

Quantified results sought for problem description, solution(s) and business interest

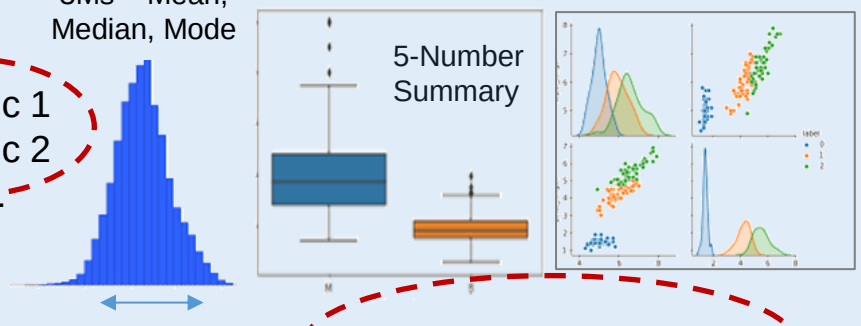
Deeper insights thru Hypothesis Testing, Probability Models, Regression, classification, Correlation etc

Predictions
Inference
Generalization

Inferential Statistics
(Probability models, Hypothesis testing Regression & Correlation Analysis)

3Ms – Mean, Median, Mode

Statistic 1
Statistic 2
……



Variance
Std. Dev

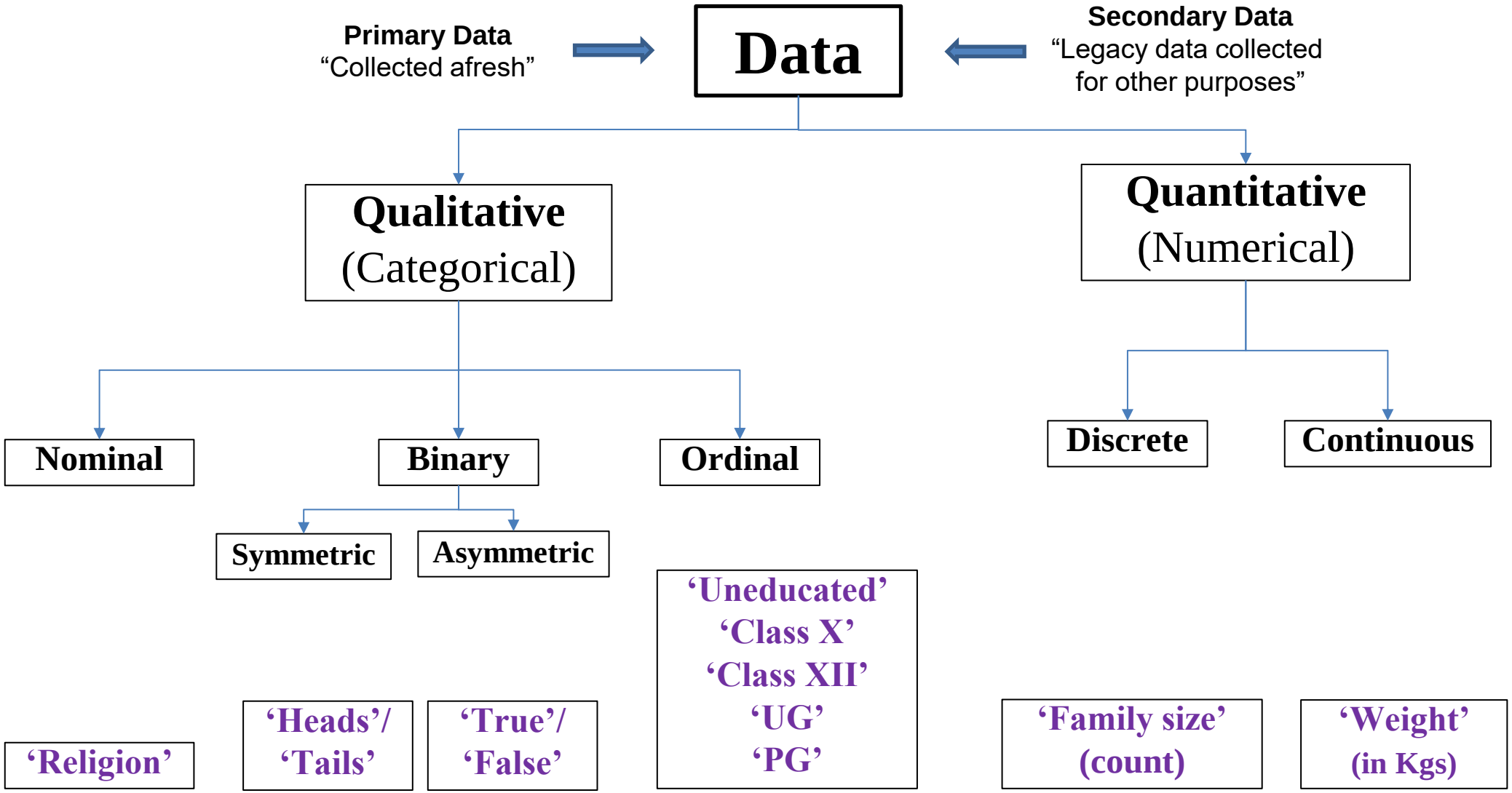
Distribution
Central Tendency
Variation

Descriptive Statistics

EDA to draw visual Insights and statistic

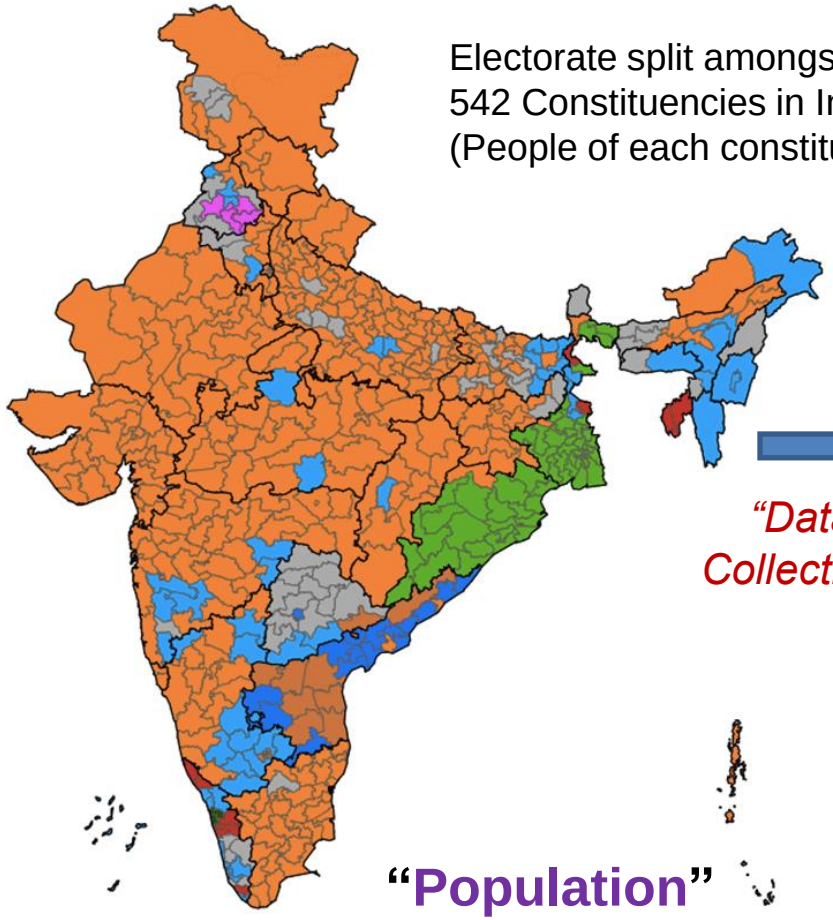
Data Types?

greatlearning



Statistics – Key Terms
(A Psephologist's Example)

greatlearning



Electorate split amongst
542 Constituencies in India
(People of each constituency)

"Population"
(~600M voters out of 900M eligible to vote)

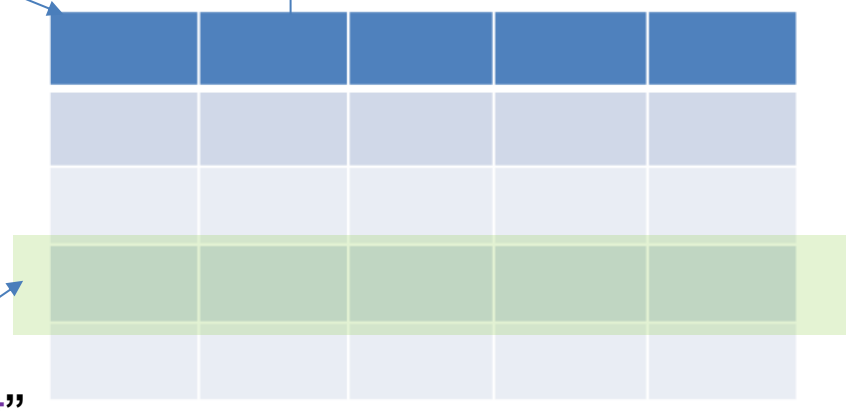
"Data Collection"

"Sample"
(Survey 1400 people Per constituency)
750K people sampled

['Age', 'Gender', 'Const', 'Caste', 'Religion', ….. 'Vote pref", 'Reason']

"Dataset"

"Attribute"



"Data Object"

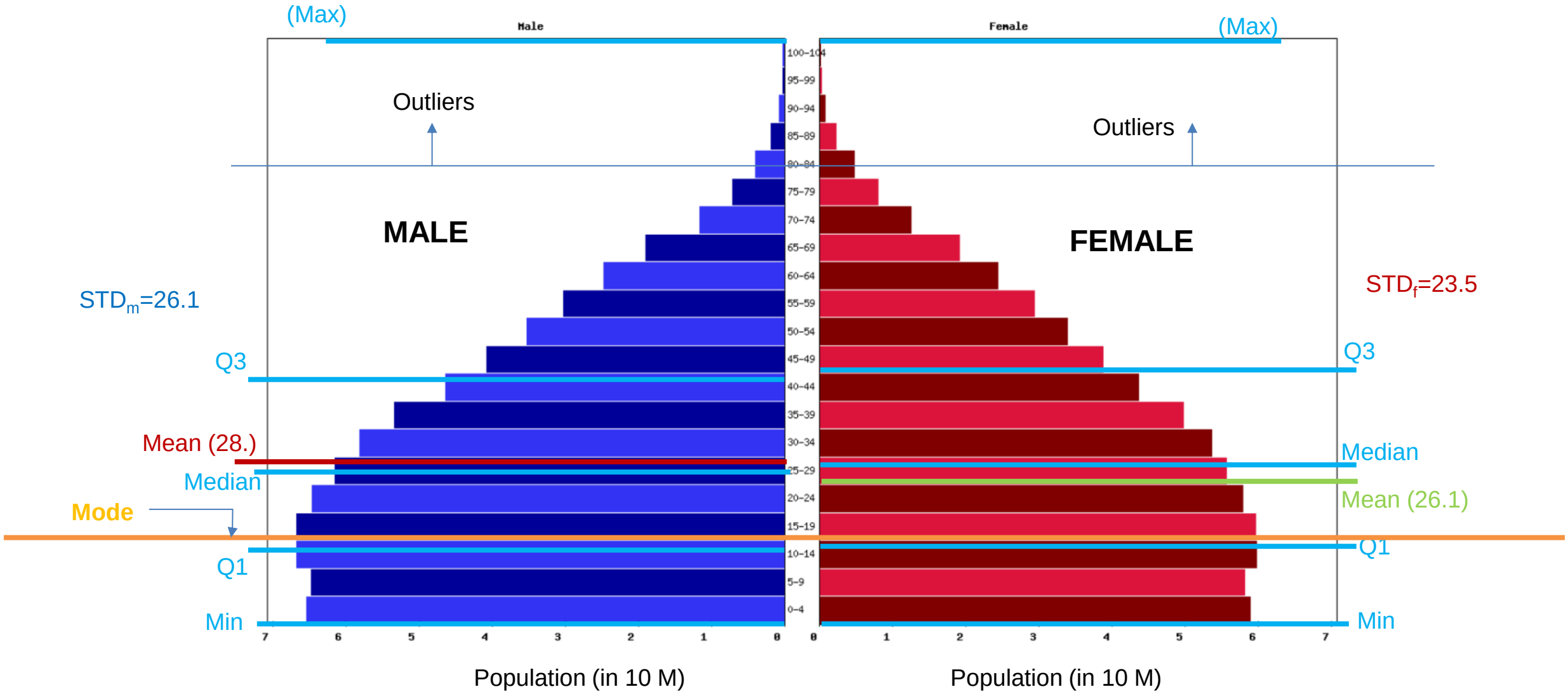
EDA → 'Descriptive Statistics'

Sample "Statistics" – Age & Gender distribution,…..

Predictive Models → 'Inferential Statistics'

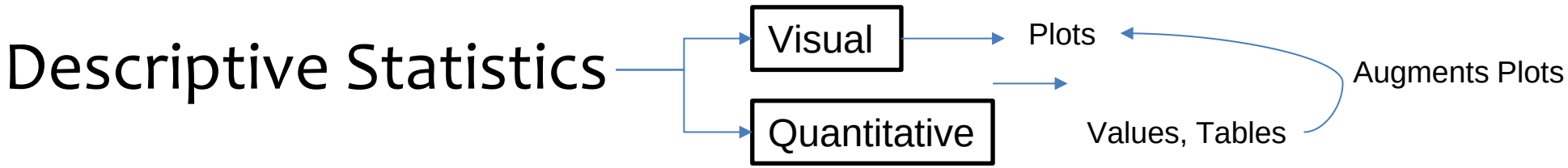
Population "Parameters" – %Vote share, Const. Seat winner etc

Applied Statistics Concepts:
Descriptive Stats example
Population distribution by age in India



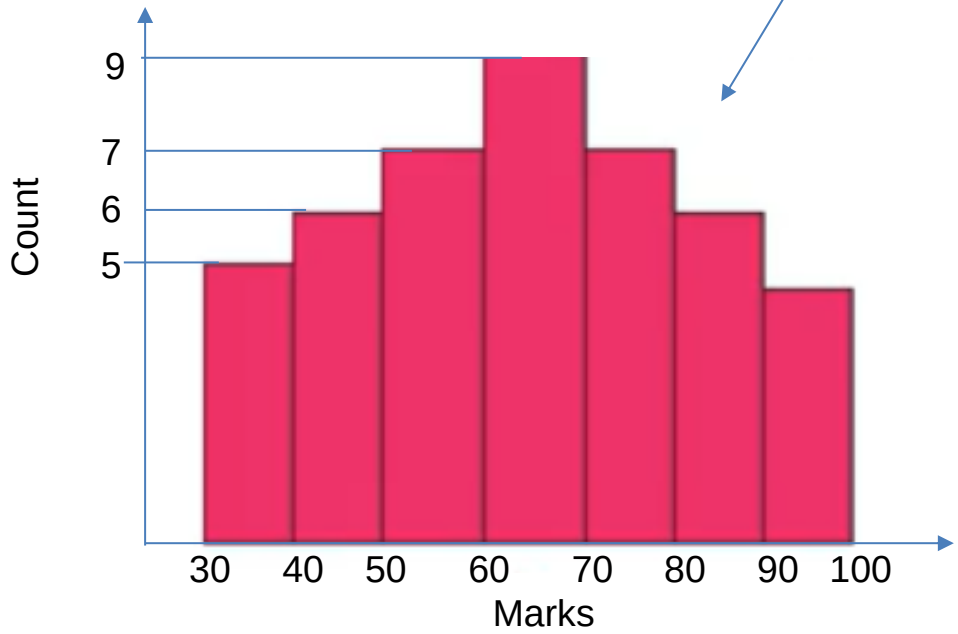
Descriptive Statistics Aspects

greatlearning

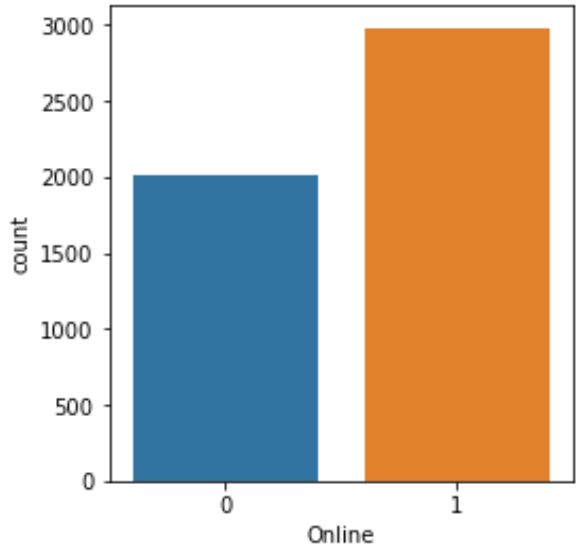


Univariate Analysis

1. Frequency Distribution via **histograms** and **bar plots**



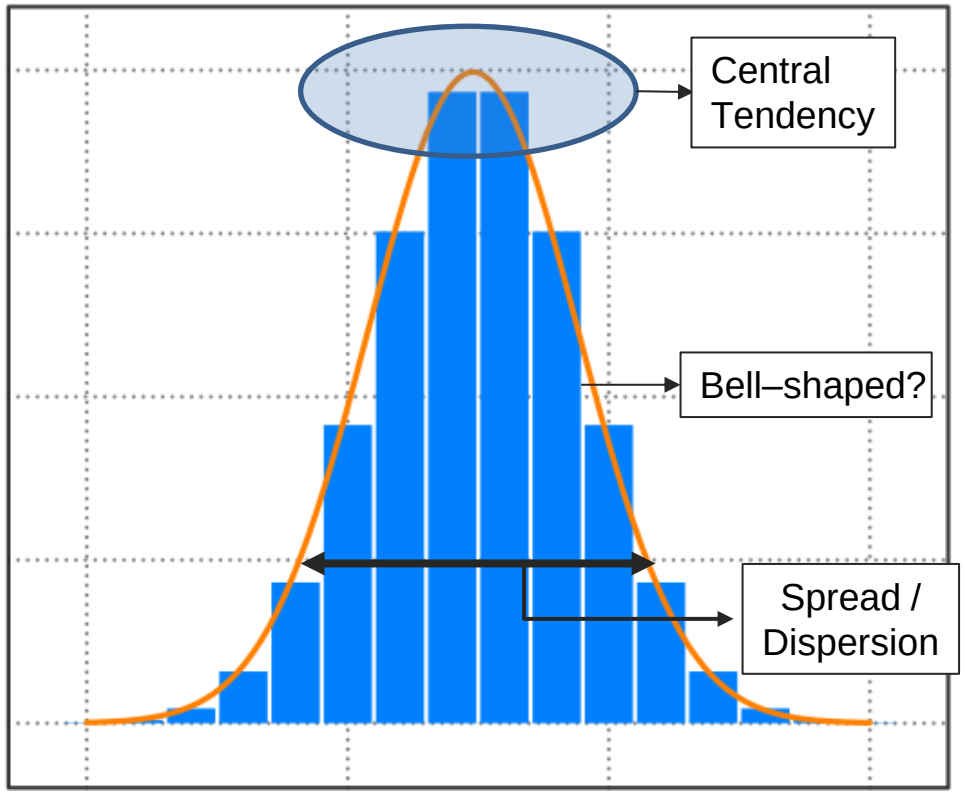
For Numerical Attributes



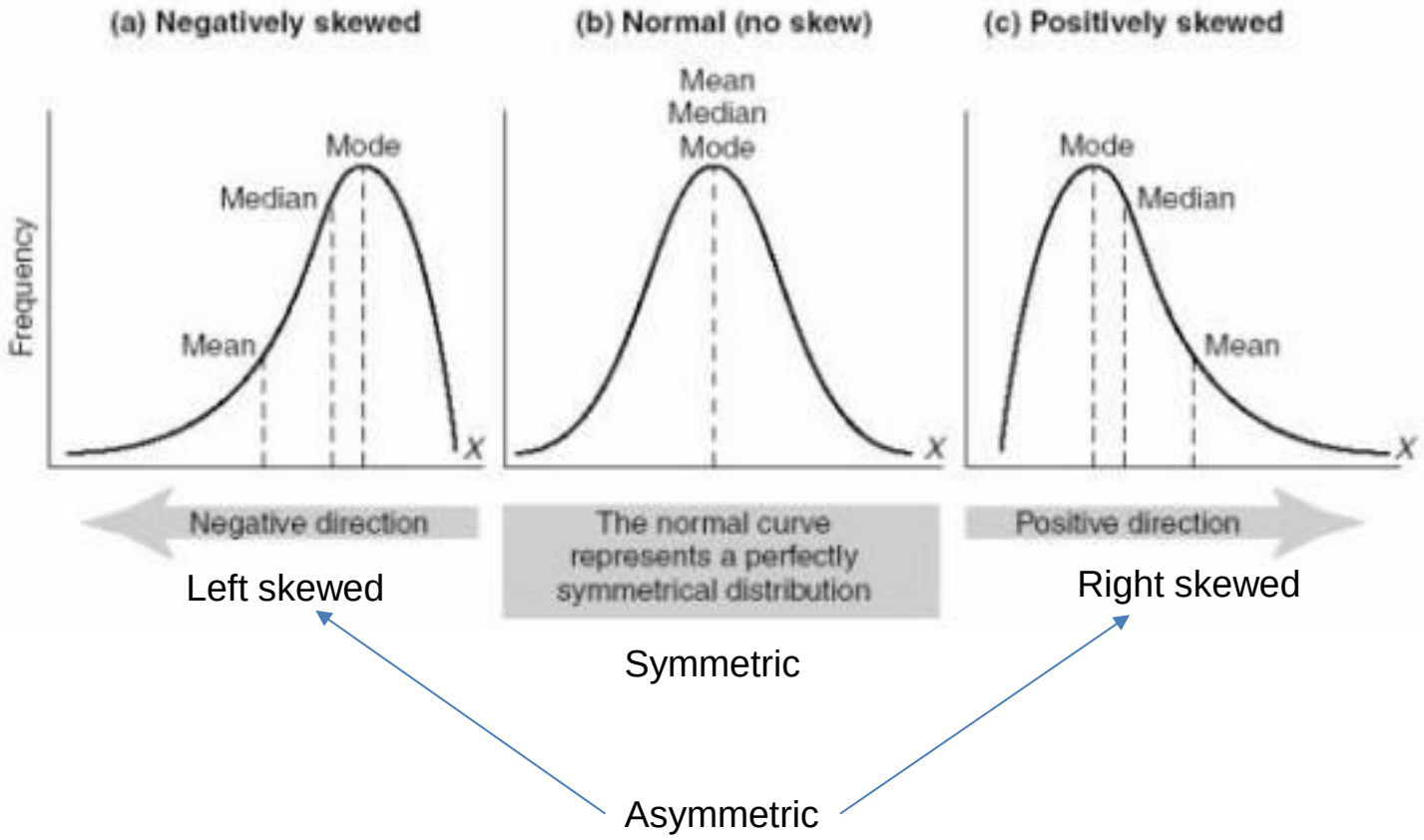
For Categorical Attributes

Descriptive Stats (cond.)

Shape and Spread of Frequency Distribution



Shape



Descriptive Stats (cond.)

Central Tendency and 3 Ms

greatlearning

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of "Central tendency"

3Ms:

1. Mean

The statistical mean refers to the mean or average that is used to derive the central tendency of the data.

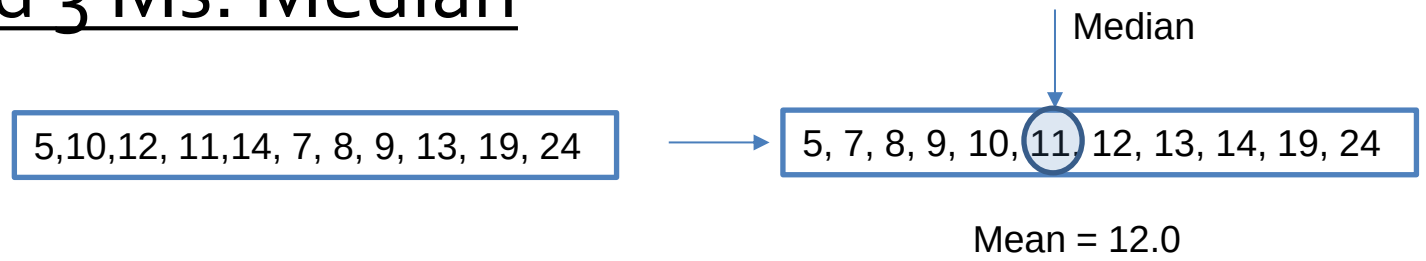
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \lambda + X_n}{n}$$

Descriptive Stats (cond.)

Central Tendency and 3 Ms: Median

greatlearning

2. Median



The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for original data, in which values are ranked relative to each other but are not measured absolutely.

$$Median = \left(\frac{n+1}{2}\right)^{th} term \longrightarrow$$ For Odd number of terms sequentially arranged

$$Median = Avg\left[\left(\frac{n}{2}\right) th\ term + \left(\frac{n+2}{2}\right) th\ term\right] \longrightarrow$$ For Even number of terms sequentially arranged

Descriptive Stats (cond.)

Central Tendency and 3 Ms: Mode

3. Mode → The Central Tendency Measure used for Categorical variable

The most frequent value in the data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.

$$Mode = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

Where,

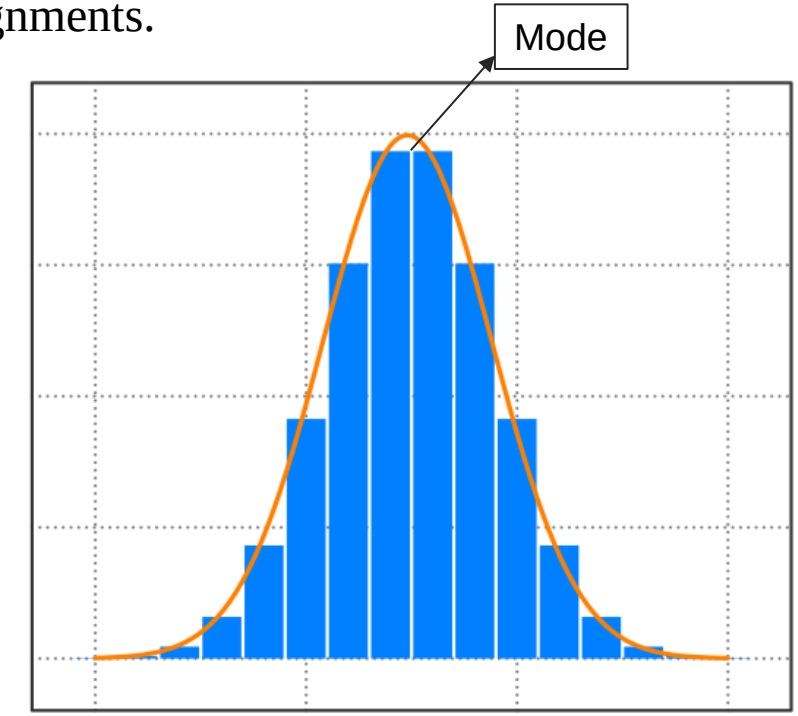
l = Lower Boundary of modal class

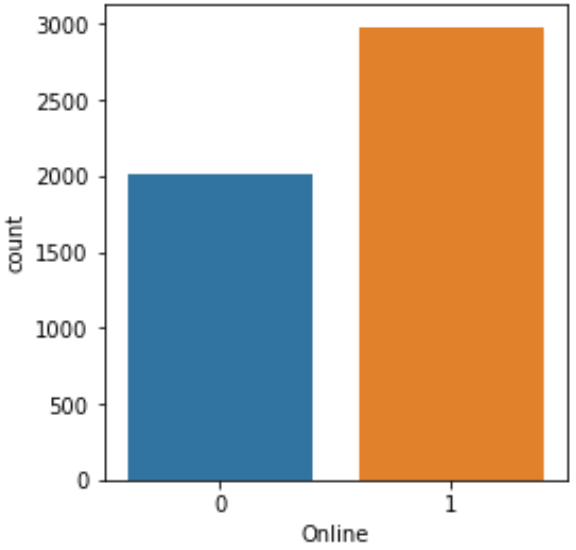
h = *size of model class*

f_m = *Frequency corresponding to modal class*

f_1 = *Frequency preceding to modal class*

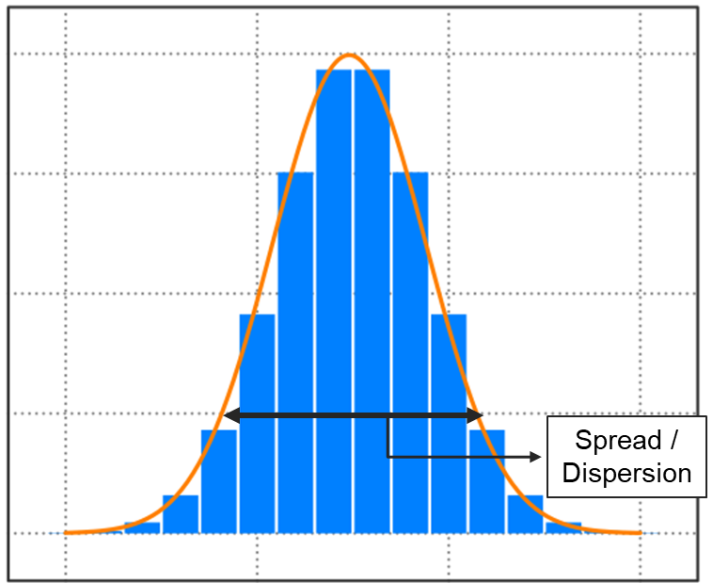
f_2 = *Frequency proceeding to modal class*





Descriptive Stats (cond.)

Measures of dispersion: Range



Measures of dispersion

Measure of dispersion indicate how large the spread of distribution in around the central tendency.

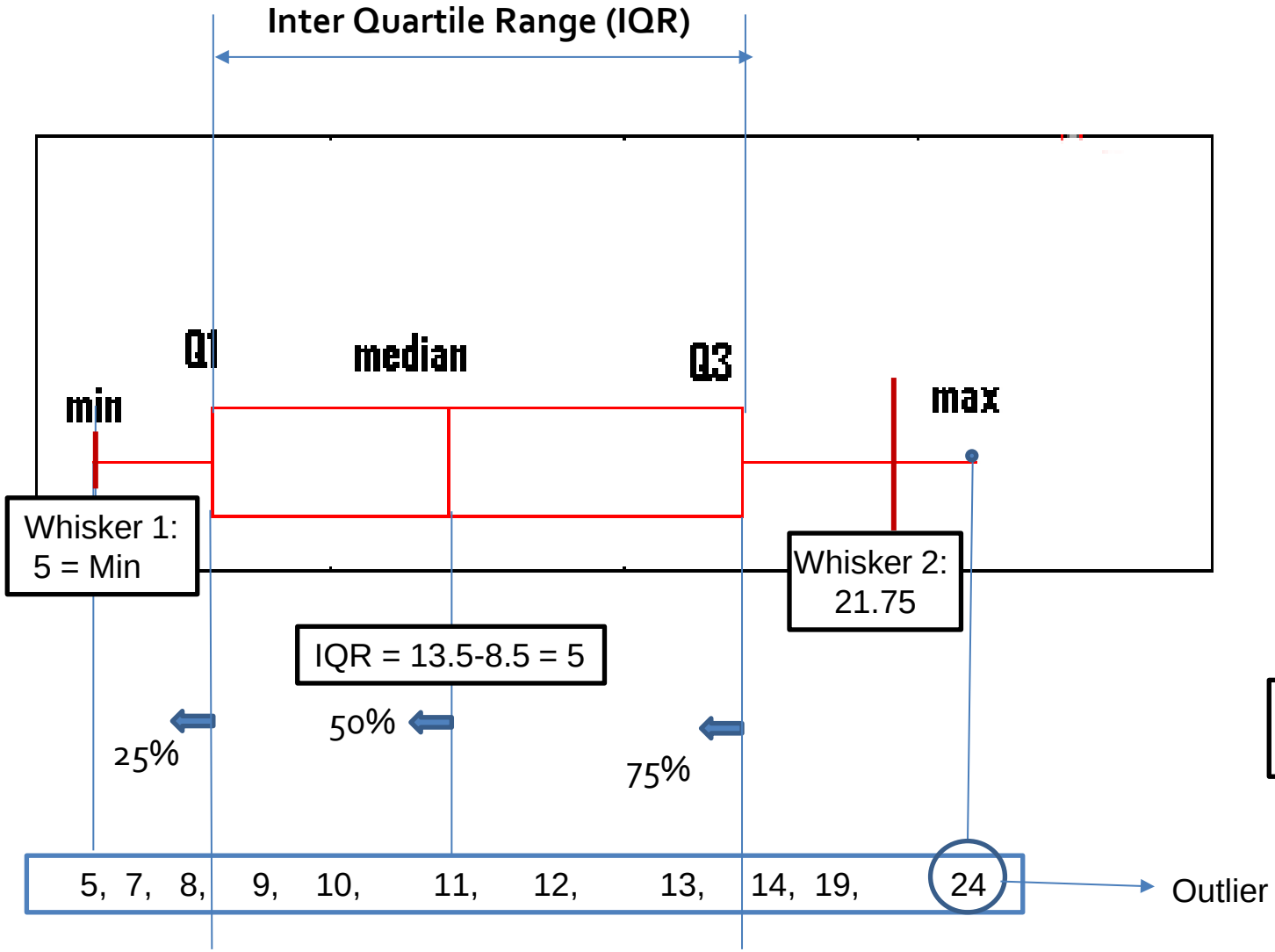
Range

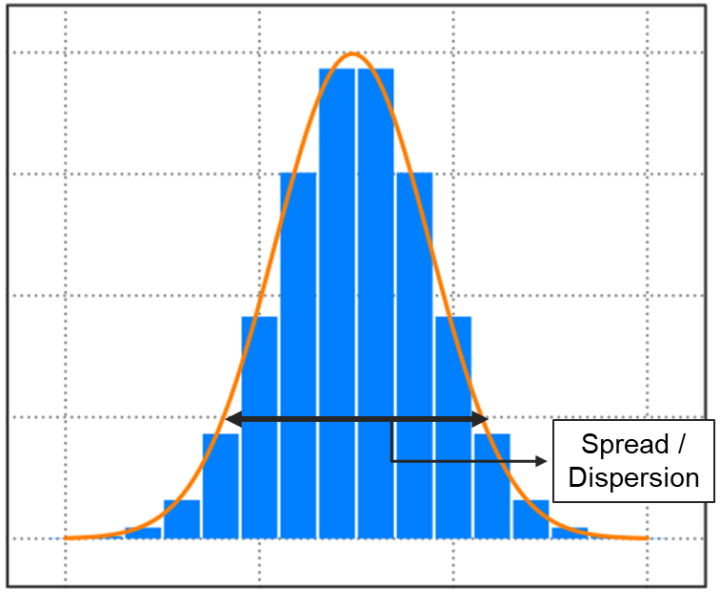
Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in dataset.

range = X(maximum) - X (minimum)

Descriptive Stats (cond.)

Measures of dispersion: IQR





Whisker 1: Q1 - 1.5*IQR
Whisker 2: Q3 + 1.5*IQR

Descriptive statistics – Standard Deviation

greatlearning

- Interpreting variance (a squared term) is not intuitive. Instead we under root it to get Standard deviation which has the same units as variable.

- Standard deviation, is a measure of average spread i.e., on an average what is the difference between any data point and the central value of the variable.

$$S_d = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{(N-1)}}$$

Variance

- Coefficient variation is defined as ratio of standard deviation to mean

$$CV = \frac{S}{\bar{X}} \text{ for the sample data and } = \frac{\sigma}{\mu} \text{ for the population}$$

Descriptive Stats (contd.)

Bivariates

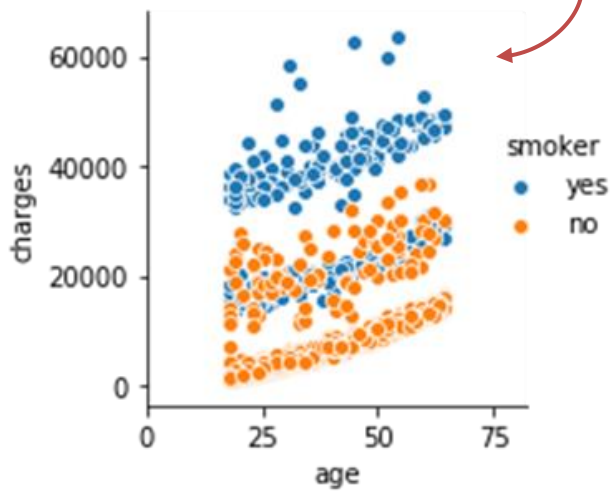
Data Pair Types:

X \ Y	Categorical	Numerical
Categorical	C-C	C-N
Numerical	N-C	N-N

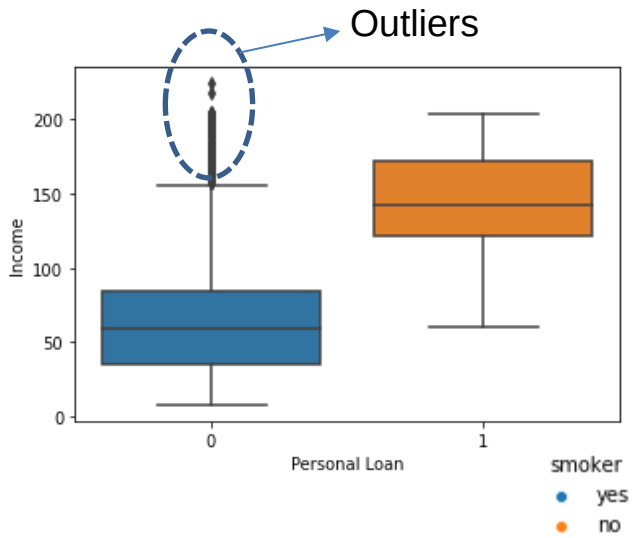
P.Loan / Education	No	Yes
Level 1	2003	93
Level 2	1221	182
Level 3	1296	205

Crosstab
(*pd.crosstab(X,Y)*

Pivot Table
for multiple C-Cs
pd.pivot_table()



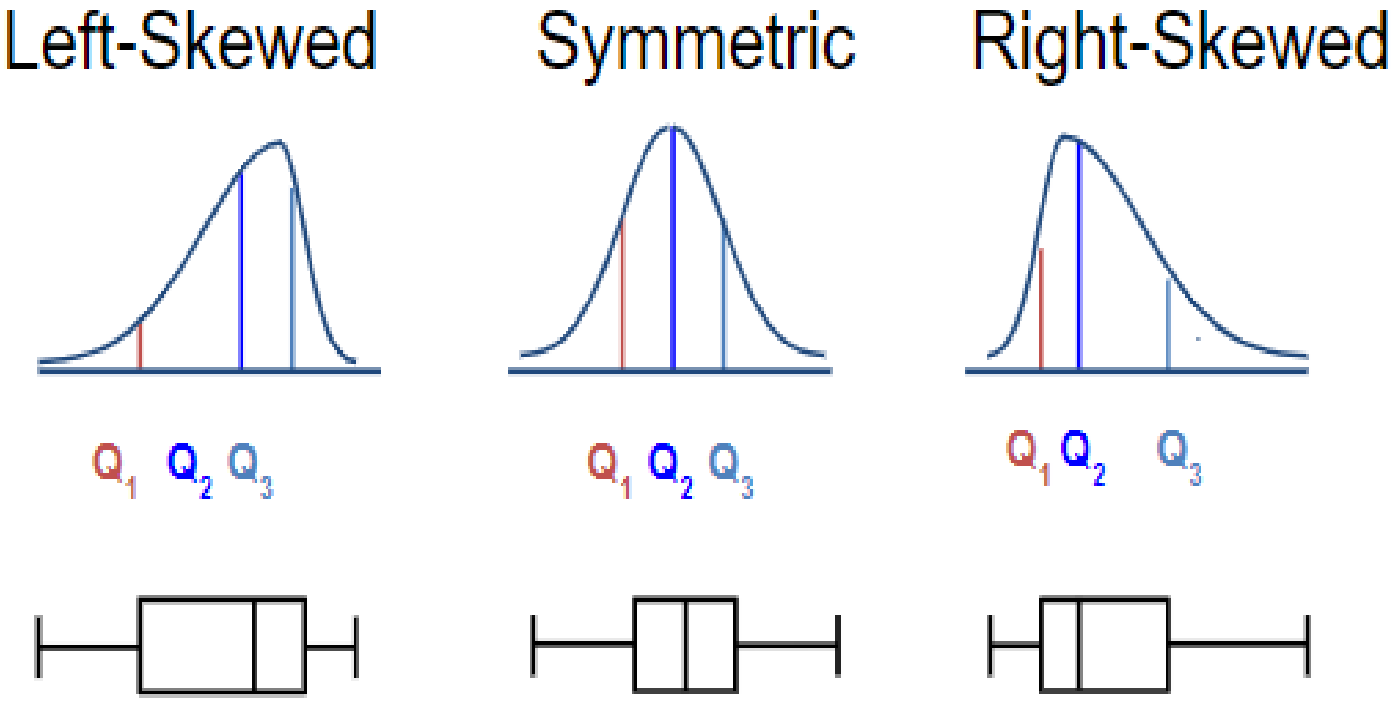
Scatter Plot (Bi-variate)
Pair Plot (Multi-variate)



Box Plots
(*sns.boxplot()*
sns.stripplot(),
sns.swarmplot(),
sns.violinplot() etc)

Descriptive Stats (contd.)

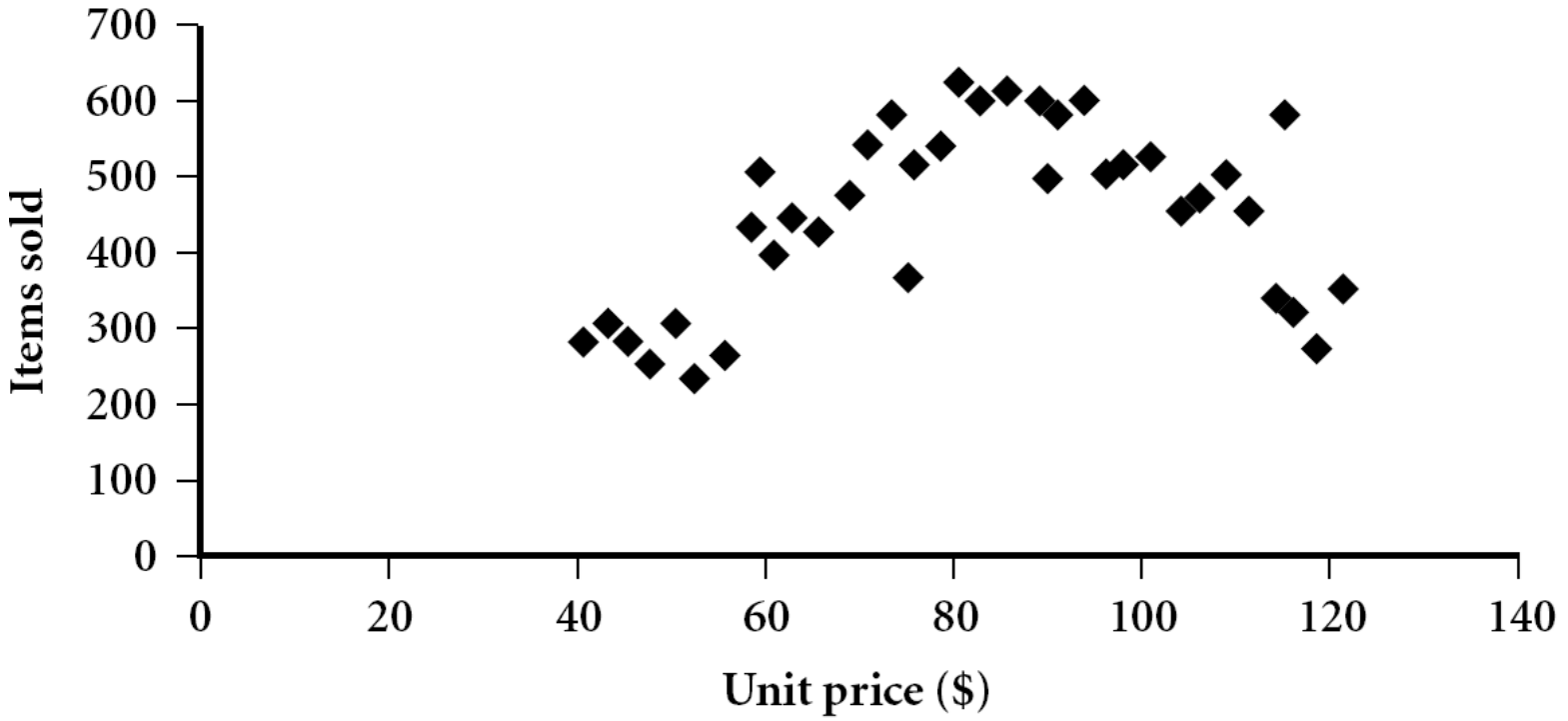
Distribution Shape and the Boxplot



Scatter Plot

greatlearning

Provides a first look at bivariate data to see clusters of points, outliers, etc
Each pair of values is treated as a pair of coordinates and plotted as points in the plane



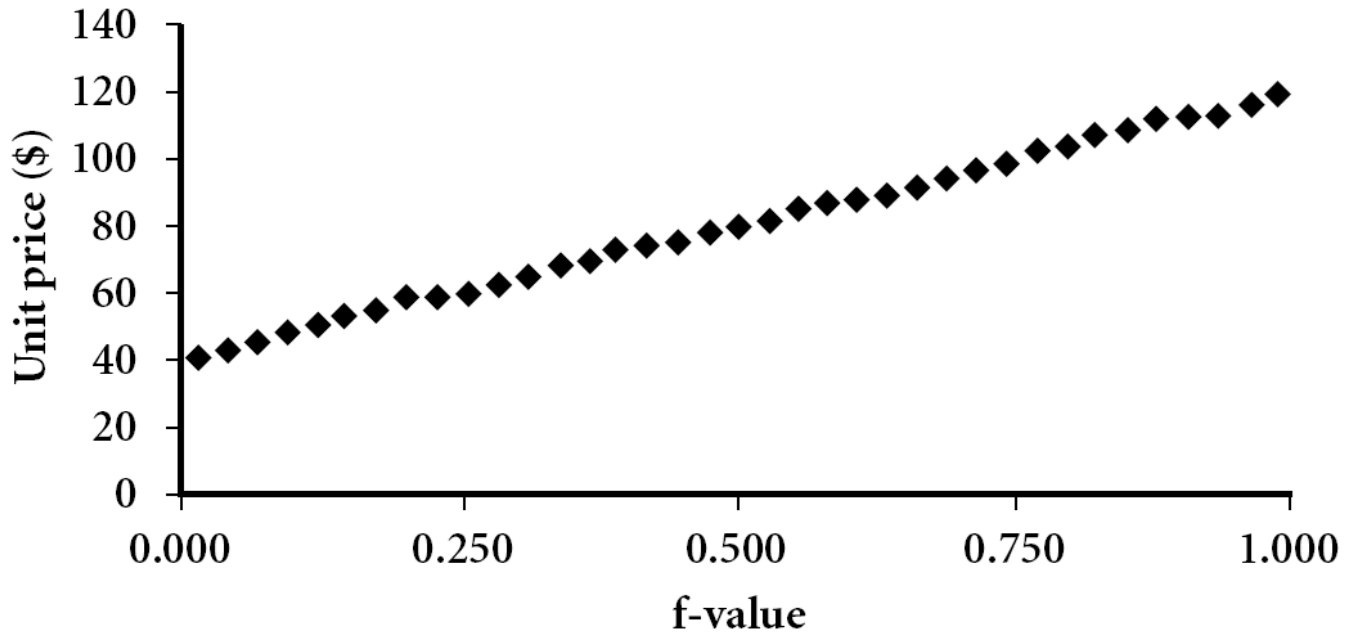
Descriptive Stats (contd.)
Quantile Plot

greatlearning

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

Plots **quantile** information

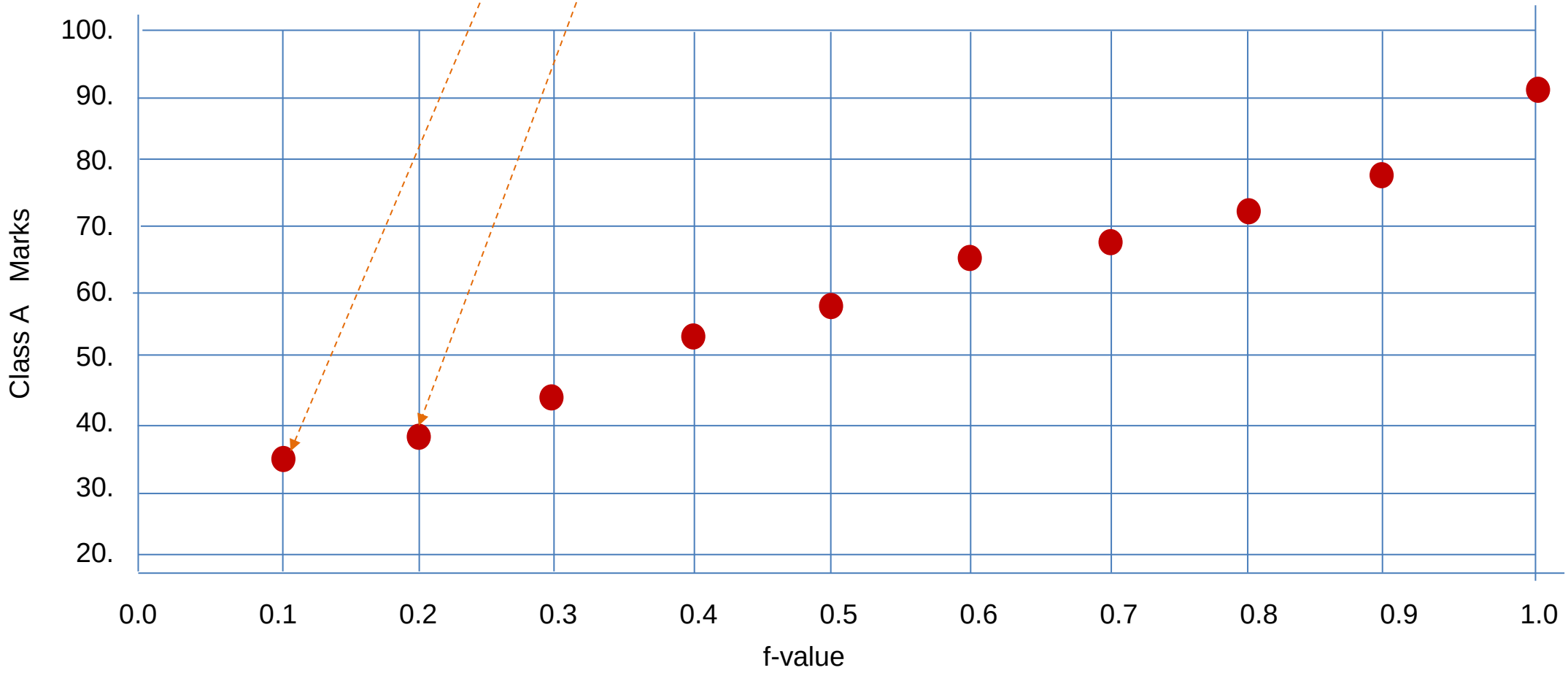
For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



Descriptive Stats (contd.)

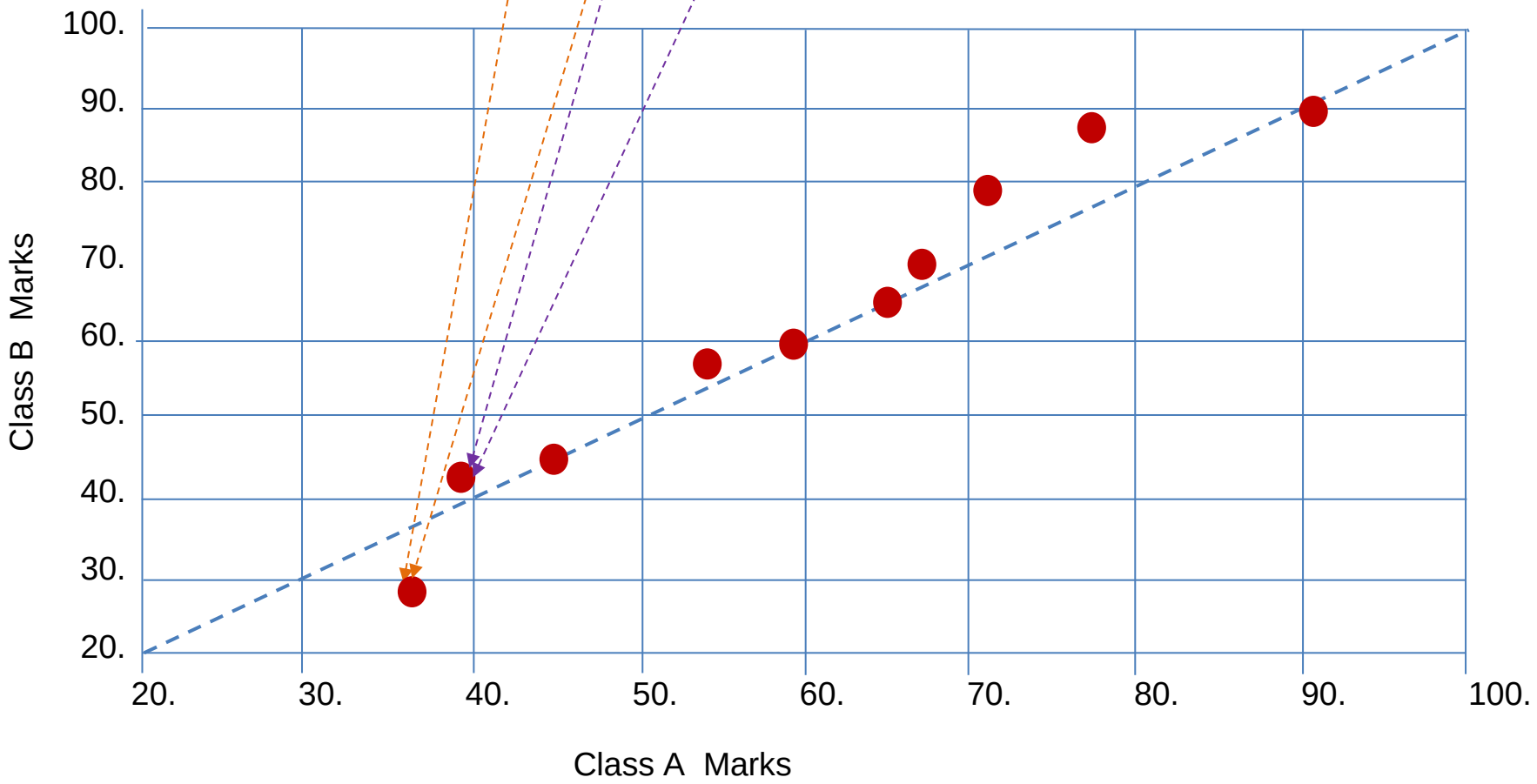
Example of Quantile plot

- Marks Class A = [35. 38. 44. 53. 58. 65. 67. 72. 77. 91.]
- Marks Class B = [29. 43. 44. 57. 59. 65. 69. 77. 84. 90.]



Descriptive Stats (contd.)
Example of Q-Q plot

- Marks Class A = [35. 38. 44. 53. 58. 65. 67. 72. 77. 91.]
- Marks Class B = [29. 43. 44. 57. 59. 65. 69. 77. 84. 90.]

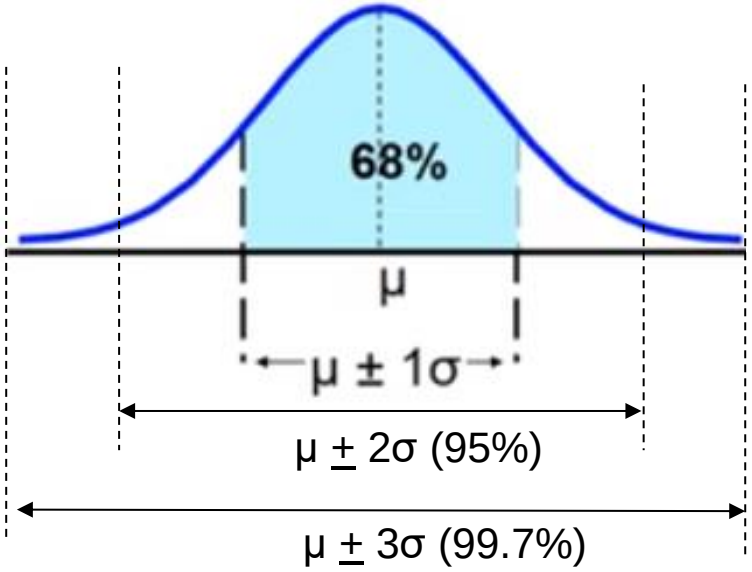


Descriptive Stats (cond.)

The Empirical Rule

The Empirical rule

- The empirical rule approximates the variation of data in a bell-shaped distribution.
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean 95% within 2 σs and 99.7% within 3 σs



"Bell-Shaped" → "Normal Distribution" → "Gaussian"

Descriptive Stats (cond.)

The Empirical Rule and Chebyshev Rule

greatlearning

Chebyshev Rule

- Regardless of how the data are distributed, at least $(1-1/k^2)x100\%$ of the values will fall within k standard deviations (for k>1)

- For example, when k=2, at least 75% of the values of any data set within 2σs

Descriptive Stats (contd.)
Measures of Associate amongst Bivariates

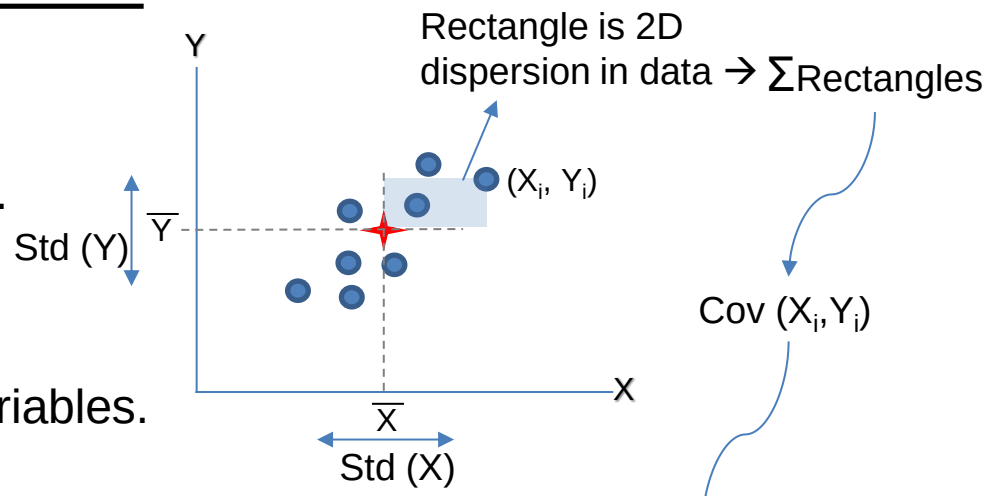
greatlearning

Covariance

- *Covariance* is a measure of association between 2 variables.
- It represents association in units of the two variables.

Correlation

- Correlation is also a measure of association between two variables.
- Moreover, it is a dimensionless quantity and thus enables comparison beyond units.
- Coefficient of correlation is also known as ***Pearson's*** coefficient

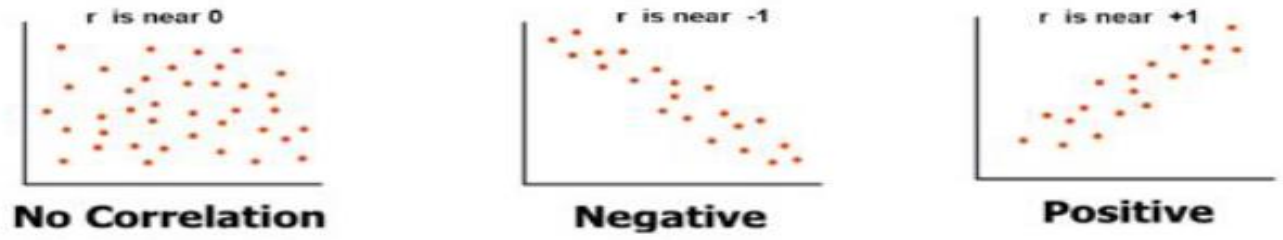


$$Cov(X,Y) = \frac{\sum_N (Xi - \bar{X})(Y_i - \bar{Y})}{(N-1)} \quad \text{units}^2$$

$$r_{xy} = Corr(X,Y) = \frac{Cov\,(X,Y)}{Std(X) * Std(Y)}$$

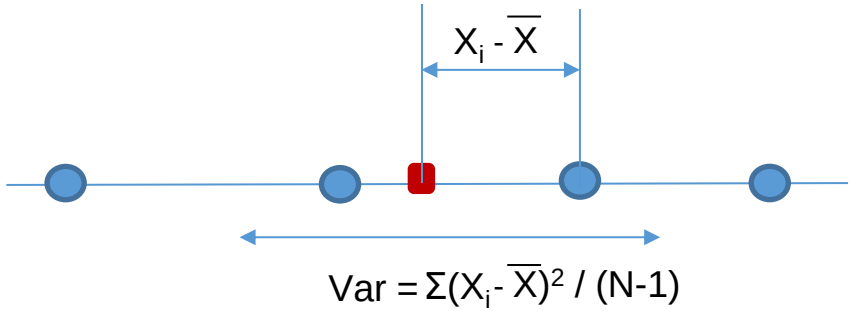
Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

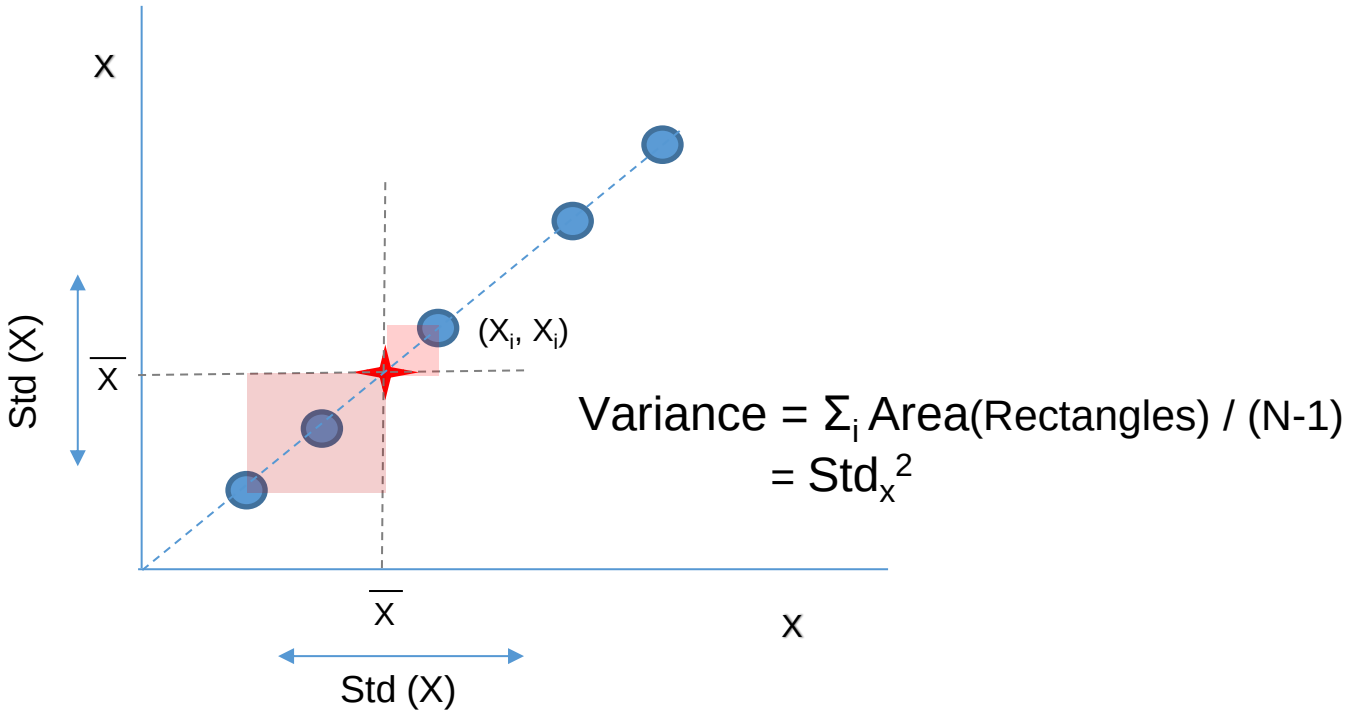
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



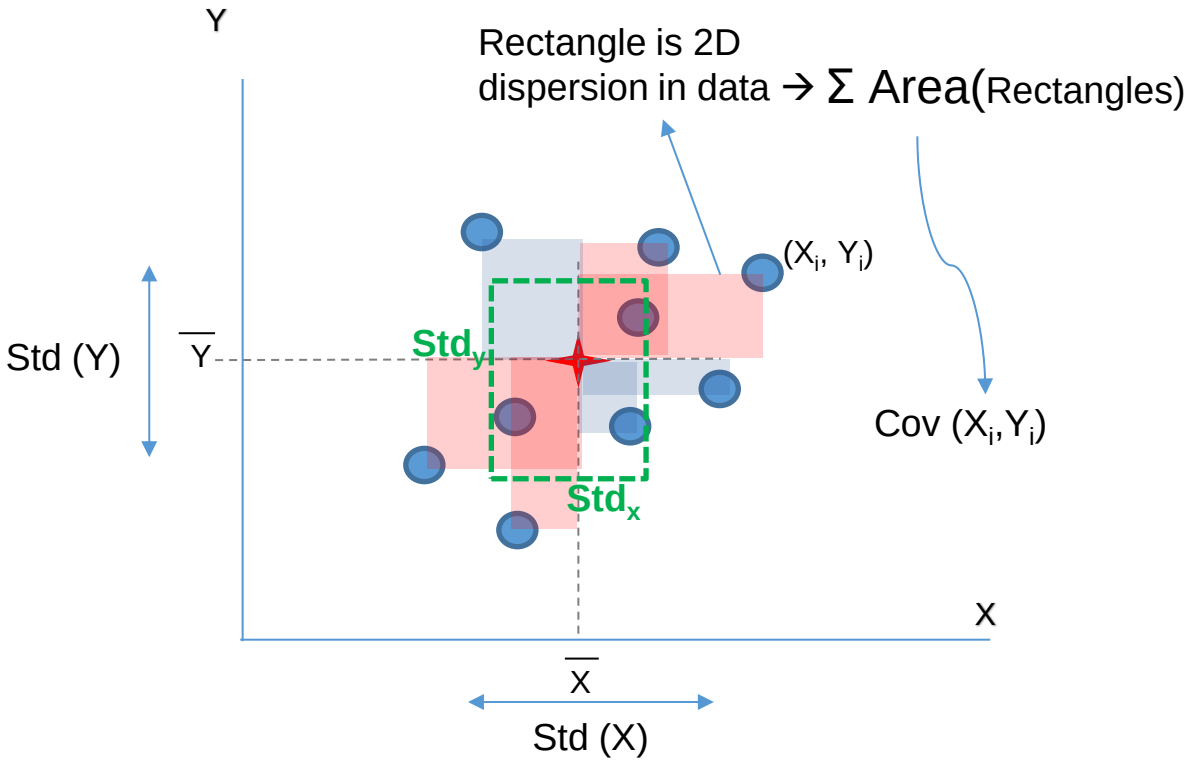
Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! Nonlinear models may be better in such cases

Pearson's Correlation





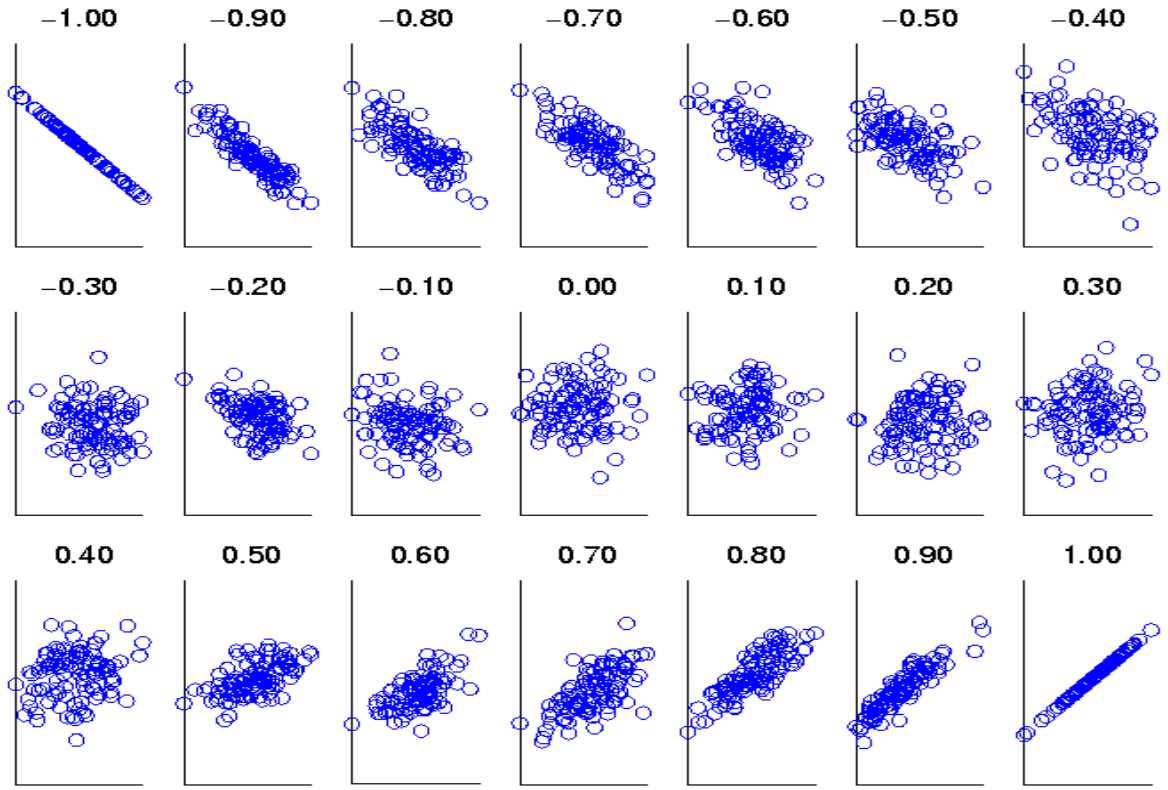
Variance = Σ_i Area(Rectangles) / (N-1)
= Std_x^2



$$Cov(X,Y) = \frac{\sum_N (Xi - \bar{X})(Y_i - \bar{Y})}{(N-1)}$$

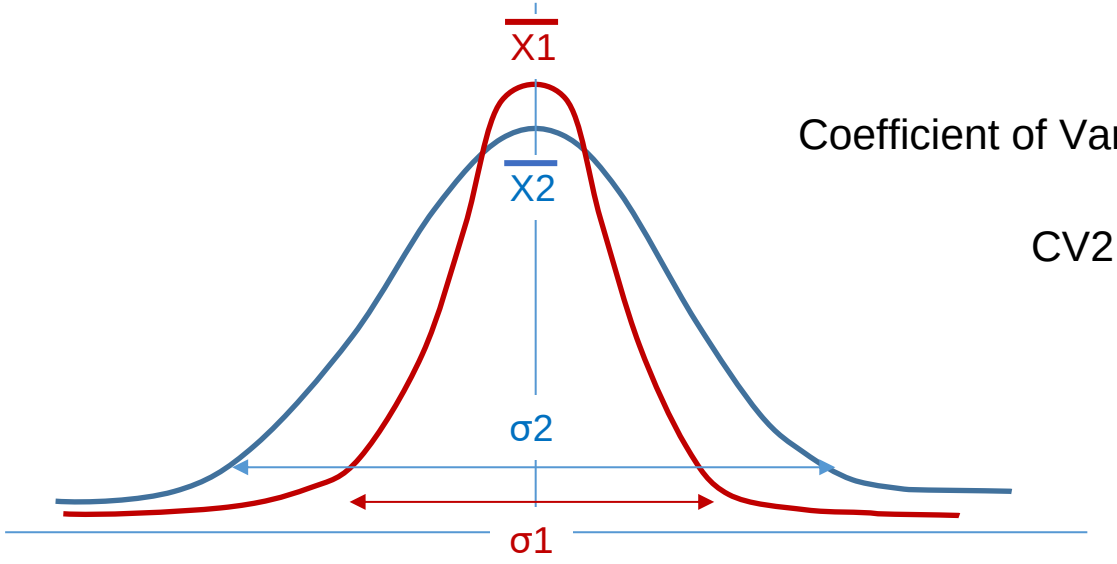
$$r_{xy} = Corr(X,Y) = \frac{Cov\,(X,Y)}{Std(X) * Std(Y)}$$

Descriptive Stats (contd.)
Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1

Coefficient of Variation, CV (will cover in session 2)



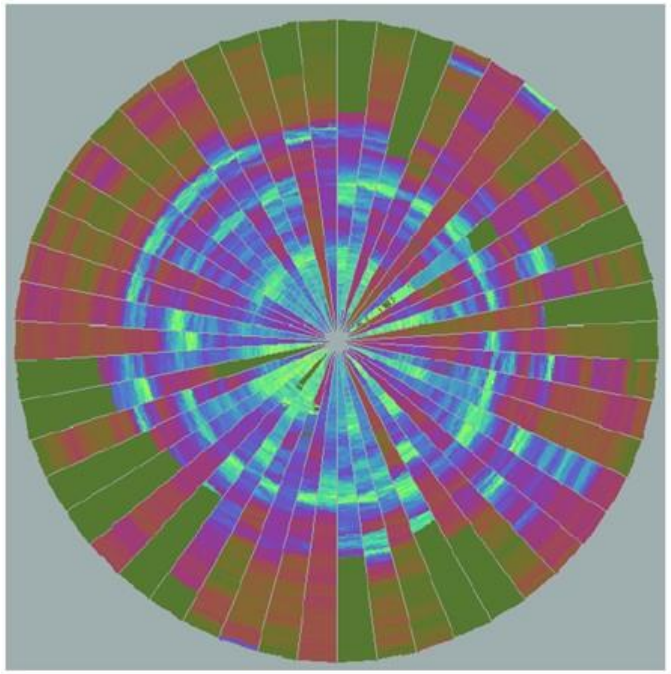
Coefficient of Variation is a measure of relative dispersion

$$CV2 = \frac{\sigma 2}{\overline{X2}} > \frac{\sigma 1}{\overline{X1}} = CV1$$

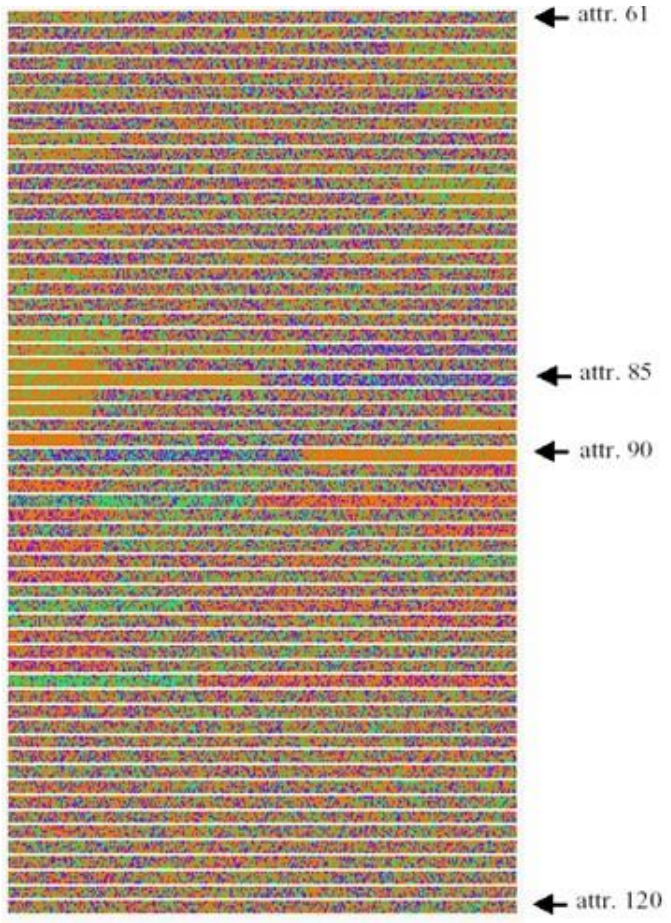
Typical goal is to achieve lower CVs with process quality metrics, predictable parameters, ML model accuracy etc → robustness, lower risk etc

Descriptive Statistics
Data Visualization (Advanced)

greatlearning



Tracking 50 stock prices over time



DNA map with 120 attributes

Pixel Oriented Visualization

Heirarchical Visualization

Applied Stats Concepts: Marginal & Conditional Probability using Covid 2nd wave key data from INDIA

greatlearning
Learning for Life

14-Apr to 11-May 2021	MH	KA	MH+KA	INDIA
COVID +VE DEATH	19425	7028	26453	84721
COVID +VE TOTAL	1665076	957327	2622403	9408672
COVID TESTS TAKEN	6909029	4091141	11000170	46705190

1. What is the probability that a randomly tested Indian from KA will test +ve? (Positivity rate in KA)
   Compare the positivity rate of KA with MH and India overall.

2. What is the probability that a randomly tested Indian will test positive and die? (Crude death rate (per 1000) in India?)

3. What is the probability that a randomly tested Indian will be Covid +ve and from MH?
   (Significance of MH in India Covid context)

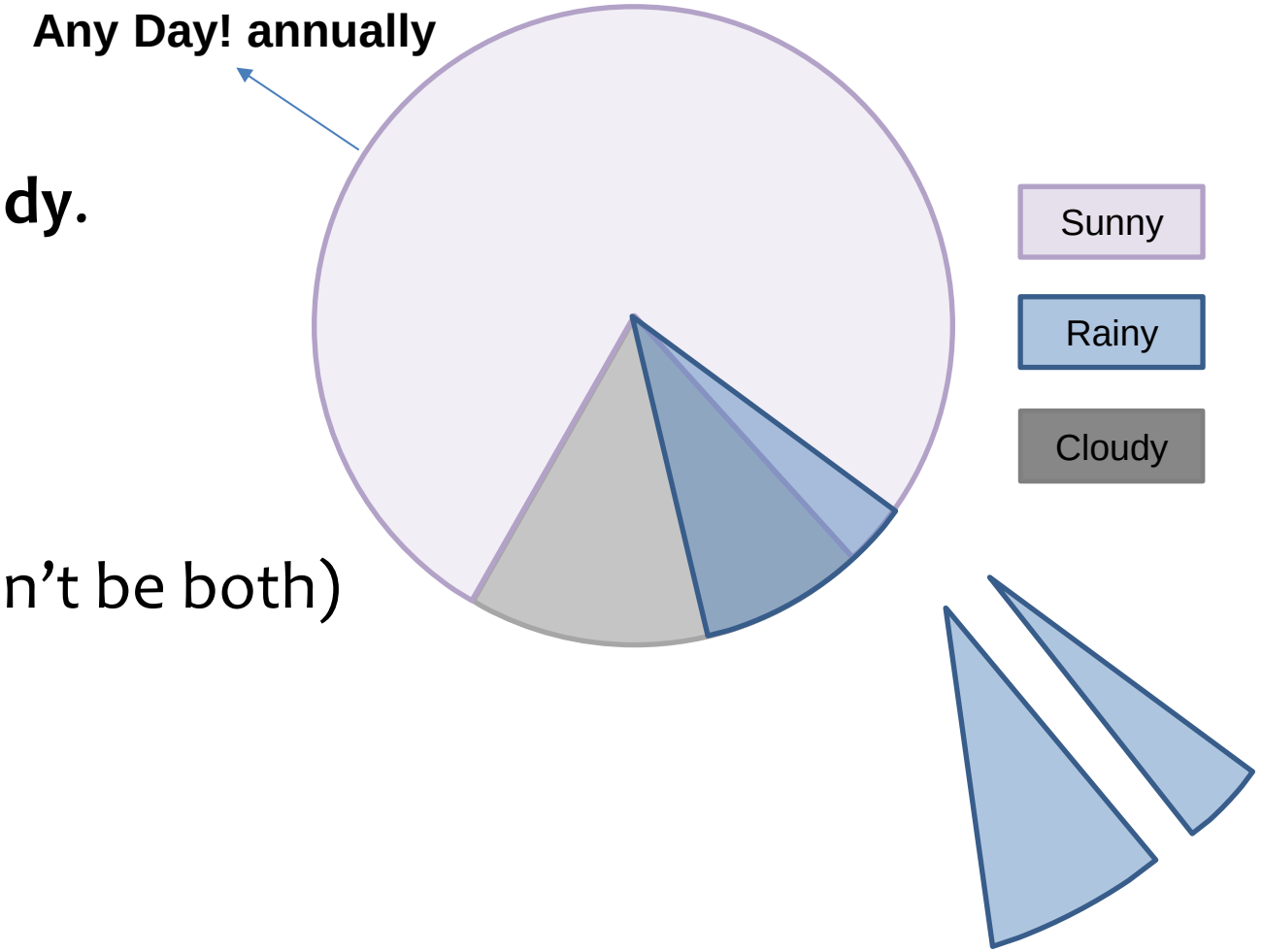
4. What is the probability that a Covid +ve patient from MH or KA will die? (Case fatality rate in MH & KA together)

5. What is the probability that a dead Covid patient is from KA or MH? (% of deaths from KA & MH together)

Simple Example of Bayes' theorem

What kind of a day is it?

greatlearning



A **Day** can either be **Sunny** or **Cloudy.**

But it can **Rain** in either condition.

R = **Raining,** P(R) + P(Ř) = 1,

What type of day is it?

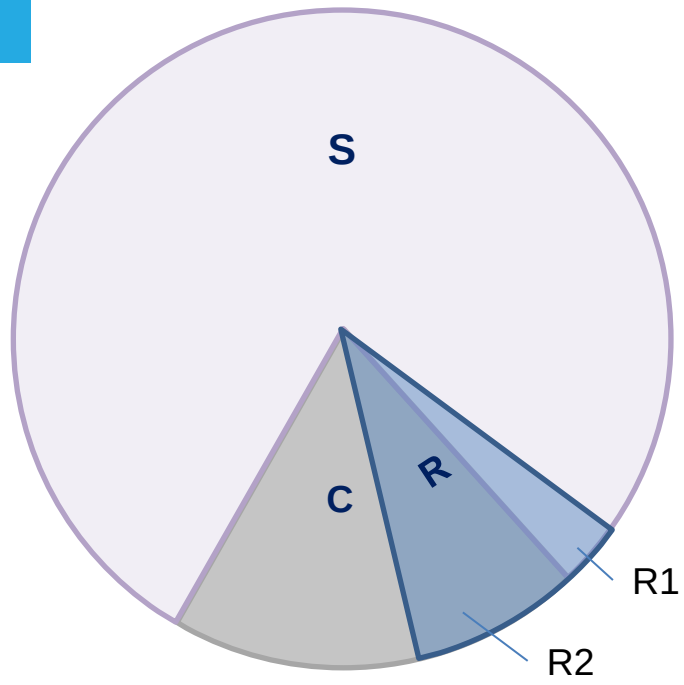
D = Either Sunny or Cloudy **Day**(can't be both)

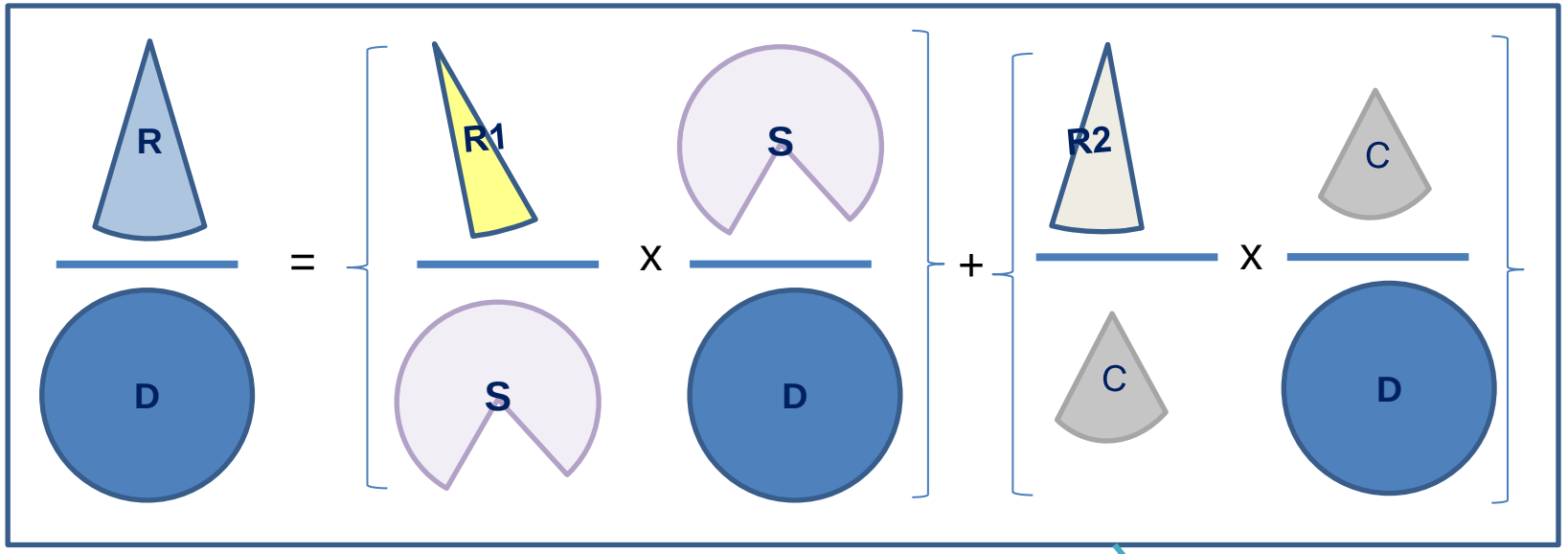
S = **Sunny** Day

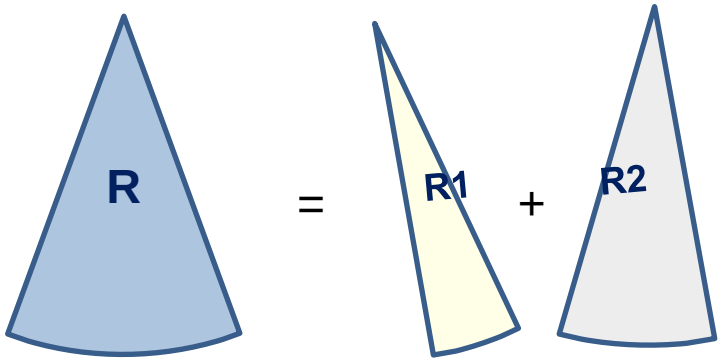
C = **Cloudy** Day

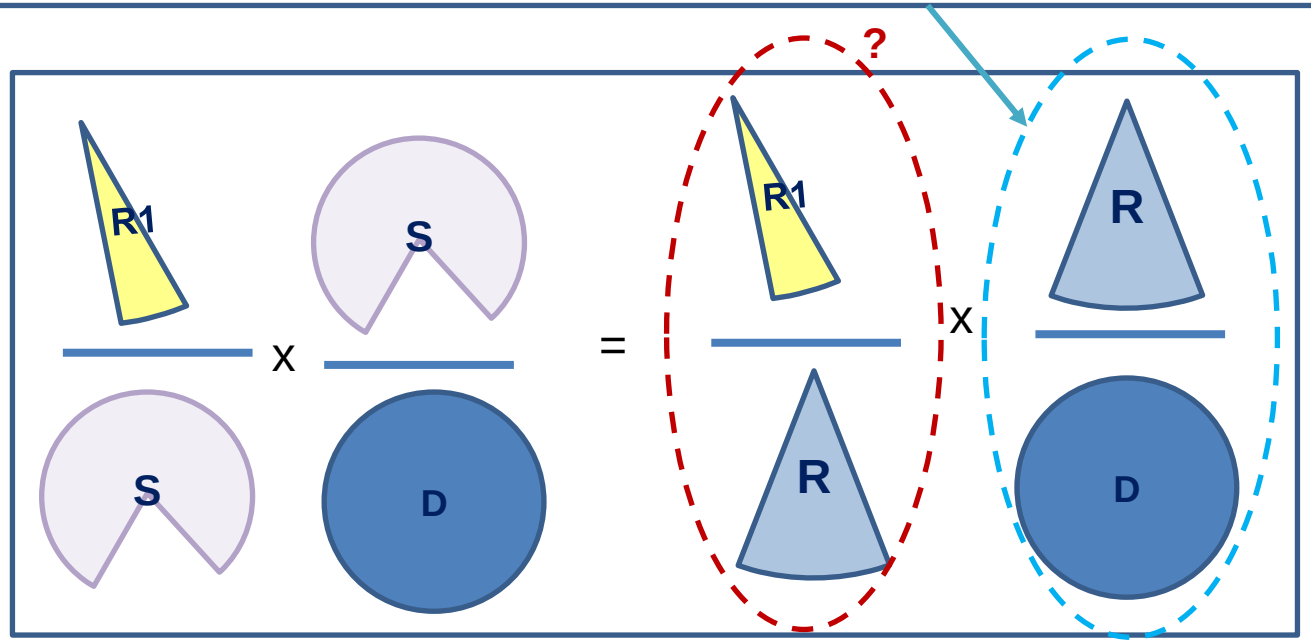
Geometric Visualization of Bayes' theorem

greatlearning









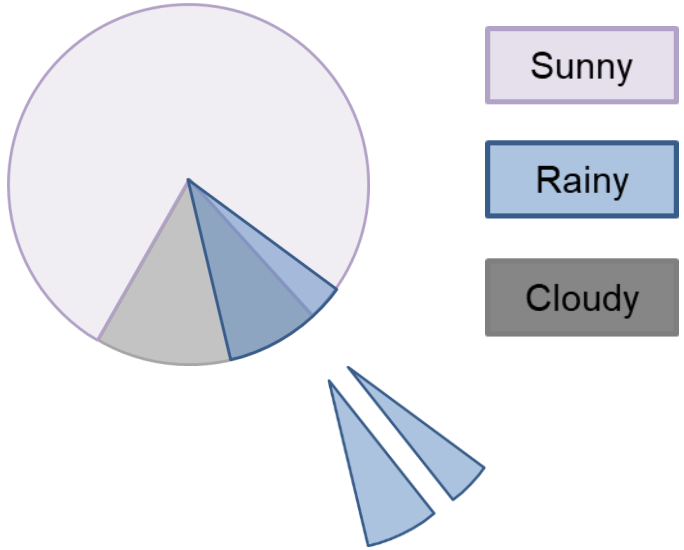
Simple Example of Bayes' theorem

greatlearning

P(Raining) = P(Raining|Sunny Day)*P(Sunny Day) + P(Raining|Cloudy day)* P(Cloudy day)

P(Raining ∩ Sunny) = P(Sunny ∩ Rainy)

P(Raining|Sunny)*P(Sunny) = P(Sunny|Raining)*P(Raining)



So,

P(R)=P(R|S)*P(S)+P(R|C)*P(C) &

P(R|S)*P(S) = P(S|R)*P(R) ➔ P(S|R) = P(R|S)*P(S)/ P(R)

Therefore, P(S|R) = P(R|S)*P(S) / {P(R|S)*P(S)+P(R|C)*P(C)} → Bayes' theorem

P(S|R) = P(Sunny | Raining)

Bayes' Theorem

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)}$$

- where:

  B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

  A = new event that might impact $P(B_i)$

Bayes Theorem Practical Example

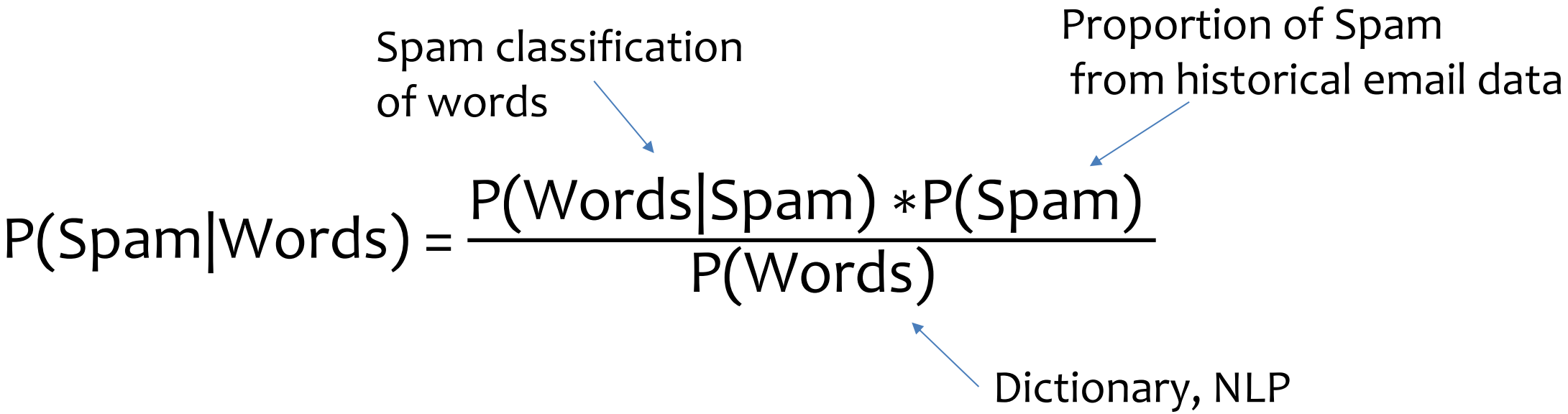
greatlearning

A lab is performing a test for a disease say “D” with two results “Positive” & “Negative”. They guarantee that their test result is 99% accurate. If a patient has the disease, they will give a positive test 99% of the time. If a patient does not have the disease, they will test negative 99% of the time. It is given that 3% of all the people have this disease in the city. A new patient goes to the Lab for a test and his test gives a “positive” result. What is the probability that the new patient actually has the disease?

Bayes' Theorem Practical Example

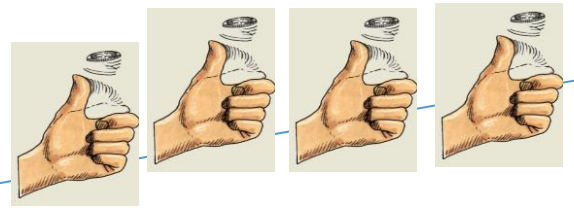
greatlearning

P("An Email is Spam")?



Probability Example (IPL tosses)

greatlearning
Learning for Life



Experiment

Coin Tosses in 4 IPL matches during a weekend

Coin Toss outcomes for the 4 matches are **Independent Events**

Coin Toss outcome for a match is **Mutually Exclusive**

Probability of this outcome = $P(H) = \frac{1}{2} = P(T)$

H H T T

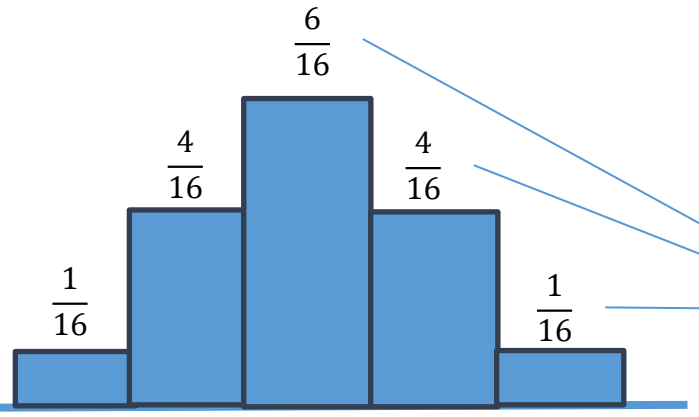
M1 M2 M3 M4

Outcome/ a combination of outcomes

Event

HHHH, TTTT, THHH, HTHH, HHTH, HHHT, HHTT, HTHT,
TTHH, THTH, HTTH, THHT, TTTH, HTTT, THTT, TTHT

Sample Space
(All possible Outcomes)



Possible Values of Event

$C(4,x)*p^x*(1-p)^{(4-x)}$

{x =0,1,2,3,4}
X→ # of Hs

Probability Distribution ⟶ Binomial Type

Binomial Distribution

greatlearning
Learning for Life

$C(n,x)*p^{x}*(1-p)^{(n-x)}$

x → number of successes
p → probability of a success
n → finite number of trials

Real World Examples:

1. Number of defective parts from an manufacturing line (Quality control measured say by ppm)
2. Estimating covid deaths in a locality (size of pandemic)
3. Number of fraudulent credit card transactions
4. Number of spam emails per day (gmail?)
5. Estimating medication side effects

The mean μ of the Binomial Distribution is given by $\mu = E(x) = np$

The Standard Deviation σ is given by

$$\sigma = \sqrt{np(1-p)}$$

For the example problem in the previous two slides,
Mean $= 7 \times 0.6 = 4.2$.
Standard Deviation $= \sqrt{4.2(1-0.60)} = 1.30$

Poisson Distribution

greatlearning
Learning for Life

Real World Examples:

1. Number of Calls per hour handled at a Call Center (# of call center operatives?)
2. Number of arrivals per hour/day at a Restaurant or Bank (Floor space and customer service windows?)
3. Number of website visits per unit time (bandwidth req?)
4. Number of network outages per week in a locality?
5. Number of cars passing thru a toll-gate?

Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where

P(x) = Probability of x successes given an idea of λ

λ = Average number of successes

e = 2.71828(based on natural logarithm)

x = successes per unit which can take values 0, 1, 2, 3,……..∞

λ is the Parameter of the Poisson Distribution.

Mean of the Poisson Distribution is = λ

Standard Deviation of the Poisson Distribution is = $\sqrt{\lambda}$

Normal Distribution

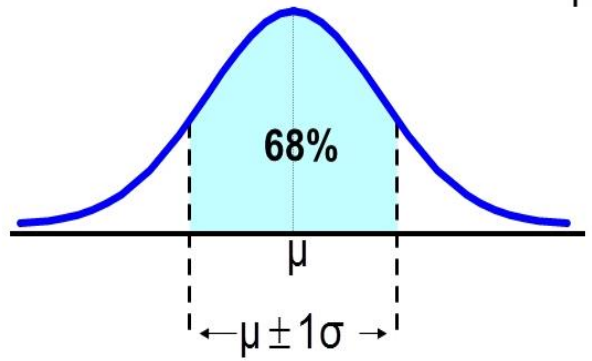
The Standard Normal Variable is defined as follows:

$$z = \frac{x - \mu}{\sigma}$$

Please note that Z is a pure number independent of the unit of measurement. The random variable Z follows a normal distribution with mean=0 and standard deviation =1.

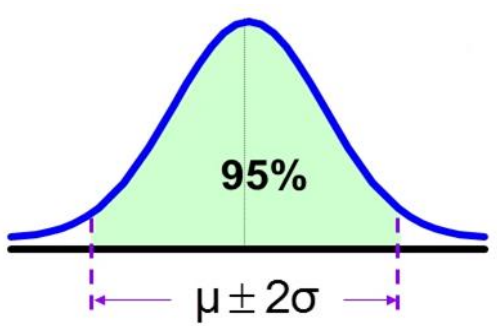
$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\left[\frac{Z^2}{2}\right]}$$

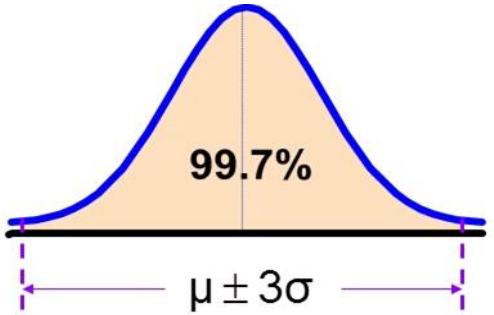
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$



Real World Examples:

1. Distribution of Student Marks in a class
2. Income distribution in an Economy
3. Rolling of 2 dices (fair)
4. Birth Weight

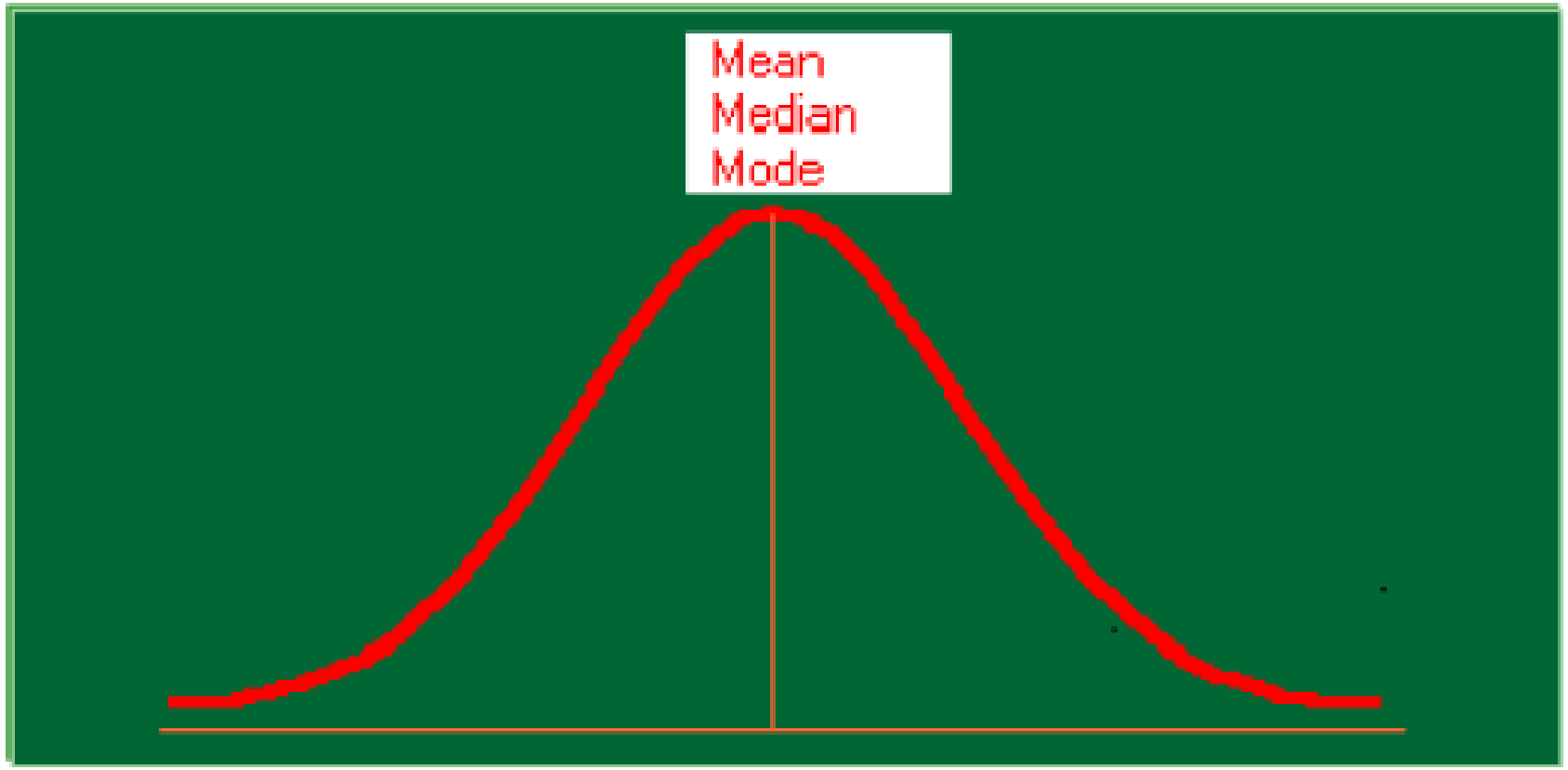




>>Go to Prob distribution Case studies

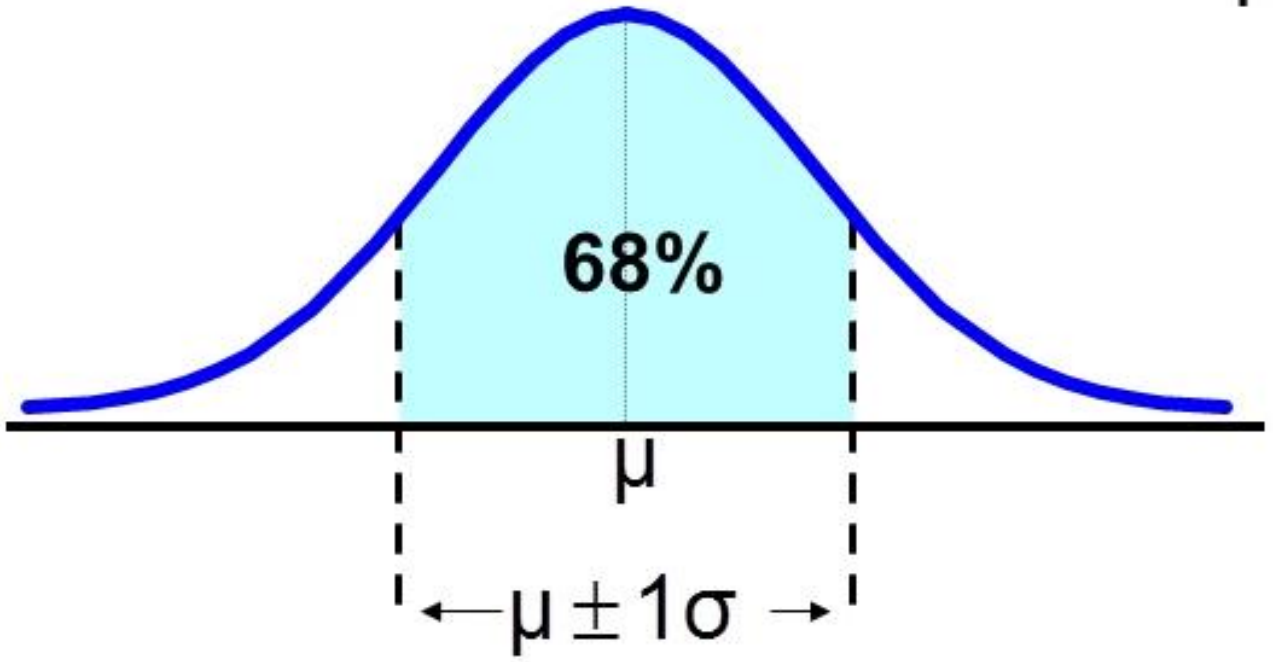
greatlearning
Learning for Life

Normal Distribution



Normal Distribution

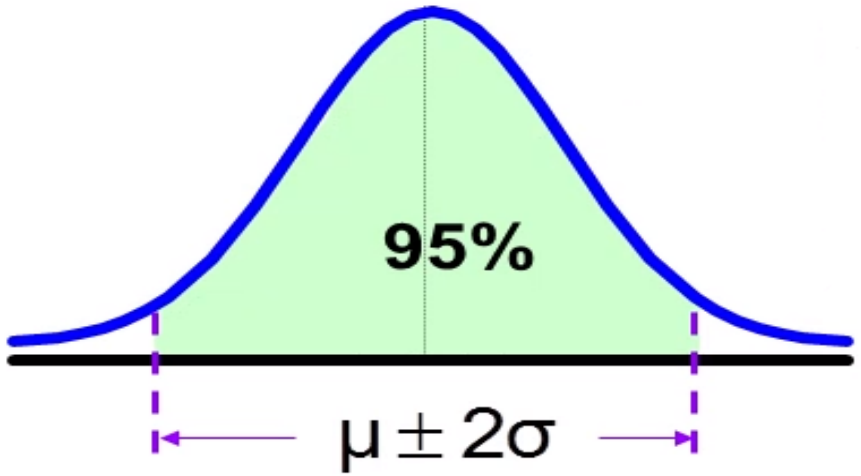
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$

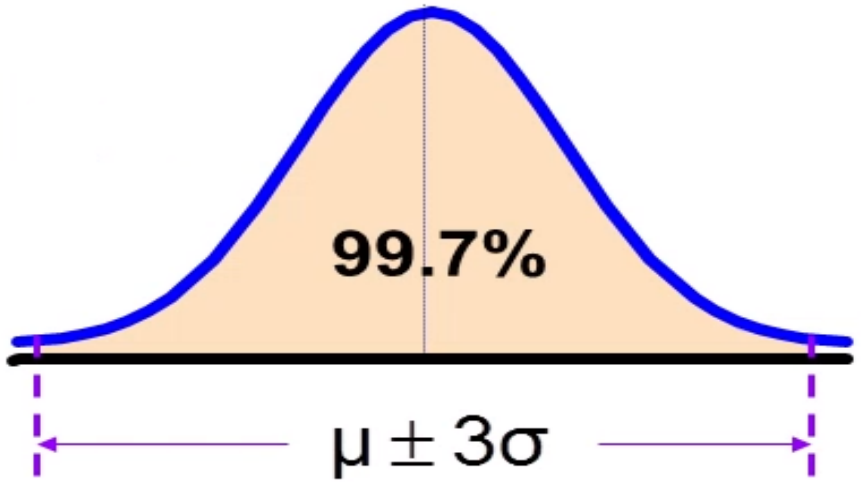


Normal Distribution

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$





Properties of Normal Distribution

greatlearning
Learning for Life

- The normal distribution is a continuous distribution looking like a bell. Statisticians use the expression "Bell Shaped Distribution".
- It is a beautiful distribution in which the mean, the median, and the mode are all equal to one another.
- It is symmetrical about its mean.
- If the tails of the normal distribution are extended, they will run parallel to the horizontal axis without actually touching it. (asymptotic to the x-axis)
- The normal distribution has two parameters namely the mean μ and the standard deviation σ

Standard Normal Distribution

The Standard Normal Variable is defined as follows:

$$Z = \frac{x - \mu}{\sigma}$$

Please note that Z is a pure number independent of the unit of measurement. The random variable Z follows a normal distribution with mean=0 and standard deviation =1.

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\left[\frac{Z^2}{2}\right]}$$

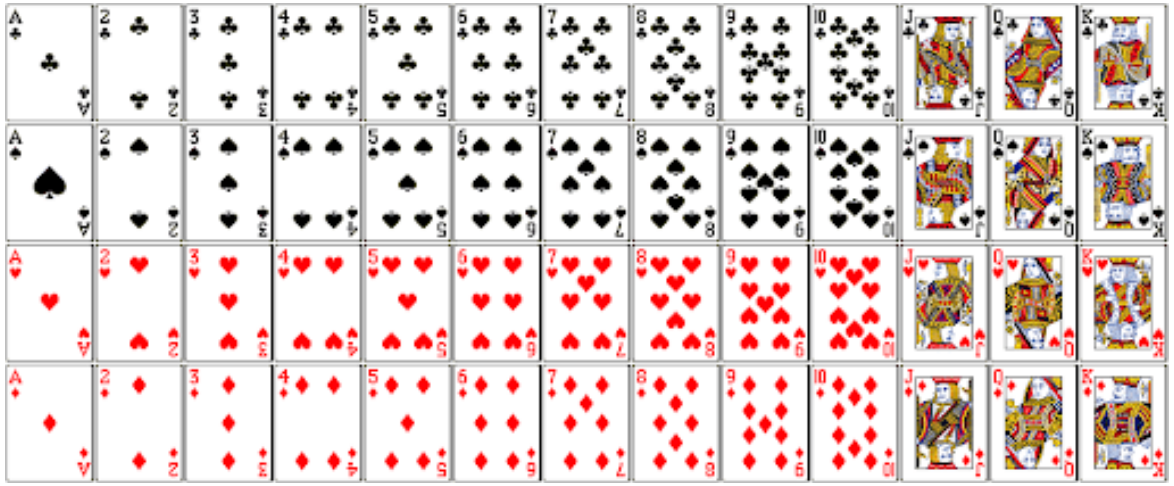
greatlearning
Learning for Life

Practice Exercises and Case Studies

Exercise 1 for Conditional Probability & Multiplication Rule

greatlearning
Learning for Life

Use P(A∩B)
or
P(AUB)



Identify the type of events and evaluate,

1. What is the probability that when the 2 identical dice are rolled together, the number you get as a sum of the numbers on the 2 dice-
    a. Is an even number?
    b. You get a total of 10?, 7?
    c. Of getting 2 odd numbers?

2. Can you construct the probability distribution for the all the outcomes (Sample space) considering the sum of numbers of the 2 dice?

3. When you draw 2 cards from a deck of 52 cards, one after the other without replacing, the probability of getting
    i. 2 consecutive diamonds?
    ii. An Ace and a Diamond considering both draws?
    iii. A King and a Queen considering both draws?

Bayes Theorem – Exercise 2

greatlearning

Bayes theorem

P(A|B), reads “A given B,” represents the probability of A if B was known to have occurred.

In many situations we would like to understand the relation between P(A|B) and P(B|A).

Practice Exercise

You are planning an outdoor event tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. Historically it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. What is the probability that it will rain tomorrow?

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

HINT:

P(F|R) = 0.9

P(F|Ř) = 0.1

P(R)=5/365

Exercise 3 for Binomial Distribution

greatlearning
Learning for Life

A bank issues credit cards to customers under the scheme of Master Card. Based on the past data, the bank has found out that 60% of all accounts pay on time following the bill. If a sample of 7 accounts is selected at random from the current database, construct the Binomial Probability Distribution of accounts paying on time.

Exercise 4 – Poisson Distribution

If on an average, 6 customers arrive every two minutes at a bank during the busy hours of working, a) what is the probability that exactly four customers arrive in a given minute? b) What is the probability that more than three customers will arrive in a given minute?

Exercise 5 - Normal distribution (to be reviewed beginning of week 3)

greatlearning
Learning for Life

At the Conclusion of a course in business statistics, a group of management students sat for a written examination. The results throw up the following information.

Marks obtained have a mean of 60 and a standard deviation of 12. There were 300 students who wrote the exam. The pattern of marks follows a normal distribution

a) What is the percentage of students who score more than 80

b) What is the percentage of students who score less than 50

c) What should be the distinction mark if the highest 10% of students are to be awarded distinction? Graphical method can be used for convenience.

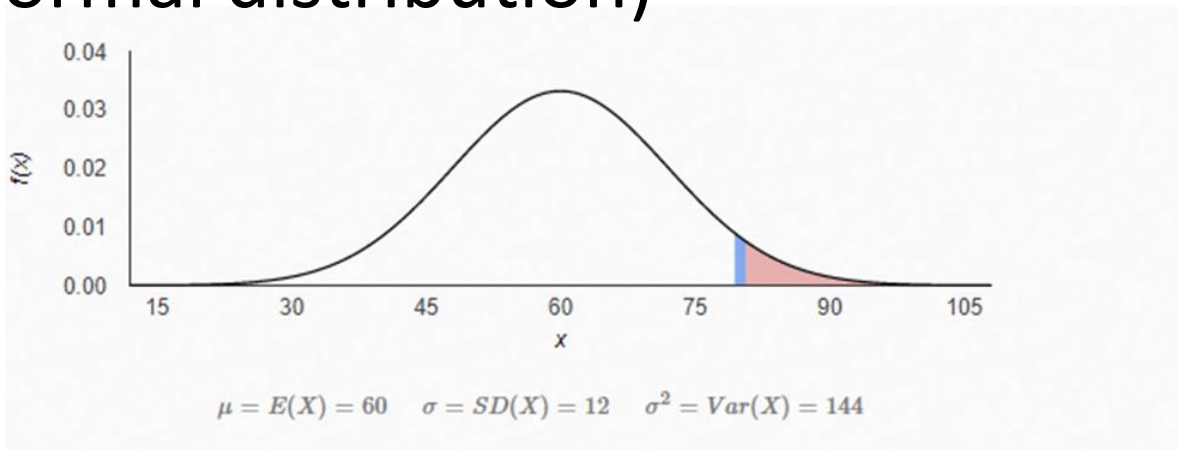
d) How many students score between 36 and 84 marks

(See solution hints in next slide after 1st trying to solve using stats.norm.cdf() and using Zscore table for (c))

Solution hints for Exercise 5 (Normal distribution)

greatlearning
Learning for Life

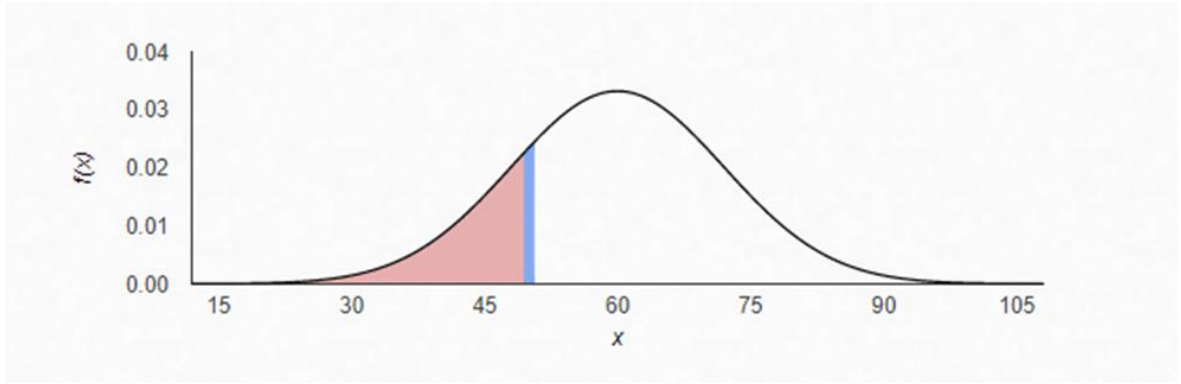
a



p(X>80)	p(z>1.67)	1-p(Z<1.67)

1-stats.norm.cdf(1.67)
0.0474

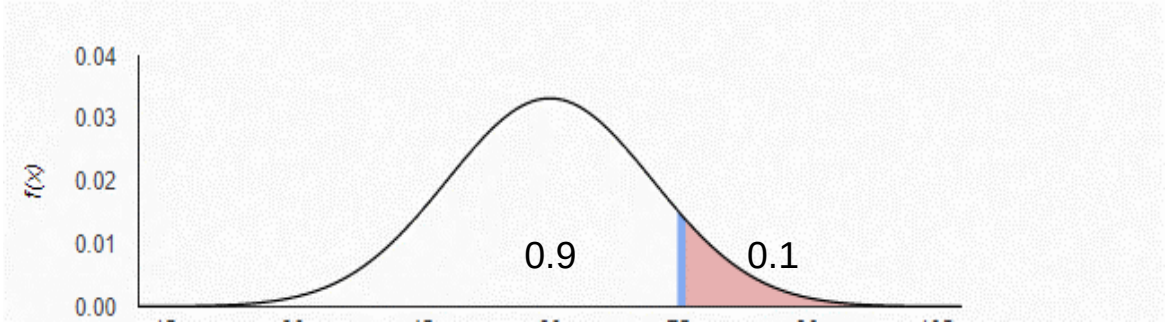
b



P(X<50)	P(z<-.833)

stats.norm.cdf(-.833)
0.2024

c



P(x>x1)	0.1	p(x<x1)	0.9
P(Z>Z1)			
Z1	1.28		
x1-mu/sigma	1.28		
x1	1.28*12+60		
	75.36		

X1	Z1

Applied Stats Concepts:
Correlation Analysis – Chi-Square Test

greatlearning
Learning for Life

Bivariate Plots –
Categorical-Categorical Variables

Family	Buyer of Car	Non-Buyer of Car	Total
Income per annum < 10L Rs			**120**
Income per annum ≥10L Rs			**80**
Total	**80**	**120**	**200**

		Chi2
48	38	2.08
32	42	3.13
72	82	1.39
48	38	2.08
		8.68

200 Urban Families

$$\chi^2 = \frac{(38-48)^2}{48} + \frac{(32-42)^2}{32} + \frac{(72-82)^2}{72} + \frac{(48-38)^2}{48} = 8.68 >> 1$$

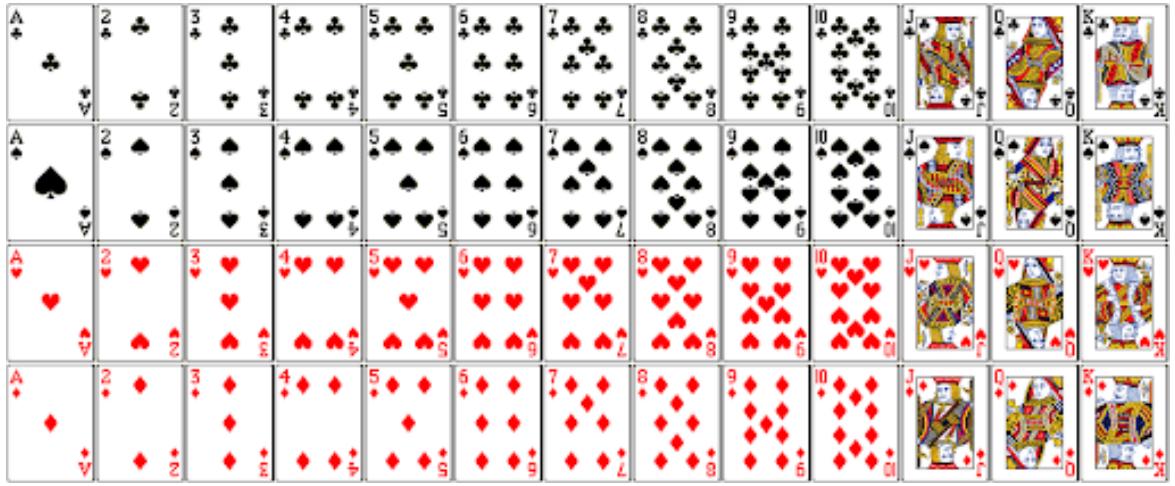
greatlearning
Learning for Life

Additional Content

Additional exercises for Conditional Probability & Multiplication Rule

greatlearning
Learning for Life

Use P(A∩B)
or
P(AUB)



Identify the type of events and evaluate,

1. What is the probability that when the 2 identical dice are rolled together, the number you get as a sum of the numbers on the 2 dice-
    a. Is an even number?
    b. You get a total of 10?, 7?
    c. Of getting 2 odd numbers?

2. Can you construct the probability distribution for the all the outcomes (Sample space) considering the sum of numbers of the 2 dice?

3. When you draw 2 cards from a deck of 52 cards, one after the other without replacing, the probability of getting
    i. 2 consecutive diamonds?
    ii. An Ace and a Diamond considering both draws?
    iii. A King and a Queen considering both draws?