ANALYTIXLABS

Data Science Foundations: Basic Statistics

https://www.analytixlabs.co.in/

Disclaimer: This material is protected under copyright act AnalytixLabs ©, 2011-2020. Unauthorized use and/ or duplication of this material or any part of this material including data, in any form without explicit and written permission from AnalytixLabs is strictly prohibited. Any violation of this copyright will attract legal actions

Know your Presenter: Chandra Mouli Kotta Kota

Chief Data Scientist

Areas of Expertise:

- Big Data Analytics
- Business Analytics
- Machine Learning
- Deep Learning
- Marketing Analytics
- Risk Analytics
- Operation Analytics
- Digital Analytics
- Business Intelligence

Chandra Mouli Kotta Kota is a former Business Consultant/Data Scientist and has worked with prestigious companies like McKinsey, Citigroup(E-serve), Genpact in the past 15 years. He has worked for clients across the globe and is an expert in Business and Big Data Analytics.

Professional Experience

- Worked on Marketing Analytics(CSI, CLM & Pricing), Risk Analytics(Credit Risk), Operation Analytics and Digital Analytics with focus on Retail/E-Commerce, Banking, Insurance, Telecom and Media clients in Asia, Australia, Europe, and US
- Hands on experience in development of Marketing & CRM Models (Acquisition, LTV Models, Cross Sell & Upsell, Attrition and MROI, Price and Promotion Models), Pricing Models(price & promo) and Risk Models (PD, LGD, EAD Models for Credit Cards, Consumer Loans and Insurance Portfolios)
- Hands on expertise in Big data and Multivariate analytical techniques including classical & machine learning algorithms including regression, instance based, regularization, Decision tree, Bayesian, clustering, Association rules, ANN, Deep learning and Ensemble algorithms
- Have used different statistical flat forms like SAS/SAS EM, R, Python, SPSS, SPSS Modeler, Hadoop, Spark, Tableau, Matlab, Julia, Salesforce, Sql, Excel, VBA, PowerBI, Cloud Platforms (AWS, GCP), MongoDB, Hbase .
- Trained and coached several client teams and various individuals on advanced analytics, Big data analytics tools and techniques as part of part of capability building programs
- Chandra Name is Featured as

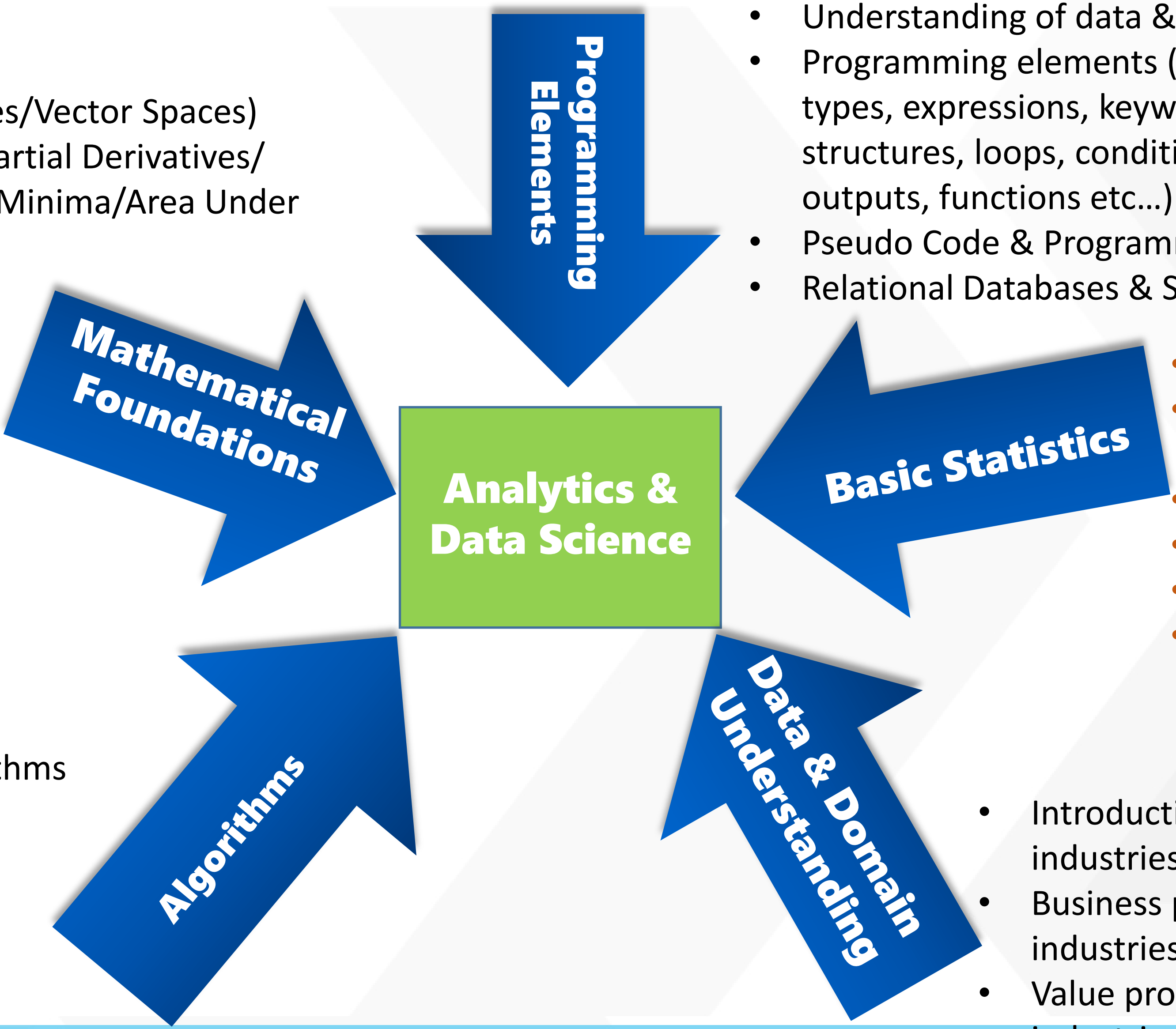
  - Top 10 data scientists (2015) in India by Analytics India Magazine (https://www.analyticsindiamag.com/top-10-data-scientists-in-india-2015/)
  - Featured as expert in Data Analytics & AI by whoknowsabout.com (https://whoknowsabout.com/blogpage/hidden-gems-boutiques-and-independent-consultants.html)

Academic Credentials

- Master of Science(Mathematics/Statistics): IIT-Madras, Chennai

Foundations of Analytics & Data Science

Foundations of Analytics & Data Science



- Linear Algebra (Matrices/Vector Spaces)
- Calculus (Derivatives/Partial Derivatives/ Integration/Maxima & Minima/Area Under the curve
- Theory of Optimization

- Introduction to Basic Analytics Tools: Excel
- Understanding of data & storage
- Programming elements (variables, constants, data types, expressions, keywords, comments, data structures, loops, conditional statements, inputs, outputs, functions etc…)
- Pseudo Code & Programming Languages
- Relational Databases & SQL

- **Basic Statistics**
- **Measures of central tendencies/ variance /Frequency/Rank**
- **Probability, Distributions,**
- **Conditional Probability**
- **Relationships**
- **Others: CLT, Confidence Intervals, Hypothesis testing etc..**

- Design of Algorithms
- Various types of Algorithms

- Introduction to data models in various industries & functions
- Business problems (Pain points) in various industries
- Value proposition of Analytics across industries & functions

What is Statistics?

There are many definitions available to define statistics. The below are few of them.

- **Statistics is a set of tools used to collect, organize, analyze and visualize data.**

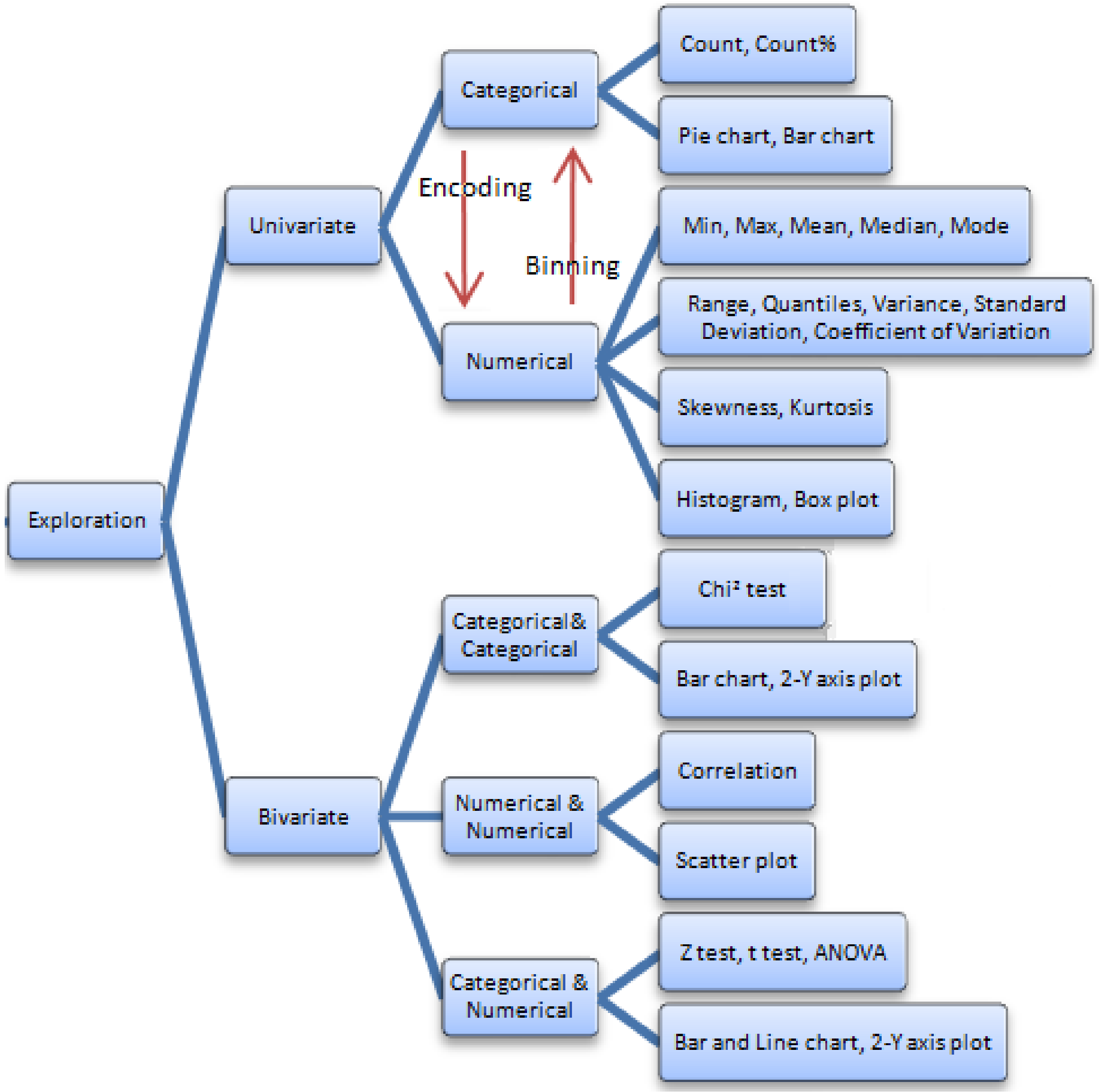
- **Statistics is a field of study using mathematical procedures to make inferences from data and decisions based on it**

Data Science Foundations: Basic Statistics

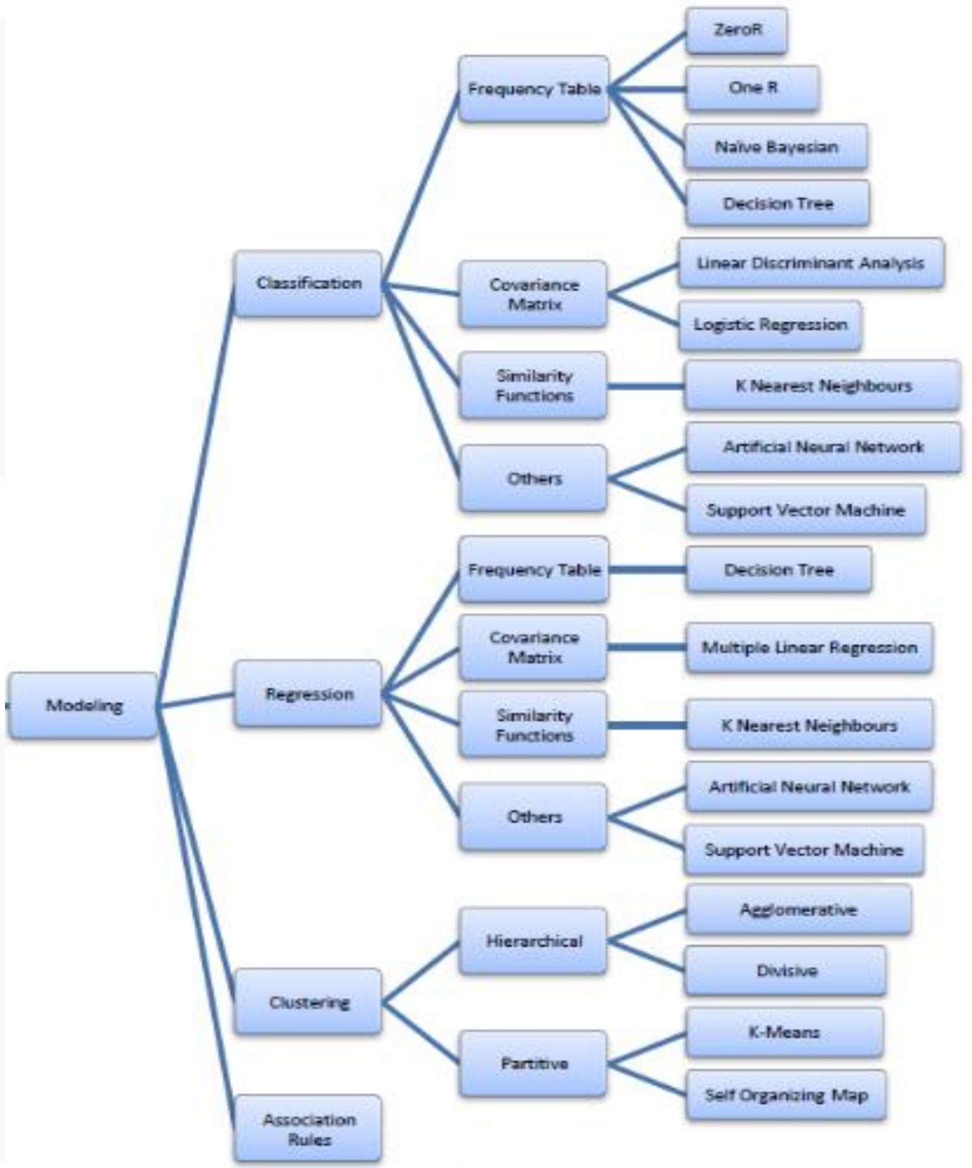
Statistics - Foundations

- What is Statistics? Why do we need?
- Types of Data (Continuous, Categorical (Ordinal & Nominal)
- Measures of central tendency (Mean vs. Median vs. Mode)
- Measure of frequency (Count, Percent, Frequency)
- Measure of position (Percentile Ranks, Quartile Ranks, Positions)
- Measures of variation/Spread (Range vs. Variance vs. Standard Deviation vs. Coefficient of Variation)
- Uni-Variate analysis vs. Bivariate analysis vs. Multivariate analysis?
- Probability & Probability Distributions (Discrete Vs. Continuous)
- Conditional Probability & Bayes Theorem
- How to identify relationship between different variables?
- Correlations & Covariance
- Descriptive stats vs. inferential statistics?
- Population vs. sample?
- Estimation of population?
- Sampling & Design of experiments

Basic Statistics – Key Topics



- Exploration
  - Univariate
    - Categorical
      - Count, Count%
      - Pie chart, Bar chart
    - Encoding
    - Binning
    - Numerical
      - Min, Max, Mean, Median, Mode
      - Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
      - Skewness, Kurtosis
      - Histogram, Box plot
  - Bivariate
    - Categorical & Categorical
      - Chi² test
      - Bar chart, 2-Y axis plot
    - Numerical & Numerical
      - Correlation
      - Scatter plot
    - Categorical & Numerical
      - Z test, t test, ANOVA
      - Bar and Line chart, 2-Y axis plot



- Modeling
  - Classification
    - Frequency Table
      - ZeroR
      - One R
      - Naïve Bayesian
      - Decision Tree
    - Covariance Matrix
      - Linear Discriminant Analysis
      - Logistic Regression
    - Similarity Functions
      - K Nearest Neighbours
    - Others
      - Artificial Neural Network
      - Support Vector Machine
  - Regression
    - Frequency Table
      - Decision Tree
    - Covariance Matrix
      - Multiple Linear Regression
    - Similarity Functions
      - K Nearest Neighbours
    - Others
      - Artificial Neural Network
      - Support Vector Machine
  - Clustering
    - Hierarchical
      - Agglomerative
      - Divisive
    - Partitive
      - K-Means
      - Self Organizing Map
  - Association Rules

How Basic Stats helps in Business Analytics/Data Science?

- Identifying Business Problem & type of problem
- Diagnostics Analysis/Root cause Analysis
- Summarize/Aggregate the data
- Data Understanding & Data Audit/Data Quality
- Understanding distributions & Identify patterns
- Define & Measure Key performance Metrics (KPI’s)
- Understand driving factors
- Cause & effect analysis
- Estimating population
- Designing experiments
- Prove or Disprove hypothesis
- Understand relationships among data (features/variables)
- Measure Effectiveness of an event (like campaigns, Changes in the organization etc)

Etc…

https://baselinesupport.campuslabs.com/hc/en-us/sections/200909805-Analyzing-Data-and-Reporting-Capabilities

Modules

Module-0

- What is Statistics?
- Definition of Statistics
- What topics comes under Foundations of Statistics
- Applications of Basic statistics in Business Analytics/Data Science

Module-1

- Types of Data (Continuous, Categorical (Ordinal & Nominal)
- Measures of central tendency (Mean vs. Median vs. Mode)
- Measure of frequency (Count, Percent, Frequency)
- Measure of position (Percentile Ranks, Quartile Ranks, Positions)
- Measures of variation/Spread (Range vs. Variance vs. Standard Deviation vs. Coefficient of Variation)

Module-2

- Uni-Variate analysis vs. Bivariate analysis vs. Multivariate analysis?
- How to identify relationship between different variables?
  - Correlations & Covariance
  - F-Statistics
  - Chi-squre Statistics

Module-3

- What is Probability?
- Conditional Probability
- What is Bayes Theorem?
- What is Probability Distributions?
- Types of Distribution(Discrete Vs. Continuous)
- Popular Distributions

Probability

Probability is simply how likely something is to happen.

The best example for understanding probability is flipping a coin:
There are two possible outcomes—heads or tails.

Probability = No.of possibility met by condition/Total possibilities

The probability of an event can only be between 0 and 1 and can also be written as a percentage.

The probability of event *A* is often written as P(A)

If P(A) > P(B), then event *A* has a higher chance of occurring than event *B*.

If *P*(*A*)=*P*(*B*), then events *A* and *B* are equally likely to occur.

Module-4

- Types of Statistics
    - Descriptive stats vs. inferential statistics?
- Population vs. sample?
- Types of Sampling Methodologies
- What is Estimation of population?

Module-5

- Data Visualization
- Uni-variate Analysis
  - Categorical variable : Bar Chart, Pie Chart, Pareto Chart
  - Numerical Variable: Box Plot, Histogram/Density Plot
- Bivariate Analysis
  - Categorical-Categorical: Stacked Chart
  - Numerical – Numerical: Scatter Plot
  - Categorical – Numerical: Panel Chart

Key Terminology We discussed till now

- What is Statistics, Application of Statistics
- Types of Data: Numerical/Continuous/Ratio, Categorical (Nominal, Ordinal)
- Measures of Central Tendencies: Mean, Median, Mode
- Measures of Frequency: N, Nmiss, Frequency, Proportions (Percentages), Cumulative Percentage
- Measures of Spread: Variance, Standard Deviation, Inter Quartile Range (IQR), Range, Coefficient of variation
- Measures of Rank: Percentiles, Quartiles
- Measures of Symmetry: Skewness (Skewness ~0, symmetric, else non-symmetric)
- Measures of Peaked ness: Kurtosis
- Types of Analysis (Univariate, Bivariate, Multivariate)
- Metrics related to Relationships (Chi-Square, Correlation, F-Statistic)
- Probability, Probability Distribution
- Conditional Probability, Bayes Theorem
- Types of Distribution (Discrete, Continuous)
- Popular Distributions (Normal, Log Normal, Uniform, Binomial, Bernoulli, Poisson etc…)
- Types of Statistics (Descriptive, Inferential)
- Types of Estimation (Point Estimation, Range Estimation)
- Population vs. Sample
- Types of Sampling Methodologies (Simple Random Sample vs. Stratified Sampling)
- How to Estimate Population using Sample?

Quick Summary of Analysis & Visualizations

Univariate Analysis:

- Categorical Variable:
  - N
  - N Missing's
  - Proportions
  - Mode
  - Pareto Analysis
  - Outliers (Percentage of observations with <0.5% of observations)
  - 80-20 Rule (20% of categories represents 80% of data)
  - **Visualization:** Bar Chart, Pie Chart, Pareto Chart
- Numerical Variables:
  - N, Nmiss
  - Central Tendencies: Mean, Median
  - Distribution: Min, P1, P5, P10, P25, P50, P75, P90, P95, P99, Max
  - Spread: Range, Variance, Std, CV
  - Symmetry/Peakedness: Skewness, Kurtosis
  - **Visualization:** Histogram/Density plot, Box Plot
- Time Variable:
  - Time Periods, Seasonality, Trend
  - **Visualization:** Line Chart

Bi-Variate Analysis:

- Categorical – Categorical
  - Cross Tab Analysis (N, Proportions, Column %, Row %)
  - Chisqure to find the relationship
  - **Visualization:** Stacked Chart
- Categorical – Numerical
  - Aggregation (Group by)
  - F-Statistics to find the relationship
  - **Visualization:** Panel Chart
- Numerical – Numerical
  - Correlation, Covariance
  - **Visualization:** Scatter Plot

References

- https://www.saedsayad.com/
    - Click on the topic in the above link to deep dive it.

Exercise

Descriptive Analytics (Exploratory Analysis) for Insurance Data

- Perform any exploratory Analysis as per your understanding?

Examples of few analysis:

- What is the breakdown of the number of Customers by renew offer type and response?
- What is the breakdown of the number of Customers by renew offer type and response where response is Yes?
- Team want to Know where they are responding from?
- What is the breakdown of the number of customers by policy Type and State?
- What is the breakdown of the number of customers by policy and State?
- What are the values of Customer Life time value and customer by state?
- Identify the right target market and course-correct (if required):

Apendix

Univariate – BiVariate - Multivariate

- A variate is the key focus of the experiment and this varies from one observation to other
  - Univariate, Bi-Variate, Multi-Variate

Product	Weight
P1	54
P2	36
P3	41
P4	26
P5	54
P6	39
P7	29
P8	[illegible]

Univariate

Product	Weight	Volume (cc)
P1	30	347
P2	53	548
P3	21	256
P4	23	278
P5	38	485
P6	25	[illegible]
P7	27	328
P8	[illegible]	264

Bi-Variate

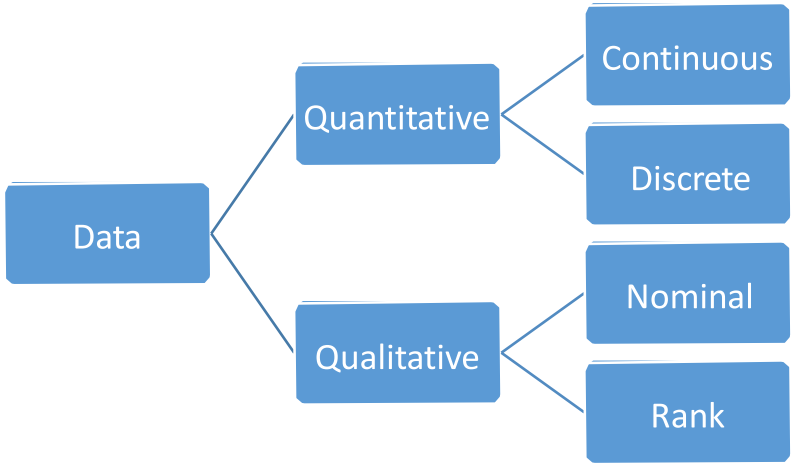
Product	Weight	Volume (cc)	Price	Demand	Tax
P1	24	336	$ 42.0	5889	$ 58.0
P2	53	307	$ 11.0	5908	$ 50.0
P3	33	433	$ 59.0	4987	$ 60.0
P4	58	295	$ 54.0	4715	$ 60.0
P5	37	365	$ 21.0	3785	$ 58.0
P6	25	377	$ 17.0	[illegible]	$ 52.0
P7	40	515	$ [illegible]	3197	$ 54.0
P8	52	309	[illegible]	2081	$ 59.0

Types of Data

Categorical data

Is non-numeric, can be observed but not measured

E.g. Favorite color, Place of Birth



Data

Quantitative

Continuous

Discrete

Qualitative

Nominal

Rank

Quantitative data

Is numerical data which can be measured

Discrete

Random variable which takes only isolated values in its range of variation. For example number of heads in 10 tosses of a coin

Continuous

Random variable which takes any value in its range of variation. For example, height of a person

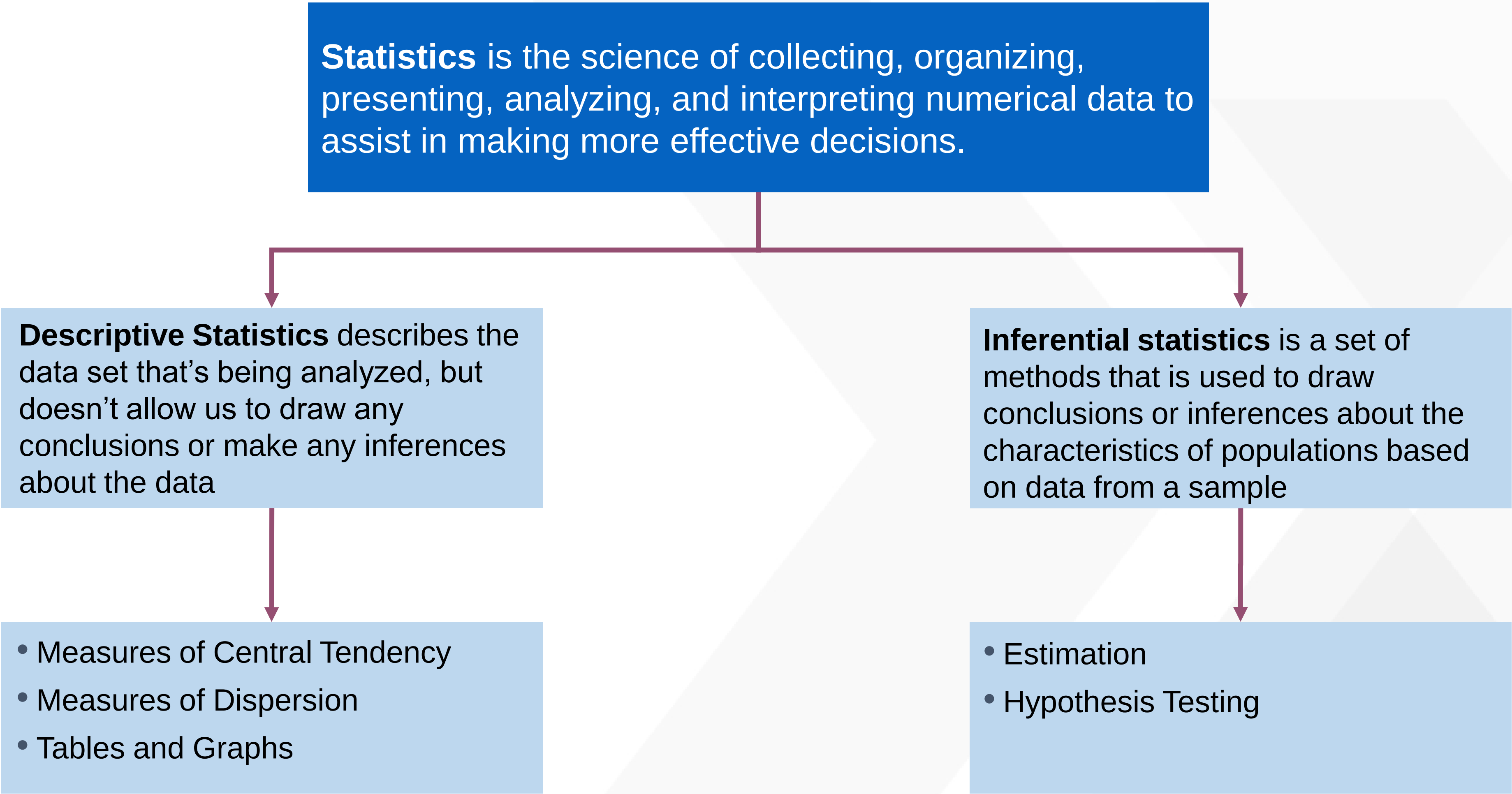
Nominal

- Values do not have ordering
- Example categorical variables like color, nationality and so on

Ordinal

- Values are ordered
- Example Satisfaction scores

Types of Statistics



Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data to assist in making more effective decisions.

Descriptive Statistics describes the data set that's being analyzed, but doesn't allow us to draw any conclusions or make any inferences about the data

- Measures of Central Tendency
- Measures of Dispersion
- Tables and Graphs

Inferential statistics is a set of methods that is used to draw conclusions or inferences about the characteristics of populations based on data from a sample

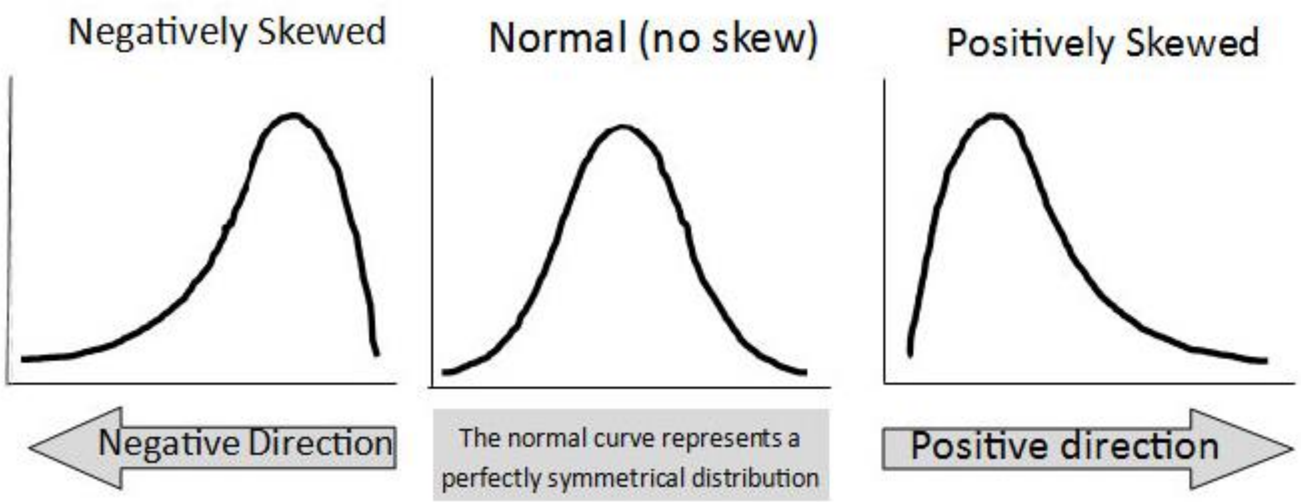
- Estimation
- Hypothesis Testing

Descriptive Statistics

Central Tendency: is the middle point of distribution. Measures of Central Tendency include Mean, Median and Mode

Dispersion: is the spread of the data in a distribution, or the extent to which the observations are scattered.

Skewness: When the data is asymmetrical ie the values are not distributed equally on both sides. In this case, values are either concentrated on low end or on high end of scale on horizontal axis.



If the trail is to the right or positive end of the scale, the distribution is said to be “positively skewed”.
If the distribution trails off to the left or negative side of the scale, it is said to be “negatively skewed”.

Measures of Central Tendency

MEAN

It is just the average of the data, computed as the sum of the data points divided by the number of points

- ➕ It is the easiest metric to understand and communicate
- ➖ Mean is prone to presence of outliers

Example: What is a typical student in the class doing?

MEDIAN

It is the value in the middle of the data set, when the data points are arranged from smallest to largest.

Tricky circumstances:

If there is an even number of data points, you will need to take the average of the two middle values.

- ➕ Median is a more "robust" to presence of outliers
- ➖ It is more complicated to communicate

Example: To compare performance of any single student against group

MODE

It is the most common value in the data set.

Tricky circumstances:

If no value occurs more than once, then there is no mode

If two, or more, values occur as frequently as each other and more frequently than any other, then there are two, or more, modes.

- ➕ Not very practical since it is affected by skewness
- ➖ Most real life distributions are multimodal

Example: A parent wanting to know whether their child is better or worse than typical child at his grade level

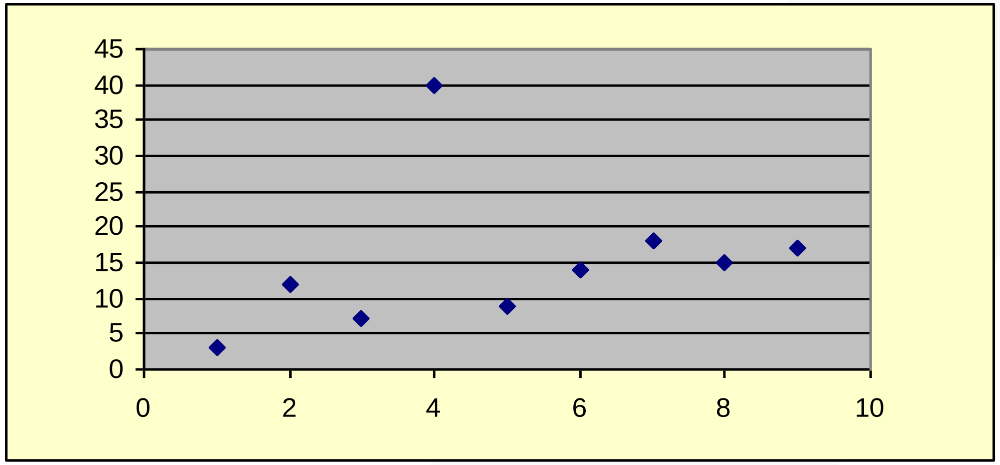
Outliers

An outlier is an observation that is numerically distant from the rest of the data.

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution.

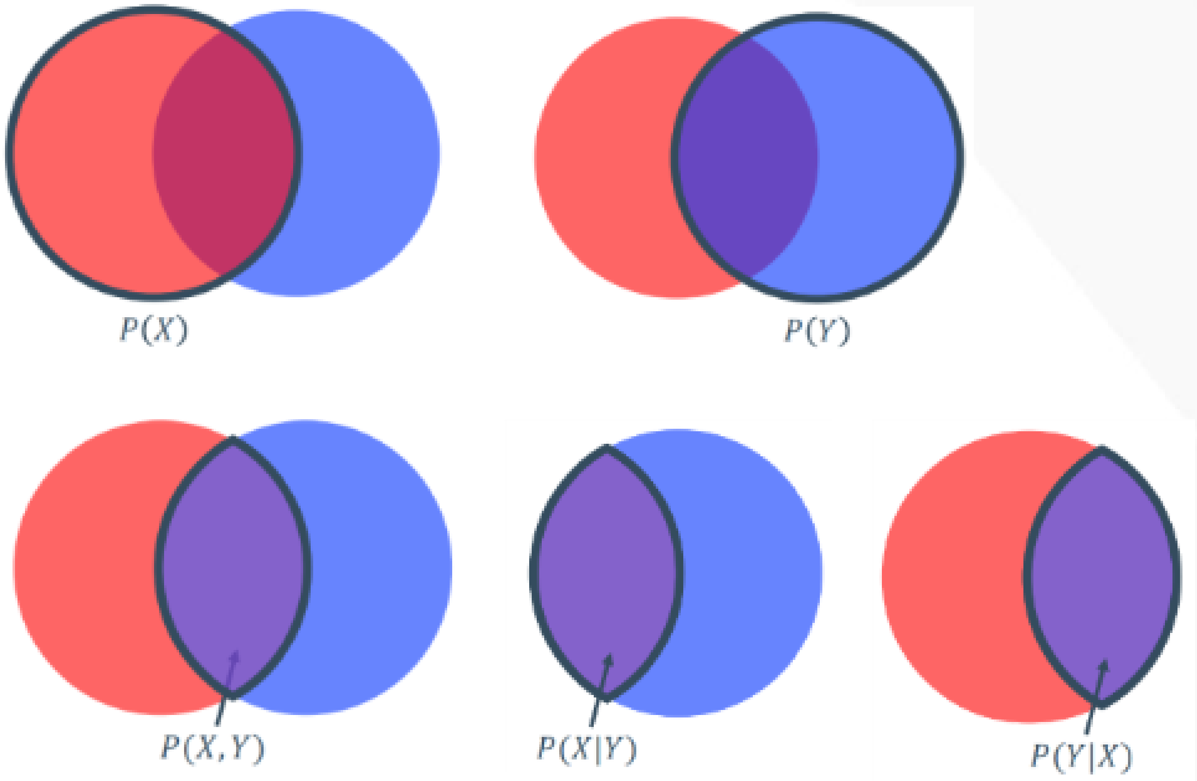
Example: Bill Gates makes $500 million a year. He's in a room with 9 teachers, 4 of whom make $40k, 3 make $45k, and 2 make $55k a year. What is the mean salary of everyone in the room? What would be the mean salary if Gates wasn't included? Mean <u>With</u> Gates: **$50,040,500** Mean <u>Without</u> Gates: **$45,000**

A **Scatterplot** is useful for "eyeballing" the presence of **outliers**.



We can also use Standard Deviation to identify Outliers!

Probability Basics



LinSingle event probability:

$$P(X)\ ,P(Y)$$

Joint event probability:

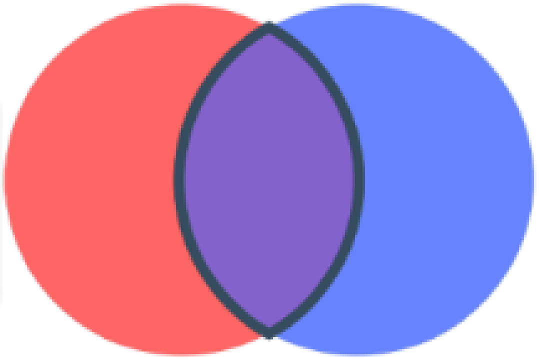
$$P(X,Y)$$

Conditional probability:

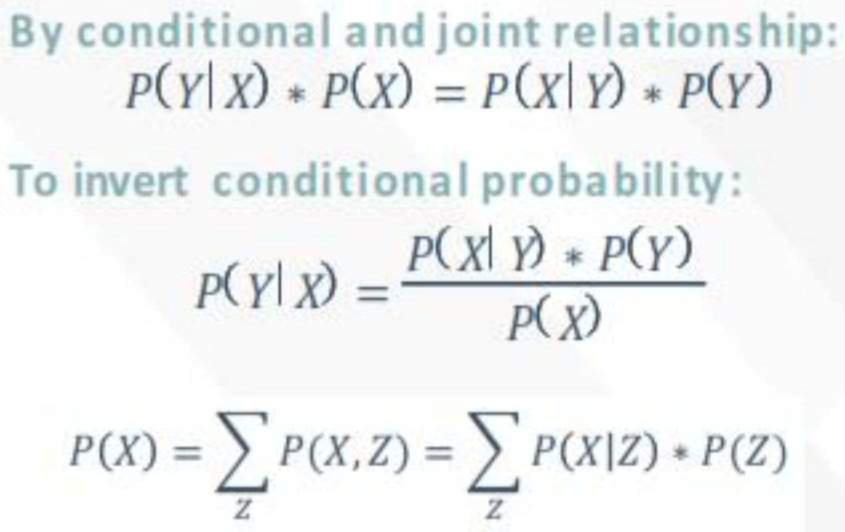
$$P(X|Y)\ ,P(Y|X)$$

Joint and conditional relationship:

$$P\ (X,Y\ )=P\ (Y|X)*P\ (X)=P\ (X|Y)\ *\ P(Y)$$



Probability Basics



By conditional and joint relationship:

$$P(Y|X) * P(X) = P(X|Y) * P(Y)$$

To invert conditional probability:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

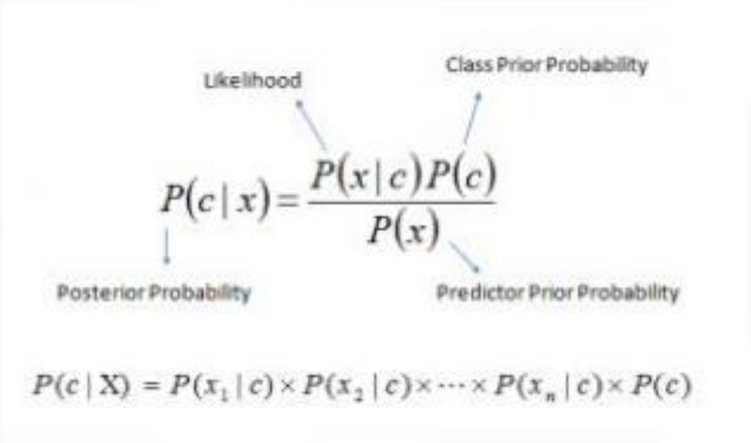
$$P(X) = \sum_{Z} P(X,Z) = \sum_{Z} P(X|Z) * P(Z)$$

Bayes Theorem

- Given training data D, posterior probability of a hypothesis c, P(c|x), follows the Bayes theorem

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

$$posterior = \frac{likelihood * prior}{evidence}$$



Likelihood

Class Prior Probability

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

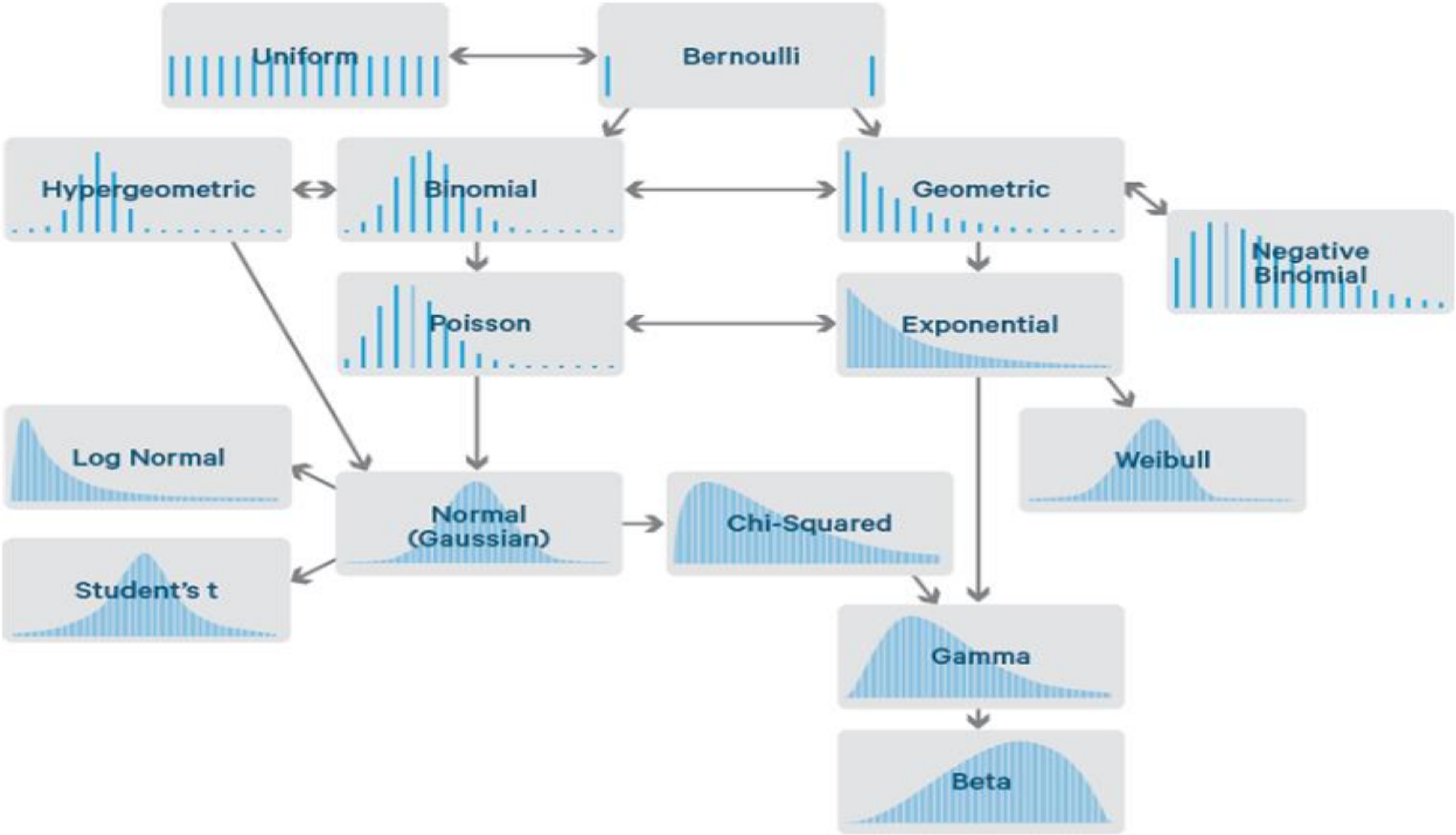
Posterior Probability

Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

- The goal is to identify the hypothesis with the maximum posterior probability
- Practical difficulty: require initial knowledge of many probabilities, significant computational cost

Popular Probability Distributions



Some useful distributions

Discrete:

- Binomial Distribution
  - The binomial probability distribution is useful when a total of *n* independent trials are conducted and we want to find out the probability of *r* successes, where each success has probability *p* of occurring.
- Poisson Distribution
  - The basis for the calculation of the probability of rare events

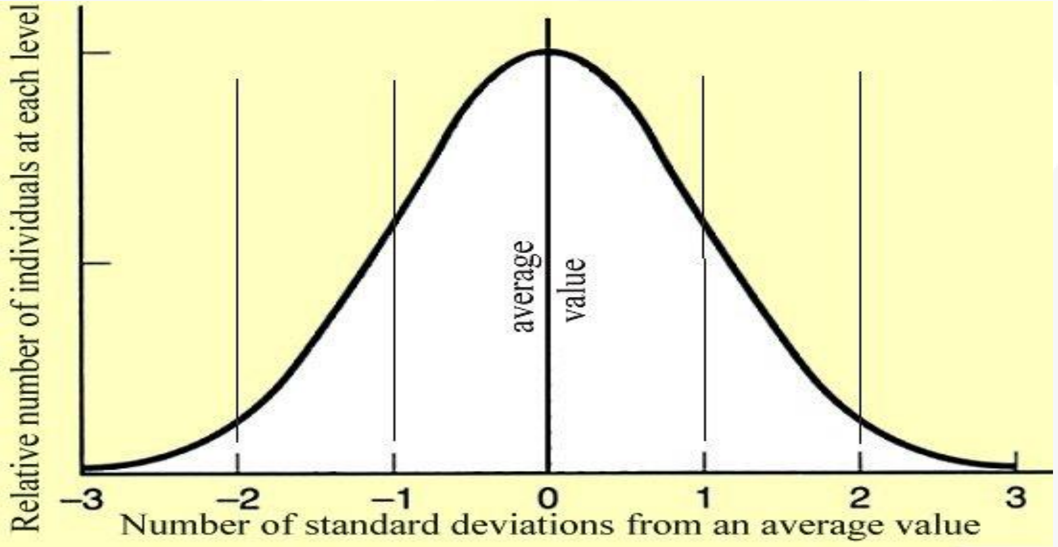
Continuous:

- Exponential Distribution
  - Time between two successive m/c failures (by accident and not by wear & tear)
- Weibull Distribution
  - Reliability and warranty analysis
- Gamma Distribution
- Insurance & weather prediction and analysis
- Normal Distribution

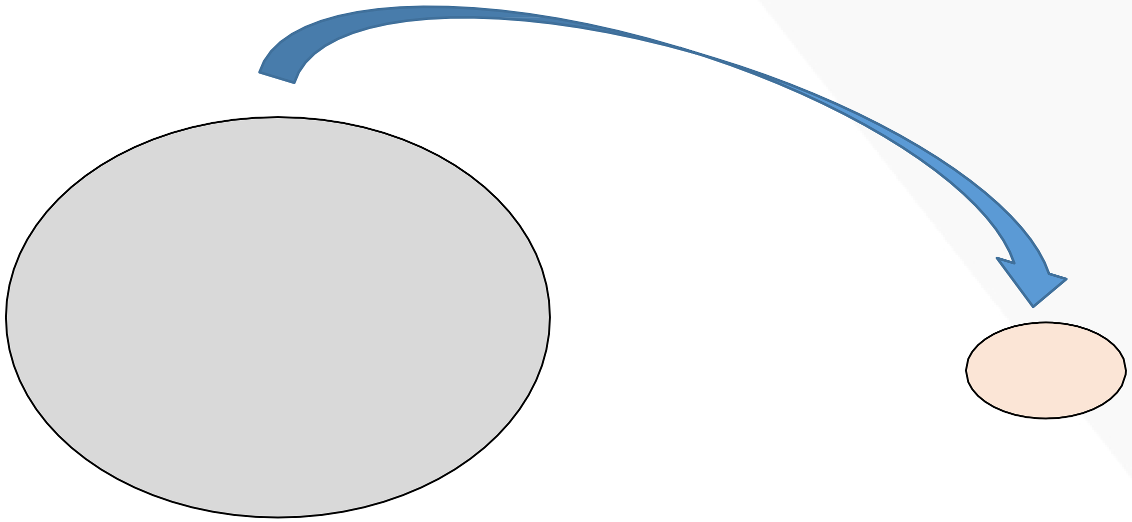
Normal Distribution

Normal distribution is a pattern for the distribution of a set of data which follows a ***bell shaped curve***. This also called the *Gaussian distribution*.
The normal distribution is a theoretical ideal distribution. Real-life empirical distributions never match this model perfectly. However, many things in life do approximate the normal distribution, and are said to be "normally distributed."

- Normal Distribution has the mean, the median, and the mode all coinciding at its peak
- The curve is concentrated in the center and decreases on either side ie most observations are close to the mean
- The bell shaped curve is symmetric and Unimodal
- It can be determined entirely by the values of mean and std dev
- Area under the curve = 1
- **The empirical *68-95-99.7 rule* states that for a normal distribution:**
  - **68.3% of the data will fall within 1 SD of mean**
  - **95.4% of the data will fall within 2 SD's of the mean**
  - **Almost all (99.7%) of the data will fall within 3 SD's of the mean**



What is Population & Sample?



Population

Sample

Population **Parameter**	Sample **Statistic**
μ, σ^2, σ	X **s^2, s,**

Simple random sample

&

Stratified random sample

Populations and Samples

So far we have determined the results associated with individual observations or sample means when the true population parameters are known. In reality, the true population parameters are seldom known. We now learn how to infer **levels of confidence**, or a measure of accuracy on parameters, estimated using samples

POINT ESTIMATOR

- If we take a sample from a population, we can estimate parameters from the population, using sample statistics
- Example: Sample mean (x) is our best estimate of the population mean (µ)
- Whereas, we really don’t know how close the estimate is to the true parameter
- The mean annual rainfall of Melbourne is 620mm per year

INTERVAL ESTIMATOR

- If we estimate a range or interval within which the true population parameter lies, then we are using an interval estimation method
- This is the most common method of estimation. We can also apply a level of how confident we are in the estimate
- In 80% of all years Melbourne receives between 440 and 800 mm rain

Sampling methodologies

Sampling is required because it is seldom possible to measure all the individuals in a population. Researchers hence, use samples and infer their results to the population of interest

Eg: Election polls, market research surveys, etc

For a sample to be a "good sample", it is imperative that there is a good sample size and there is no biasness in the sample.

Simple Random Sample

is one in which every member of the population is equally likely to be measured

Eg: Allocate a number to each member of the population and use a random number generator to determine which individuals will be measured

Stratified Sampling

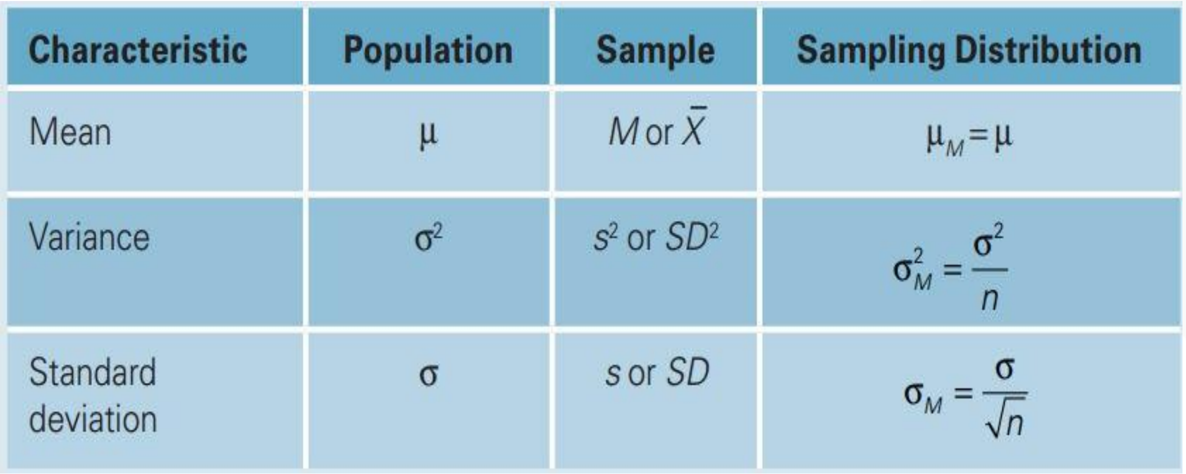
separates the population into mutually exclusive groups and randomly samples within the groups

Eg: Randomly select a number of people within each demographic cell, while maintaining overall proportions like gender ratio, income ratio, etc

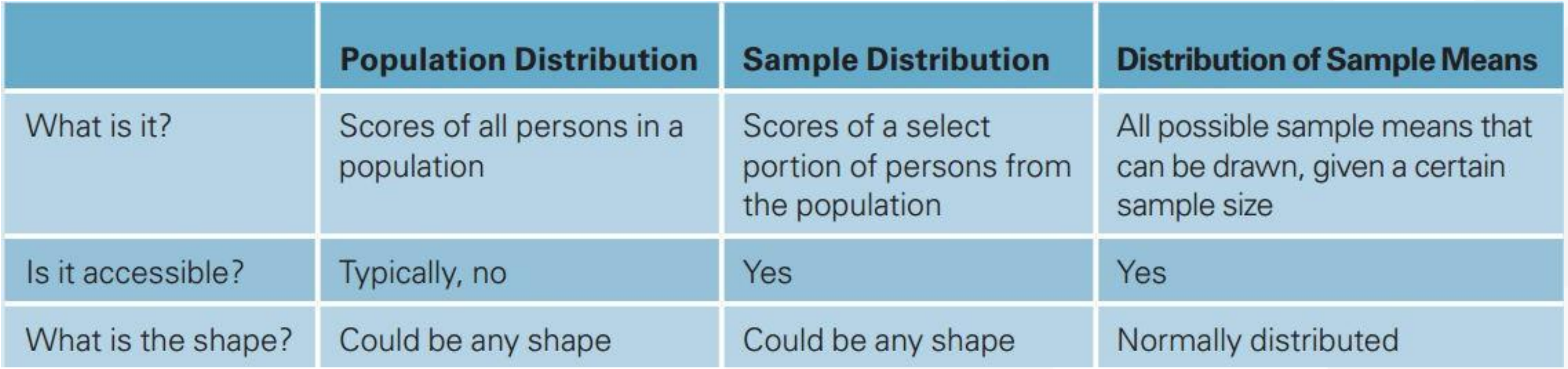
Other methodology

Cluster sampling: is used when there is a considerable variation within each group but the groups are essentially similar to each other. Here we divide the population into groups, or clusters, and then select a random sample of these clusters.

Characteristics of the Sampling Data



Characteristic	Population	Sample	Sampling Distribution
Mean	μ	M or $\overline{X}$	$\mu_M = \mu$
Variance	σ^2	s^2 or SD^2	$\sigma_M^2 = \frac{\sigma^2}{n}$
Standard deviation	σ	s or SD	$\sigma_M = \frac{\sigma}{\sqrt{n}}$



	Population Distribution	Sample Distribution	Distribution of Sample Means
What is it?	Scores of all persons in a population	Scores of a select portion of persons from the population	All possible sample means that can be drawn, given a certain sample size
Is it accessible?	Typically, no	Yes	Yes
What is the shape?	Could be any shape	Could be any shape	Normally distributed

Correlation

What is the relationship between two variables?

Relationship between hours studying (X) and grades on a midterm (Y)?

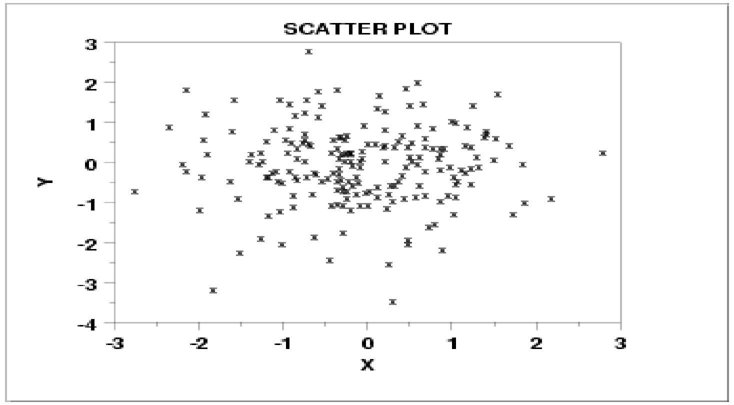
Relationship between self-esteem (X) and depression (Y)?

The relationship between two variables over a period, especially one that shows a close match between the variables' movements

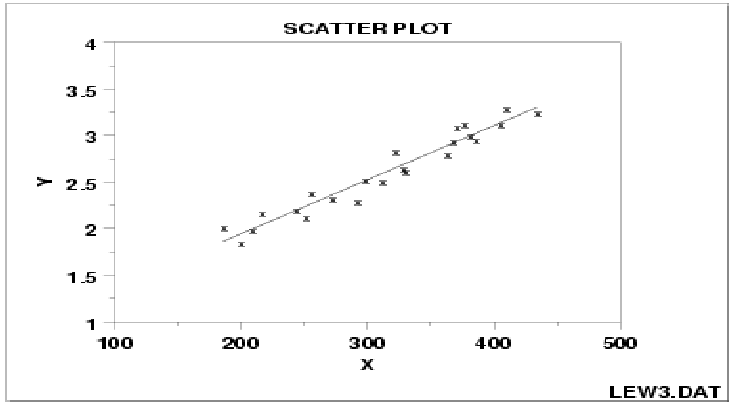
Direction and strength of relationship between two variables

Graphical representation of data in a bivariate setup

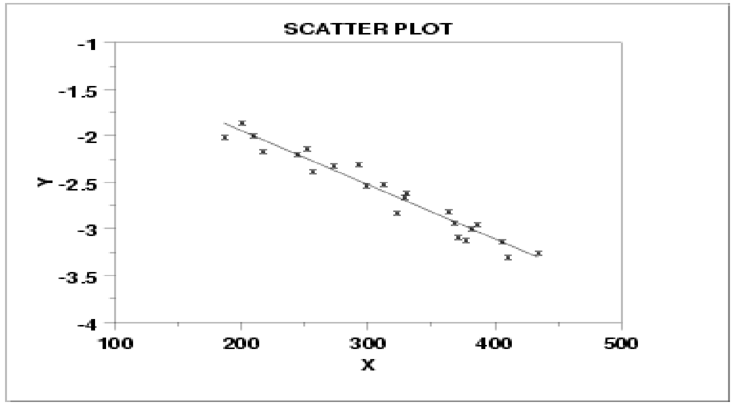
No association



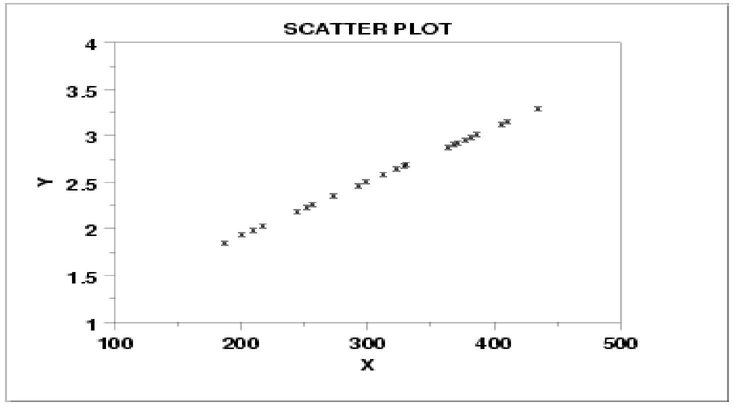
Strong linear relationship



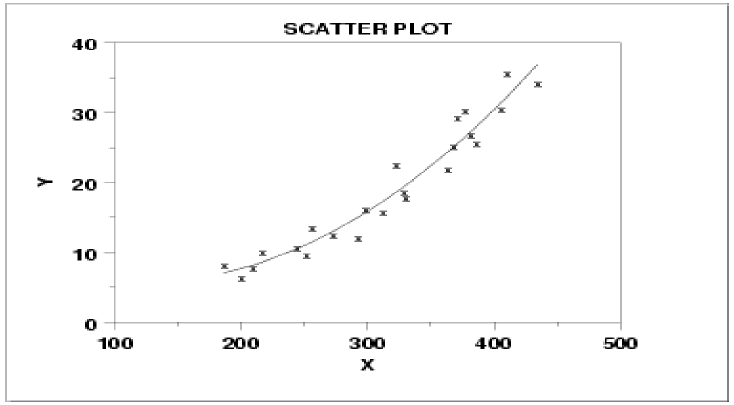
Strong linear relationship



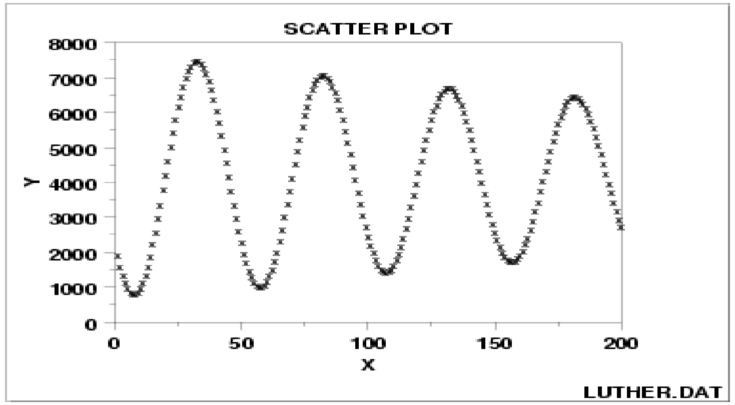
Exact linear relationship



Quadratic relationship

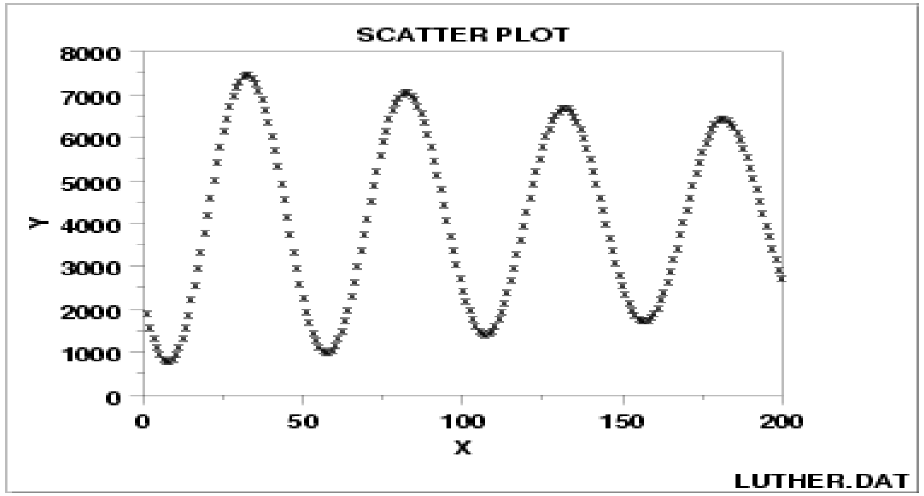


Sinusoidal relationship (damped)



Correlation measures may be misleading in certain scenarios

Correlation and independence



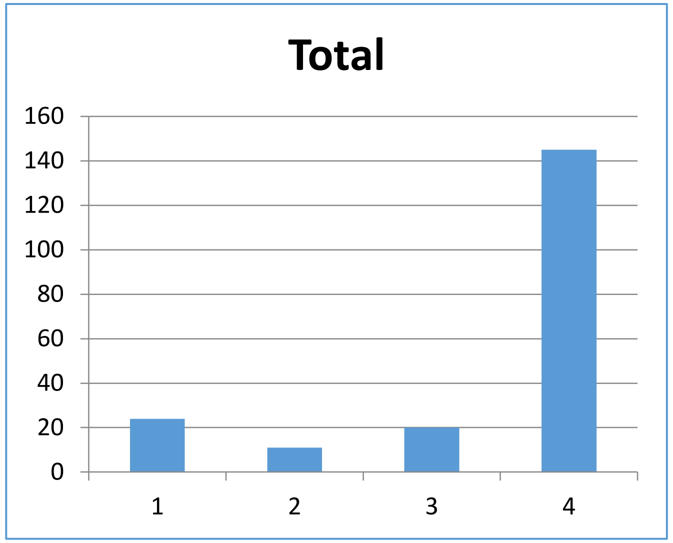
No correlations: **Does not imply no association**

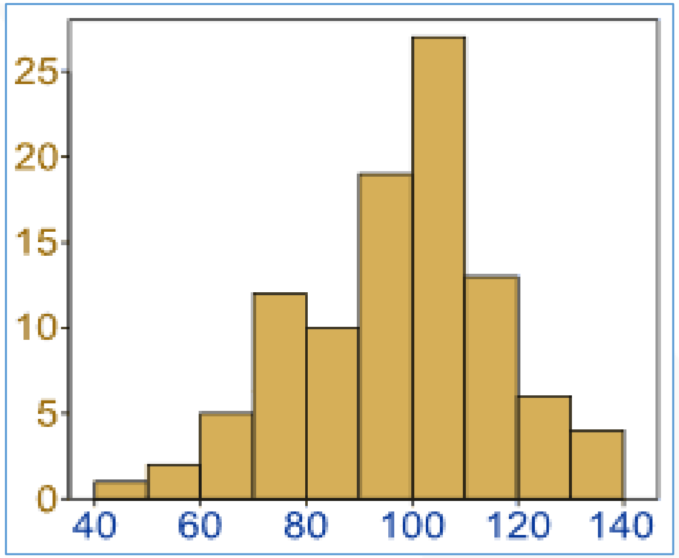
Spurious correlation

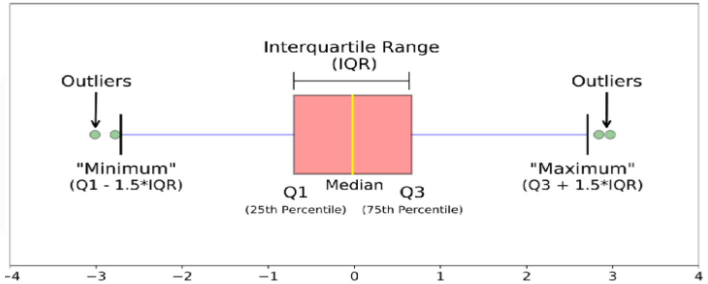
A **spurious relationship** is a mathematical relationship in which two events or variables have no direct causal connection, yet it may be wrongly inferred that they do, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "confounding factor" or "lurking variable")

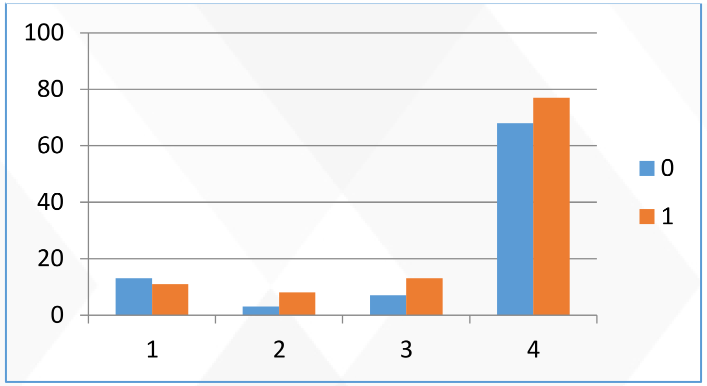
Another popular example is a series of Dutch statistics showing a positive correlation between the number of storks nesting in a series of springs and the number of human babies born at that time. Of course there was no causal connection; they were correlated with each other only because they were correlated with the weather nine months before the observations

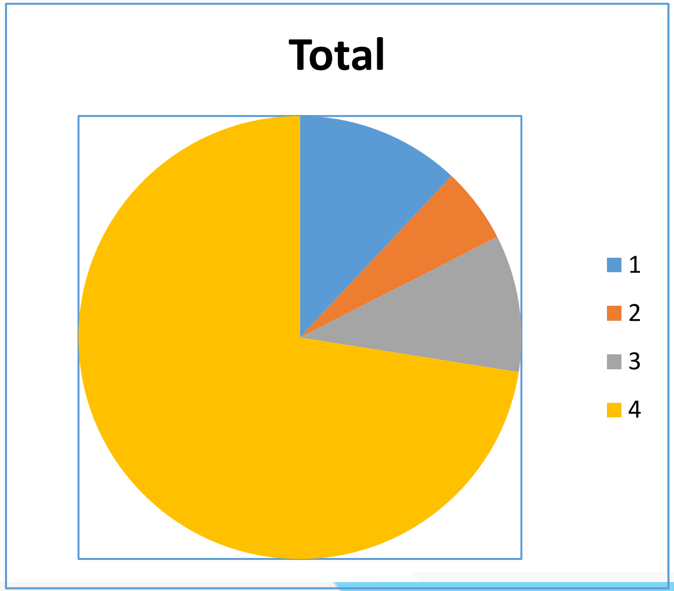
Types of Visualizations

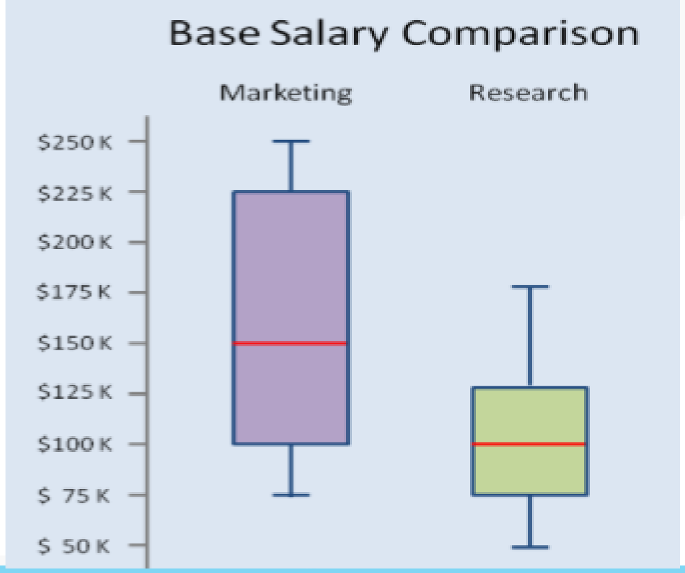


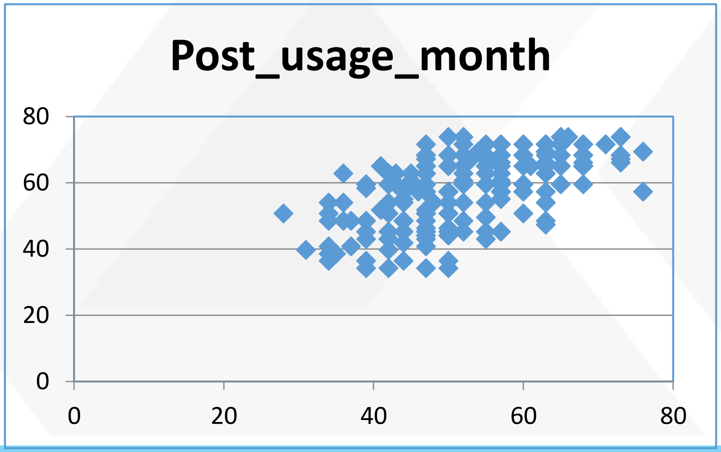












Histogram

A histogram based on relative frq density is an approximation to a probability density function

- Constructing the Histogram
  - Number of bins ~ SQRT (Observations)
- Frequency
- Relative Frequency
  - The relative frequencies are found by dividing the observed frequency in each bin by the total number of observation (even for Cumulative)
- Relative Frequency Density
  - bin height = bin frq/bin width, this is Called frequency density
- Cont. vs. Discrete
  - The important point is that f(x) is used to calculate an area that represents the probability that X assumes a value in [a, b].

Key Take-away:

- *Frq > Divide by total N > Relative frq > Divide by bin width > Relative Frq Density*
- *Relative Frq Density is the Y axis, bin width is the X part, area = X.Y; if summed over then total Area=1*
- *So, it is not based on height, but based on area!*

The shape often gives insight about possible choices of probability distribution to use as a model for the population

Chandra Mouli Kotta Kota
Analytixlabs
Mail: chandra@analytixlabs.co.in
MOB: 8826587444