Data Preparation

Big Data Analytics Certification

Agenda

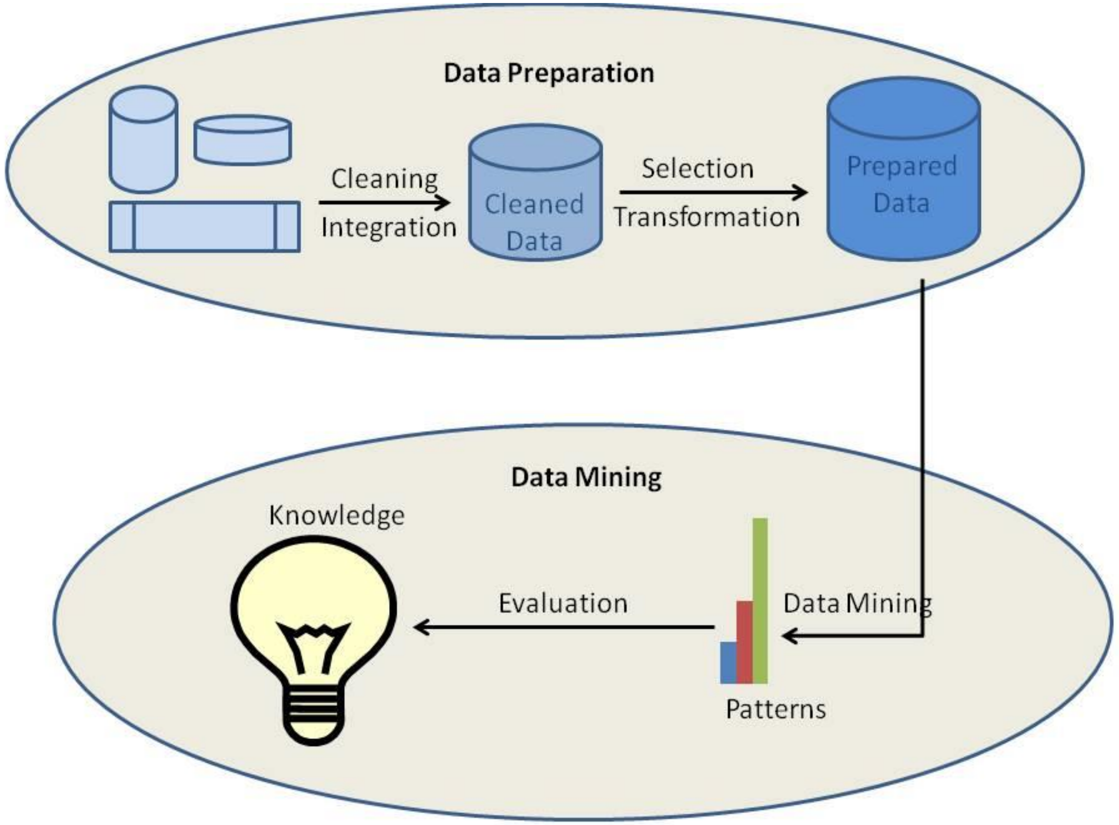
- Introduction to Data Preparation
- Types Of Data: Statistics parametric and nonparametric
- Dealing with outlier data
- Data Preparation: Probability Distribution
- Transforming the data
- What is Hypotheses
- Understand data analysis

Introduction to Data Preparation

Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis. Handling messy, inconsistent, or un-standardized data. Trying to combine data from multiple sources. Reporting on data that was entered manually.

- Once data is collected, process of analysis begins
- But, data has to be translated in an appropriate form
- This process is known as Data Preparation

Introduction to Data Preparation



Agenda

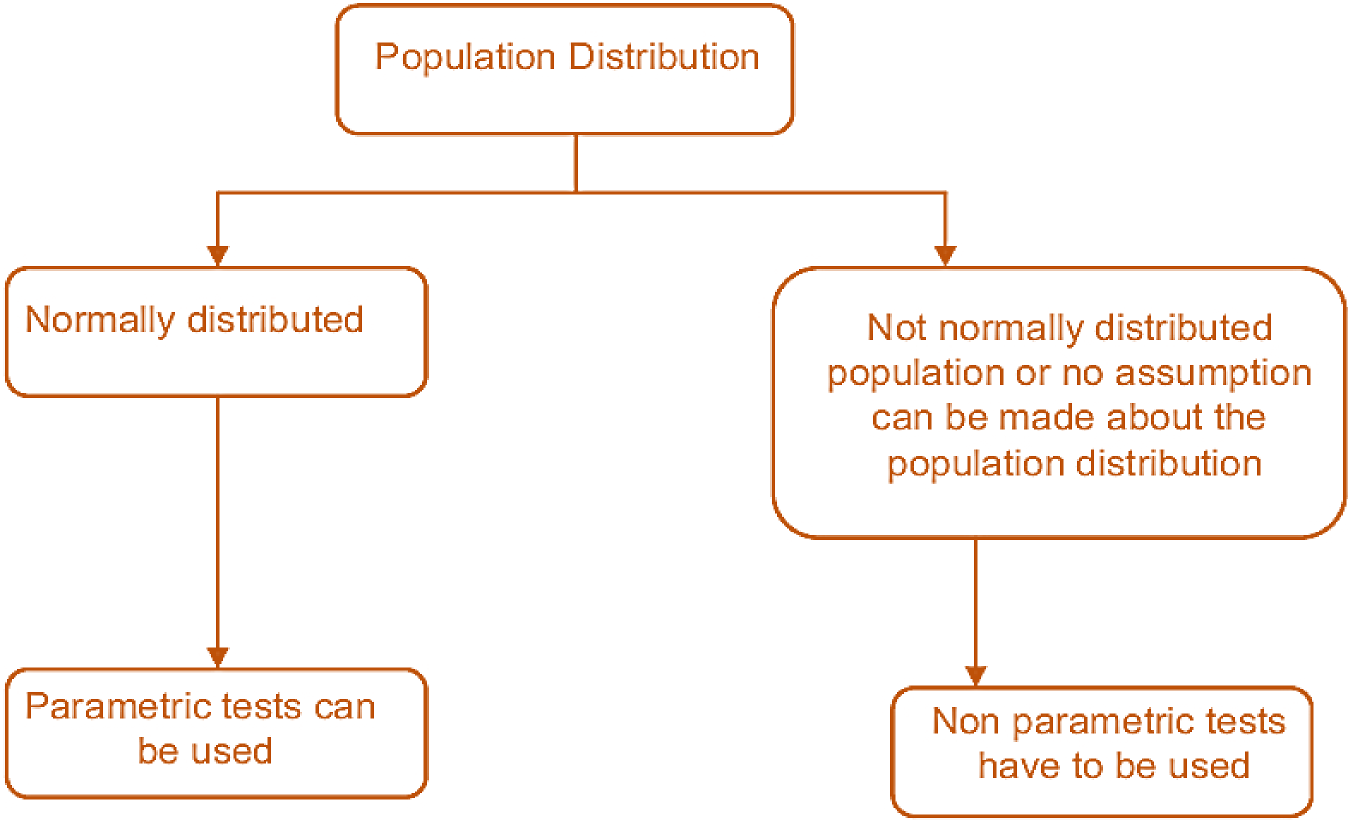
- Introduction to Data Preparation
- **Types Of Data: Statistics parametric and nonparametric**
- Dealing with outlier data
- Data Preparation: Probability Distribution
- Transforming the data
- What is Hypotheses
- Understand data analysis

Statistics Parametric and Non-Parametric Methods

- **Parametric statistics** is a branch of statistics which assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters. Most well-known elementary statistical methods are parametric.

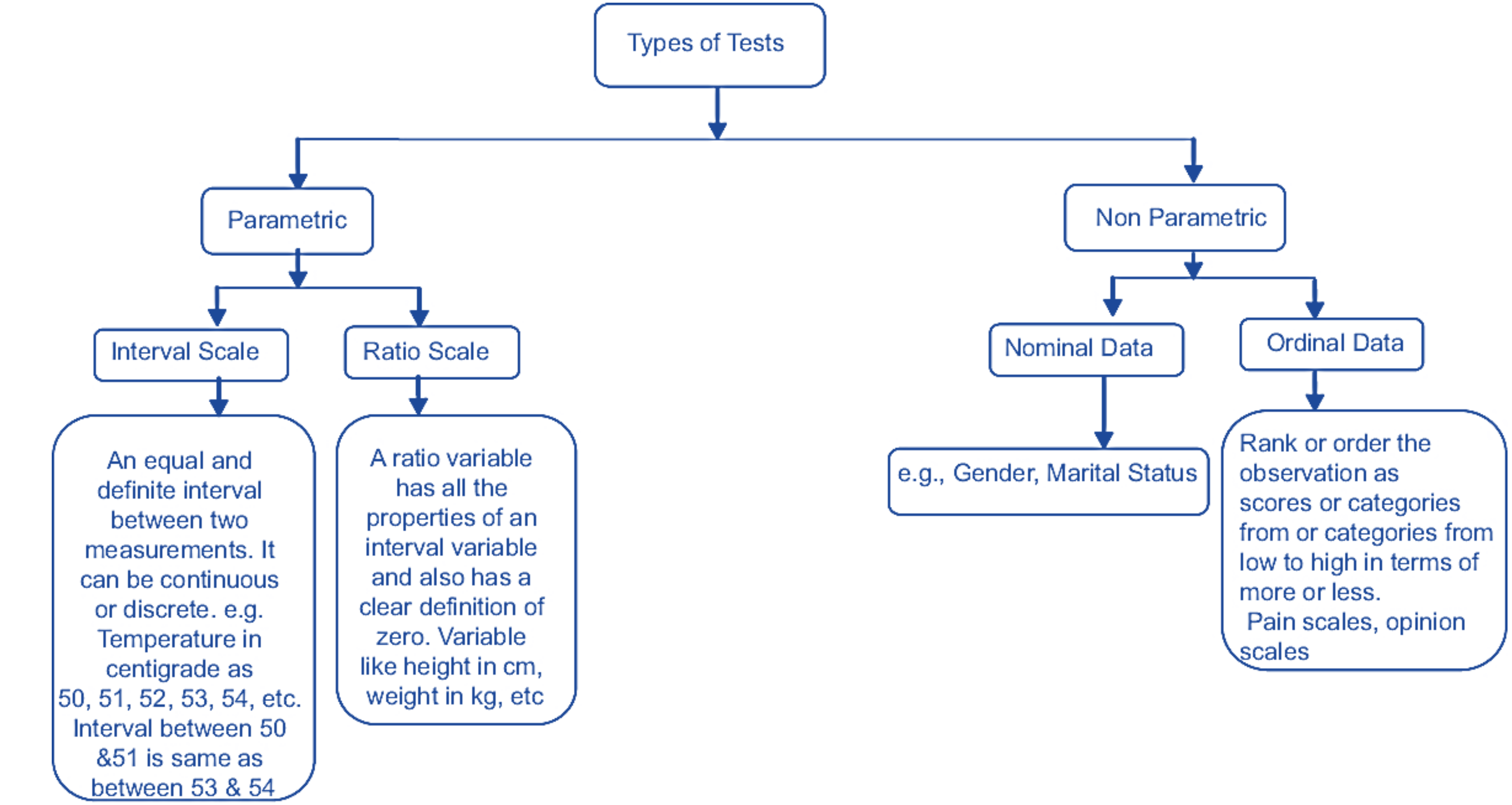
- **Non-Parametric statistics** is the branch of statistics that is not based solely on parameterized families of probability distributions (common examples of parameters are the mean and variance). Nonparametric statistics is based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified. Nonparametric statistics includes both descriptive statistics and statistical inference.

Statistics Parametric and Non-Parametric Methods



Population Distribution
Normally distributed
Not normally distributed population or no assumption can be made about the population distribution
Parametric tests can be used
Non parametric tests have to be used

Statistics Parametric and Non-Parametric Methods



Types of Tests

Parametric

Interval Scale

An equal and definite interval between two measurements. It can be continuous or discrete. e.g. Temperature in centigrade as 50, 51, 52, 53, 54, etc. Interval between 50 &51 is same as between 53 & 54

Ratio Scale

A ratio variable has all the properties of an interval variable and also has a clear definition of zero. Variable like height in cm, weight in kg, etc

Non Parametric

Nominal Data

e.g., Gender, Marital Status

Ordinal Data

Rank or order the observation as scores or categories from or categories from low to high in terms of more or less.
Pain scales, opinion scales

Statistics Parametric and Non-Parametric Methods

Parametric Assumptions:

- The observations must be independent (For example participants need to have completed the dependent variable separately, not in groups).
- The observations must be drawn from normally distributed populations
- These populations must have the same variances

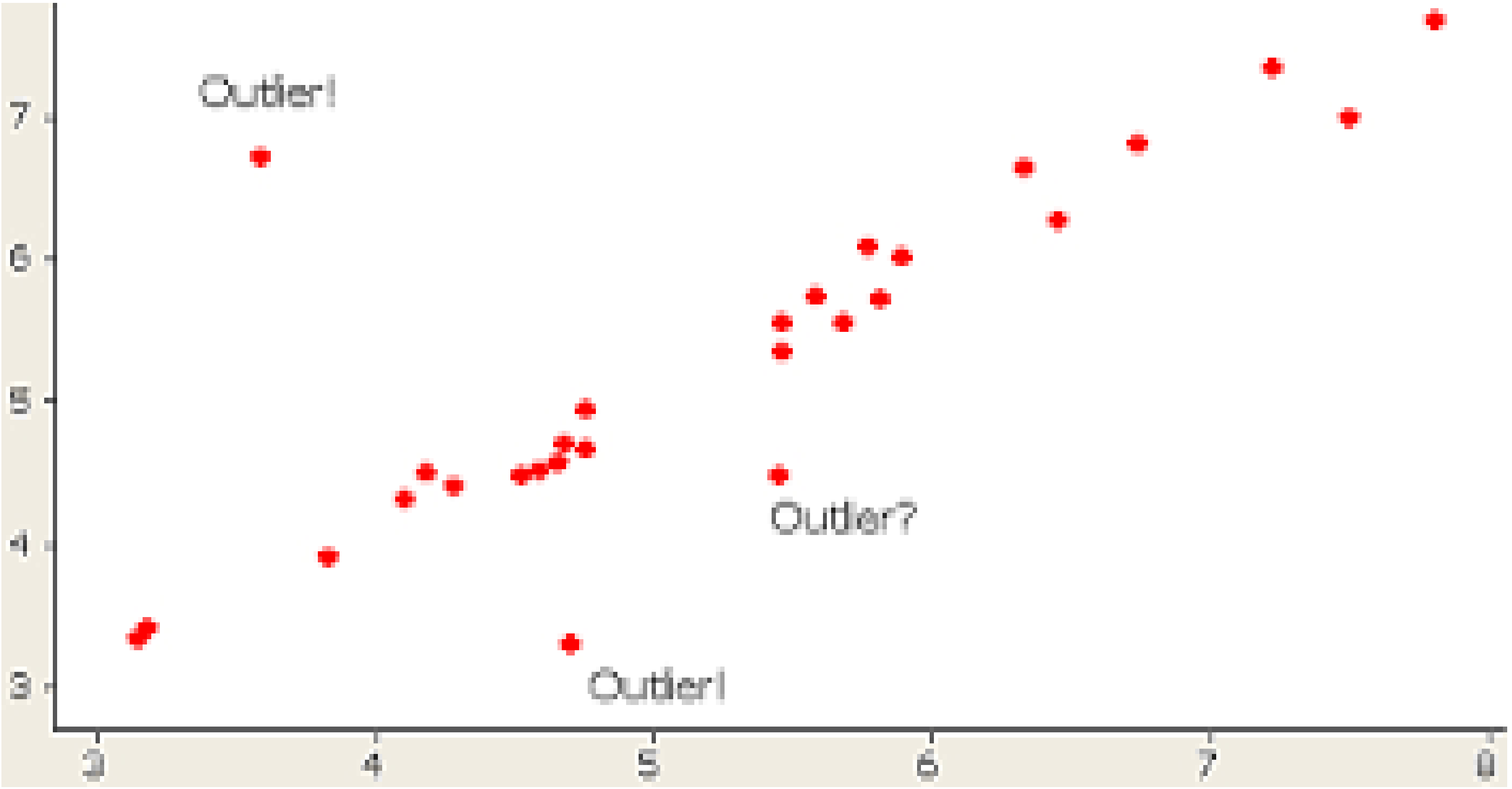
Statistics Parametric and Non-Parametric Methods

- parametric test, of course, is a test that requires a parametric assumption, such as normality. A nonparametric test does not rely on parametric assumptions like normality.
- a nonparametric test protects against some violations of assumptions and not others.
- But Many people ignore the assumptions in the data
- Many data sets have outliers
- Most of the data in this world is not normally distributed

Agenda

- Introduction to Data Preparation
- Types Of Data: Statistics parametric and nonparametric
- **Dealing with outlier data**
- Data Preparation: Probability Distribution
- Transforming the data
- What is Hypotheses
- Understand data analysis

Dealing with Outlier Data



Dealing with
Outlier Data

Here are four approaches:

- Drop the outlier records. In the case of Bill Gates, or another true outlier, sometimes it's best to completely remove that record from your dataset to keep that person or event from skewing your analysis.
- Cap your outliers data.
- Assign a new value.
- Try a transformation.

Agenda

- Introduction to Data Preparation
- Types Of Data: Statistics parametric and nonparametric
- Dealing with outlier data
- **Data Preparation: Probability Distribution**
- Transforming the data
- What is Hypotheses
- Understand data analysis

Data Preparation:
Probability Distribution

Discrete Probability Distribution

- is a distribution of probability for random variables whose values are obtained by counting (counting),
- Example :
  - Bernoulli
  - Binomial
  - Poisson

Continuous Probability Distribution

- is a distribution of probability for random variables whose values are obtained using a measuring instrument.
- Example :
  - Normal
  - Weibull
  - Gamma
  - Betha

Data Preparation

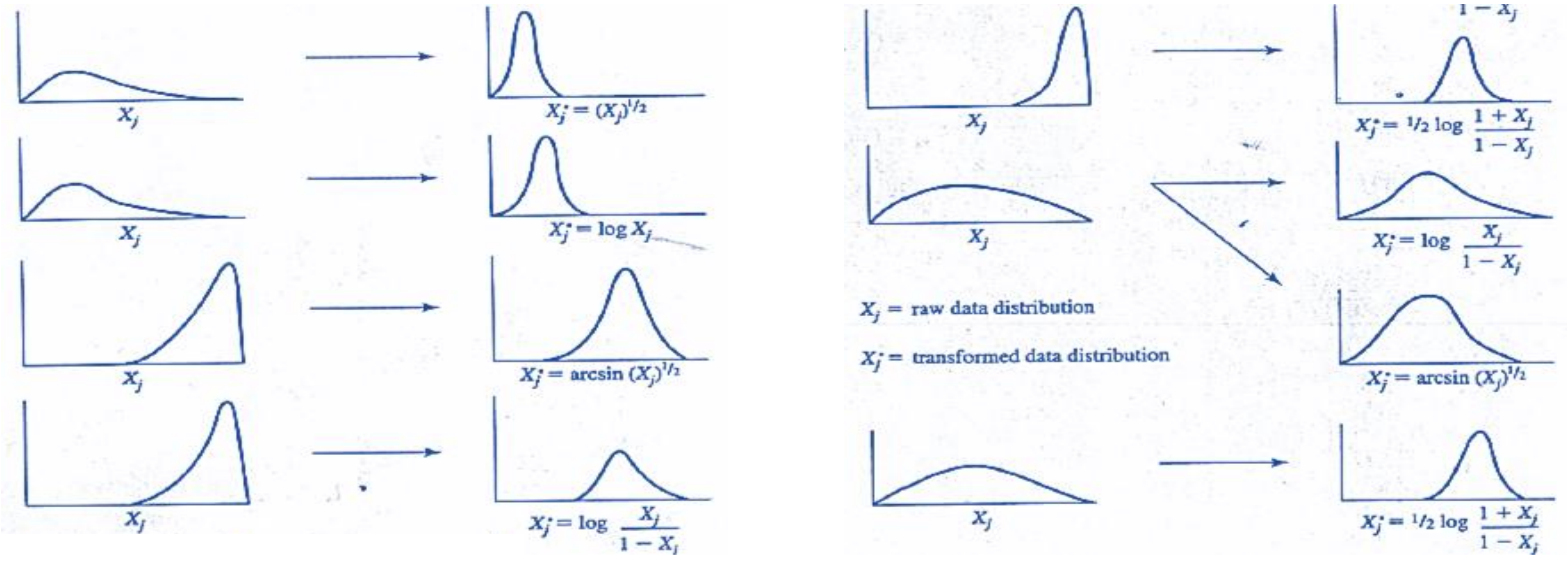
Agenda

- Introduction to Data Preparation
- Types Of Data: Statistics parametric and nonparametric
- Dealing with outlier data
- Data Preparation: Probability Distribution
- **Transforming the data**
- What is Hypotheses
- Understand data analysis

Transforming the Data

- In statistics, data transformation is the application of a deterministic mathematical function to each point in a data set — that is, each data point zi is replaced with the transformed value yi = f(zi), where f is a function.

Transforming the Data



Agenda

- Introduction to Data Preparation
- Types Of Data: Statistics parametric and nonparametric
- Dealing with outlier data
- Data Preparation: Probability Distribution
- Transforming the data
- **What is Hypotheses**
- Understand data analysis

The Null Hypothesis, H_0 and The Alternative Hypothesis, H_1

Begin with the assumption that the null hypothesis is true. Similar to the notion of innocent until proven guilty.

- Refers to the status quo
- Always contains “=” , “≤” or “⍰” sign
- May or may not be rejected

Is the opposite of the null hypothesis, e.g., The average number of TV sets in U.S. homes is not equal to 3 (H1: $\mu \neq 3$).

- Challenges the status quo
- Never contains the “=” , “≤” or “⍰” sign
- May or may not be proven
- Is generally the hypothesis that the researcher is trying to prove

Outcomes and Probabilities

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No error (1 - a)	Type II Error (β)
Reject H_0	Type I Error (a)	No Error (1 - β)

Key:
Outcome
(Probability)

Agenda

- Introduction to Data Preparation
- Types Of Data: Statistics parametric and nonparametric
- Dealing with outlier data
- Data Preparation: Probability Distribution
- Transforming the data
- What is Hypotheses
- **Understand data analysis**

Understand
Data Analysis

- Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

- Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

- In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation

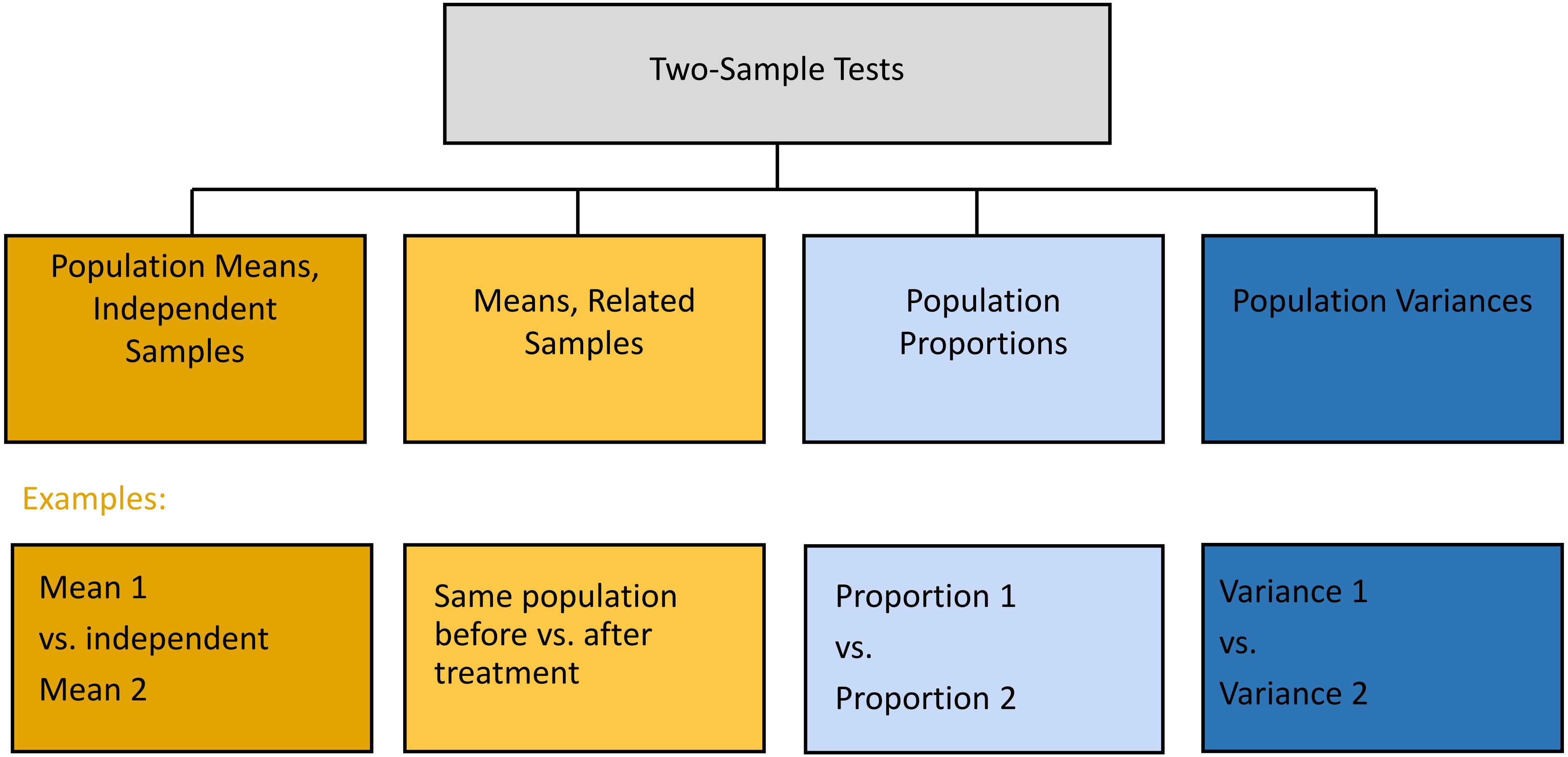
Understand Data Analysis:
Learning Objectives

Learn hypothesis testing procedures to test:

- The means of two independent populations
- The means of two related populations
- The proportions of two independent populations
- The variances of two independent populations
- The means of more than two populations

Understand Data Analysis:

Two – Sample Tests



Two-Sample Tests

Population Means, Independent Samples

Means, Related Samples

Population Proportions

Population Variances

Examples:

Mean 1
vs. independent
Mean 2

Same population before vs. after treatment

Proportion 1
vs.
Proportion 2

Variance 1
vs.
Variance 2

Understand Data Analysis:
Differences Between Independent Groups

Two samples –

compare mean value for some

variable of interest

Parametric	Nonparametric
t-test for independent samples	Wald-Wolfowitz runs test
	Mann-Whitney U test
	Kolmogorov-Smirnov two sample test

Understand Data Analysis:
Differences Between Independent Groups

	Parametric	Nonparametric
Multiple groups	Analysis of variance (ANOVA/ MANOVA)	Kruskal-Wallis analysis of ranks
		Median Test

Understand Data Analysis:
Differences Between Dependent Groups

	Parametric	Nonparametric
Compare two variables measured in the same sample	t-test for dependent samples	Sign test
		Wilcoxon’s matched pairs test
If more than two variables are measured in same sample	Repeated measures ANOVA	Friedman’s two way analysis of variance
		Cochran Q

Understand Data Analysis:
Relationships Between Variables

Two variables of interest are categorical

Parametric	Nonparametric
Correlation Coefficient	Spearman R
	Kendall Tau
	Coefficient Gamma
	Chi square
	Phi coefficient
	Fisher exact test
	Kendall coefficient of concordance

Understand Data Analysis:

Summary Table of Statistical Tests

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	X2 or binomial	X2	Macnarmar's X2	X2	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r

(Plonskey, 2001)