Non-Parametric Data and Statistical Tests

Dr. Subrata Roy

Associate Professor, Department of Commerce

Mahatma Gandhi Central University

Motihari, Bihar - 845401

The Research Procedure

- Research Question
- Use Theory
- Make Hypothesis
- Identify Data
- Identify Variables
- Collect Data
- Measure Variables
- Graph Data and Fit a Model
- Evaluating the Hypotheses
- Analyze Data and Results

Research Questions

Research Questions

- A **research question** is an answerable inquiry into a specific issue.
- It focuses the study, determines the methodology and guides all stages of inquiry, analysis and reporting.
- **How?**

  1. Select a research topic

  2. Convert the research topic into a research question. Remember that a research question is a statement of what you want to study.
- **Characteristics of a good research question**: The question will be feasible, clear, significant and ethical.

Research Questions

Question type	Question	Examples
Exploratory questions	What is the case? What are the key factors?	What are the critical success factors of a profitable company? What are the distinguishing features of a good leader? What are the reasons for the massacre on Burdwan roads?
Descriptive questions	How many? What is the incidence of x? Are x and y related?	How many people died of Dengue in Burdwan last year? Is there a correlation between parental support and academic achievement?

Research Questions

Question type	Question	Examples
Causal questions	Why? What are the causes of y?	What are the main causes of underfeeding in a rural community? Is smoking the main cause of lung cancer?
Evaluative questions	What was the outcome of X? Has X been successful?	Has the new Dengue awareness programme produced a decline in reportable Dengue cases? Has the introduction of a new refrigeration technology led to more cost-effective production?

Research Questions

Question type	Question	Examples
Predictive questions	What will the effect of x be on y?	What effect will the introduction of a new educational bill have on college teachers?
Historical questions	What led to y happening? What were the events that led up to y? What caused y?	What caused the end of old Indian currency in India during the year? What led US President to decide to start aerial bombing of Iran?

Hypothesis

Hypothesis

- One of the most difficult aspects of statistics is determining which procedure to use in what situation.
- A statistical hypothesis is an unproven statement which can be tested. A hypothesis test is used to test whether this statement is true.
- A hypothesis test examines a claim about some characteristic of a population.
- A hypothesis is a prediction regarding the possible outcome of a study

Hypothesis

- A simple hypothesis involves one independent and one dependent variables.
- A complex hypothesis involves multiple variables, such as two or more dependent and independent variables.
- A directional hypothesis is based on an established theory, while a non-directional hypothesis is used when there is insufficient scientific basis for the prediction.
- A directional hypothesis is one in which the researcher indicates the specific direction that he or she expects will emerge in a relationship in the study.
- A non-directional hypothesis is when there is no specific prediction about what direction the outcome of a study will take.
- A hypothesis H is said to be a simple hypothesis if H completely specifies the density $f(x; \theta)$ of the population; otherwise it is called a **composite** hypothesis.

Hypothesis

- An <u>associative hypothesis</u> supposes that when one variable changes, the other variable also changes.
- A <u>causal hypothesis</u> assumes the cause and effect relationship between variables.
- **Empirical hypothesis** – The formulation is based on assumption only but actually it is applied in the field or tested through statistical tools.
- **Logical hypothesis** – It is that type in which hypothesis is verified logically.

Hypothesis

- **Statistical hypothesis** - *A* statistical hypothesis is either (1) a statement about the value of a population parameter (mean, median, mode, variance, standard deviation, proportion, total) or (2) a statement about the kind of probability distribution that a certain variable obeys.
- **Null hypothesis** - If the original claim includes equality (<=, =, or >=), it is the null hypothesis. If the original claim does not include equality (<, not equal, >) then the null hypothesis is the complement of the original claim. The null hypothesis *always* includes the equal sign. The decision is based on the null hypothesis.

Hypothesis

- **Alternative hypothesis** - The hypothesis the researcher wants to test is called the alternative hypothesis H_1.
- The objective is to disprove the null hypothesis.
- The significance level is the critical probability of choosing between the null hypothesis and the alternative hypothesis

Testing of Hypothesis

1. State the research question.
2. State the statistical hypothesis.
3. Choose a level of significance
4. Select the appropriate statistics
5. Set decision rule.
6. Calculate the test statistic.
7. Decide if result is significant.
8. Interpret result as it relates to your research question.

1. State the research question

- The research question will be of academic and intellectual interest within your discipline.
- It is not a topic, fragment, phrase or sentence. It ends with a question mark!
- Clear and precisely stated. It is not too broad, nor is it too narrow.
- Open-ended as opposed to closed. It cannot be answered in a sentence or phrase.

State the research question

- List your interests as they pertain to your specific discipline or assignment.
- Now find the answers way of: Who? What? When ? Where? Why? How? So what?" and "What if…?
- These represent possible "gaps" in your knowledge; the last four are particularly tough because they are open-ended – they often lead to good research questions .

2. State the statistical hypothesis

- **Focusing Question** - The most important question you discovered from the three prior activities.

  -How can MGCUB increase its economic impact on the society?

- **Supporting Questions** - Questions that will help you explore the relationships around the focusing question in greater depth.

  -How does MGCUB currently impact the society economically? What could be done that isn't being done and why? What are the limitations?

State the statistical hypothesis

- Make out the hypothesis from the problem
- Convert the hypothesis to math
- State what will happen if the hypothesis doesn't come true
- State what will happen if the experiment doesn't make any difference
- Figure out the alternate hypothesis

Null hypothesis represents status quo.

Alternative hypothesis represents the desired result.

3. Choose a level of significance

- Whenever we draw inferences about a population, there is a risk that an incorrect conclusion will be reached
- The level of significance affirms the probability of incorrectly rejecting H_0. This error is commonly known as Type I error and we denote the significance level as α.
- Level of Significance selected is typically 0.05 or 0.01

How to Choose Level of Significance

- Entrust Type error II when we incorrectly accept a null hypothesis when it is false. The probability of committing Type error II is denoted by β.

Type I and Type II Error

	Accept null	Reject null
Null is true	Correct Decision – No Error	Type I error
Null is False	Type II error	Correct Decision – No Error

How to Choose Level of Significance

Which is good?

- Both are serious, but traditionally Type I error has been considered more serious, that's why the objective of hypothesis testing is to reject H_0 only when there is enough evidence that supports it.
- Therefore, we choose α to be as small as possible without compromising β.
- Increasing the sample size for a given α will decrease β

4. How to Select appropriate test

- The selection of a proper Test depends on
  - Scale of the data (nominal/ordinal/interval/ratio)
  - the statistic you seek to compare (means/proportions)
  - the sampling distribution of such statistic (Normal Distribution/T Distribution/χ^2 Distribution
  - Number of variables (Univariate/Bivariate/Multivariate)
  - Type of question to be answered

5. How to Set Decision Rule

- The decision rule is a statement that tells under what circumstances to reject the null hypothesis. The decision rule is based on specific values of the test statistic (say, reject H_0 if $Z \geq 1.645$).
- The decision rule for a specific test depends on three factors:

  -the research or alternative hypothesis,

  -the test statistic and

  -the level of significance.

How to Set Decision Rule

- The decision rule depends on whether an right-tailed, left-tailed, or two-tailed test is proposed.
- In a right-tailed test, the decision rule has researchers reject H_0 if the test statistic is larger than the critical value.
- In a left-tailed test the decision rule has researchers reject H_0 if the test statistic is smaller than the critical value.
- In a two-tailed test, the decision rule has investigators reject H_0 if the test statistic is extreme, either larger than an upper critical value or smaller than a lower critical value.

How to Set Decision Rule

- The exact form of the test statistic is also important in determining the decision rule.
- If the test statistic follows the standard normal distribution (Z), then the decision rule will be based on the standard normal distribution.
- If the test statistic follows the t distribution, then the decision rule will be based on the t distribution.
- The appropriate critical value will be selected from the t distribution again depending on the specific alternative hypothesis and the level of significance.
- The level of significance (α =0.05) dictates the critical value. For example, in an right tailed Z test, if α =0.05 then the critical value is Z=1.645.

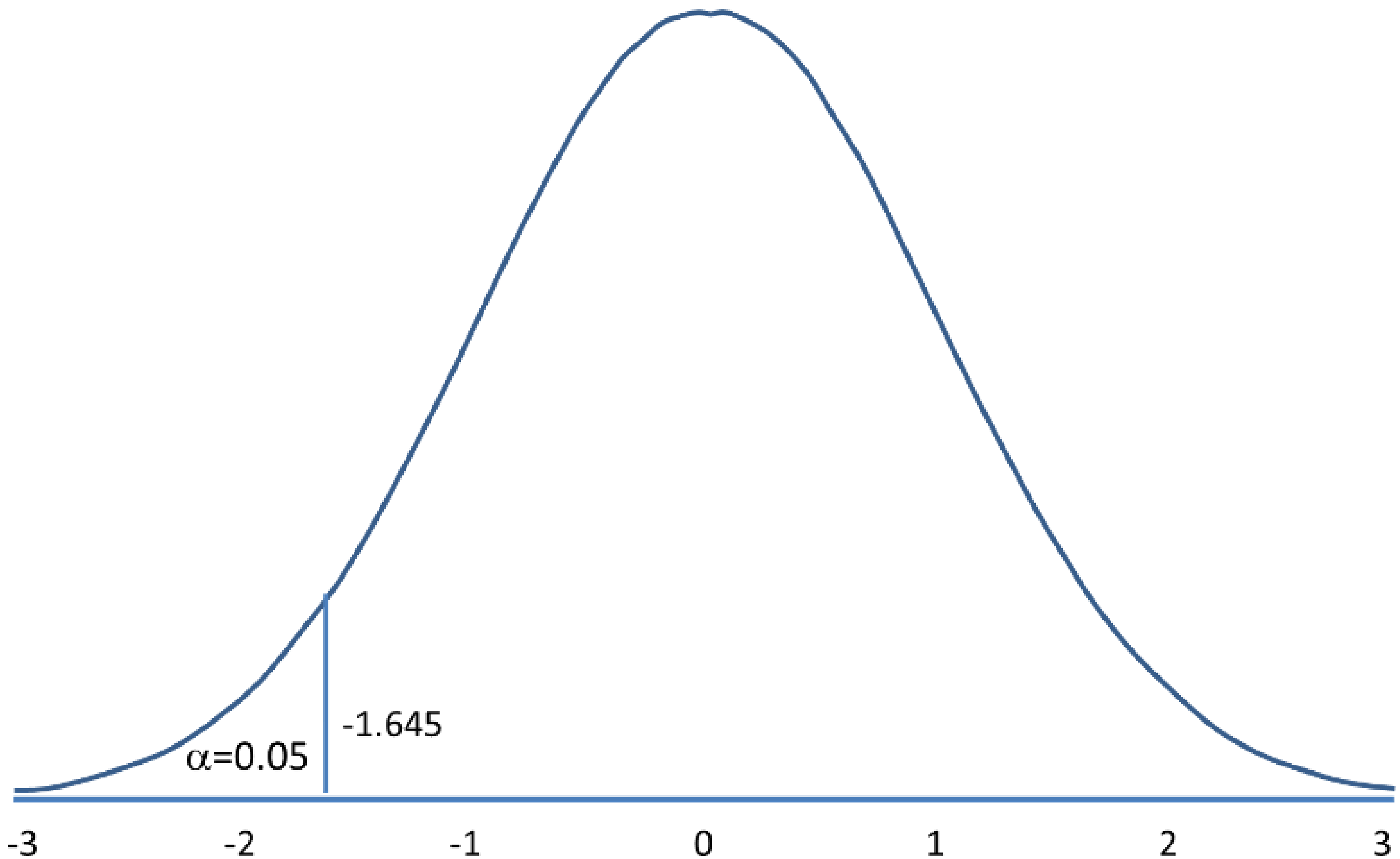
How to decide, is it one tailed test or Two tailed test in Hypothesis Testing

- Question – 1: The Department of Commerce, MGCUB claims that the dropout rate is 5%. Last year, 4 out of 80 students dropped out. Is there enough evidence to reject the Department's claim?

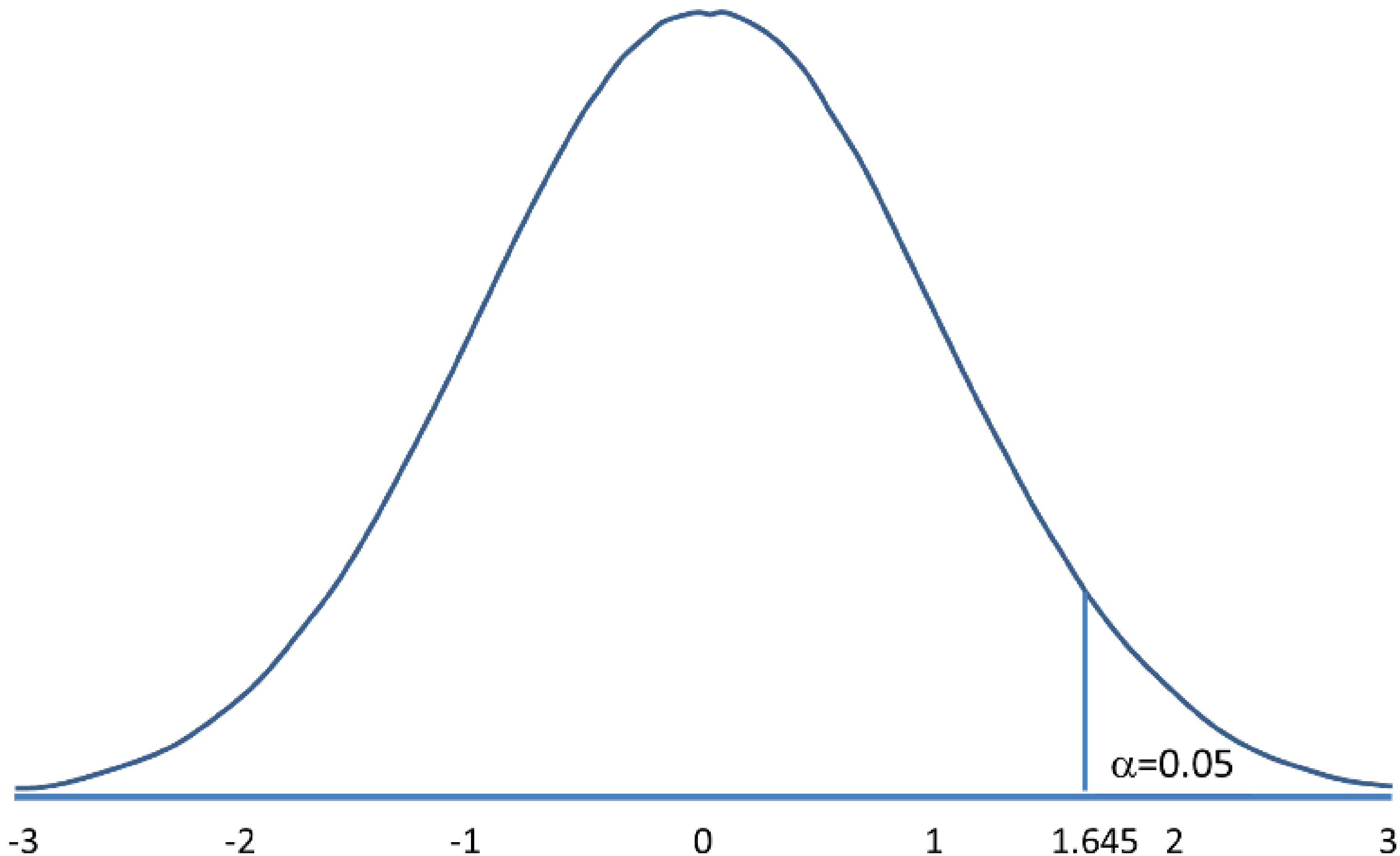
- Question – 2: The Department of Commerce, MGCUB claims that the dropout rate is less than 5%. Last year, 4 out of 80 students dropped out. Is there enough evidence to reject the Department's claim?

- Question – 3: The Department of Commerce, MGCUB claims that the dropout rate is more than 5%. Last year, 4 out of 80 students dropped out. Is there enough evidence to reject the Department's claim?

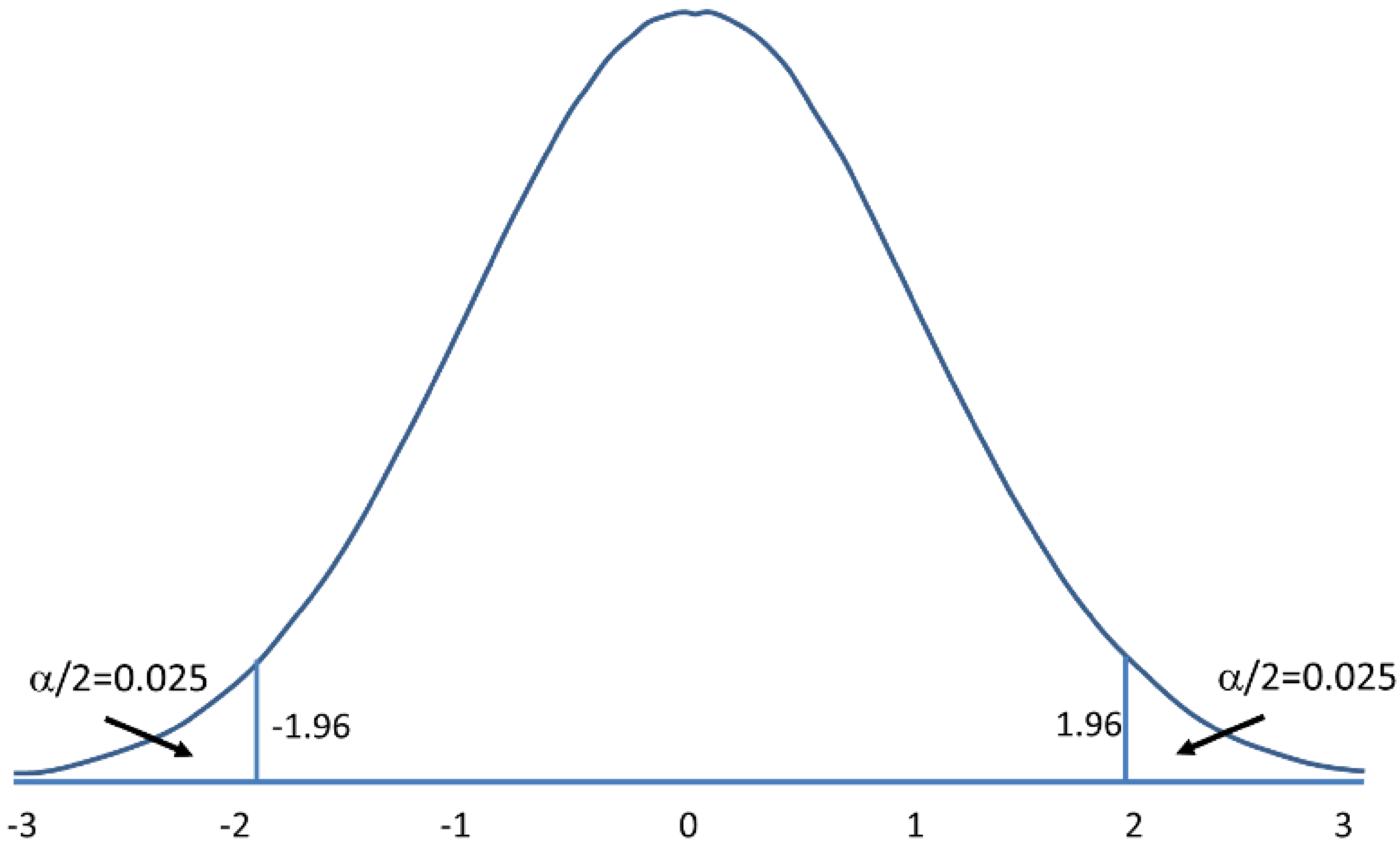
Left-Tailed Test



Right-Tailed Test



Two-Tailed Test



Guidelines for using the *P*-value to assess the evidence against the null hypothesis

P-value	Evidence against H_0
$P > 0.10$	Weak or none
$0.05 < P \leq 0.10$	Moderate
$0.01 < P \leq 0.05$	Strong
$P \leq 0.01$	Very strong

Conclusions for a Hypothesis Test

- If the null hypothesis is rejected, we conclude that the alternative hypothesis is true.
- If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis.

Data

Data

- **Data** are the facts and figures collected, summarized, analyzed and interpreted.
- The data collected in a particular study are referred to as the **data set**.
- The **elements** are the entities on which data are collected.
- A **variable** is a characteristic of the elements.
- The set of measurements collected for a particular element is called an **observation**.

Attribute

- An **attribute** refers to the quality of a characteristic.
- **Attribute** is a characteristic of an object (person, thing, etc.).
- **Attributes** are closely related to variables. A variable is a logical set of **attributes**.
- The theory of attributes deals with qualitative types of characteristics that are calculated by using quantitative measurements.
- Attributes refer to the characteristics of the item under study, like the habit of smoking or drinking.
- So ‘smoking’ and ‘drinking’ both refer to the example of an attribute.

Types of Data

- Primary data
- Secondary data
- Qualitative data
- Quantitative data

Types of Data

- **Discrete data** – it can't be broken down into smaller data values, e.g. a questionnaire with answer options of "Yes/No?" and "Male/Female?". It is counted.
- **Continuous data** – continuous data are values within a bounded or boundless interval. It is measured.
- **Ordinal data** – refers to quantities that have a natural ordering. This is used to describe data that has a sense of order, but for which we cannot be sure that the distances between the consecutive values are equal.
- **Binary data/Dichotomous data** - It is a type of data that is represented or displayed in the binary numeral system. Binary data is the only category of data that can be directly understood and executed by a computer. It is numerically represented by a combination of 0 and 1.

Types of Data

- **Polychotomous data** - A polychotomous data is a data that can have more than two values.

  Polychotomous data can be ordered, unordered, or sequential:

  Ordered polychotomous data: Data that have some kind of order, like: "1" if you earn up to Rs. 25,000, "2" if you earn Rs. 25,001-Rs.50,000 and "3" if you earn over Rs. 50,000.

  Unordered polychotomous data: Data that don't have an implied order, like: "1" for male, "2" for female "3" for trans gendered male and "4" for trans gendered female.

  Sequential polychotomous data: Data with a sequence. For example: "1" for freshmen, "2" for sophomore, "3" for junior and "4" for senior.

Types of Data

- **Nominal data** – Nominal is hardly measurement. It refers to quality more than quantity. A nominal level of measurement is simply a matter of distinguishing by name, e.g., 1 = male, 2 = female.
- **Interval data** – Interval data is like ordinal except we can say the intervals between each value are equally split.
- **Ratio data** – Ratio data is interval data with a natural zero point.

Non-parametric data/statistics

- Non-parametric data/statistics are used when the parameters of the population are not measurable or do not meet certain standards.
- In cases when the data only order the observations, so that the interval between the observations is unknown, neither a mean nor a variance can be meaningfully computed. In such cases, you need to use non-parametric tests.
- When either (1) the population is not normal and the samples are small or (2) when the data are not cardinal, the same non-parametric statistics are used.

When to Use a Nonparametric Test

- When it is clear that the outcome does not follow a normal distribution. These include situations:

  -When the outcome is an ordinal variable or a rank,

  -When there are definite outliers or

  [An outlier is an observation that lies an abnormal distance from other values in a random sample from a population]

  -When the sample size is small.

Non-parametric data/statistics

- Describe some attribute of a population
  – Use univariate stats
  - Quantitative variables (median & Inter-quartile Range)
  -Qualitative variables (Mode for nominal data)
- Test hypotheses about that attribute
  – Use univariate statistical tests
  - Quantitative variables (1 Sample Median Test)
  -Qualitative variables (Chi-square Test)
- Its relationship with some other attribute
  – Use tests of association
  - Quantitative variables (Correlation, t-test)
  -Qualitative variables (Pearson's Contigency Test)

Non-parametric data/statistics

- Differences on that attribute across populations
  - Use between groups comparisons
  - Quantitative variables (Sign Test, Mann-WhitneyTest)
  - Qualitative variables (Chi-square Test)
- Across time or across related constructs
  - Use within-groups comparisons
  - Quantitative variables (Kruskal-Wallis Test)
  - Qualitative variables (Chi-square Test)

Normal Distribution

- We will use the following to determine if a distribution is approximately normal:

1. Q-Q Plot values should lie close to the 45 line.
2. Distribution should be similar in shape to the normal curve.
3. Skew & Kurtosis should be reasonably close to 0.
4. Data points with a Z score > +3.3 or < -3.3 will be considered as outliers and removed.

Test of Normality-Kolmogorov-Sminov Test

- The Kolmogorov-Smirnov test is a nonparametric procedure that determines whether a sample of data comes from a normal distribution.
- It is mostly used for evaluating the assumption of univariate normality by taking the observed cumulative distribution of scores and comparing them to the theoretical cumulative distribution for a normally distributed variable.
- We will test the following null hypothesis:

  H_o: There is no difference between the distribution of sense of data and a normal distribution.

- If test results are significant, then reject the null hypothesis. Consequently, it is concluded that the data are not normally distributed.

Kolmogorov-Sminov Test

- Formula: D = Maximum $|Fo(X) - Fr(X)|$
- Where Fo(X) = Observed cumulative frequency distribution of a random sample of n observations.

  and Fo(X)= k/n (No. of observations ≤ X)/(Total no. of observations).
- Fr(X) = The theoretical frequency distribution.
- **Acceptance Criteria:** If calculated value is less than critical value accept null hypothesis.
- **Rejection Criteria:** If calculated value is greater than table value reject null hypothesis.

Kolmogorov-Sminov Test

	Observed (O)	Theoretical (T)	Fo(X)	Fr(X)	|Fo(X) - Fr(X)|
B.Sc.	5	12	5/60	12/60	7/60
B.A.	9	12	9/60	24/60	10/60
B.Com.	11	12	11/60	36/60	11/60
M.A.	16	12	16/60	48/60	7/60
M.Com.	19	12	19/60	60/60	60/60
Total	N=60				

Test Statistics: D = Maximum $|Fo(X) - Fr(X)|$

$$D = 11/60 = 0.183$$

The table value of D at 5% significance level is given by

$D_{0.05} = 1.36/\sqrt{n} = 1.36/\sqrt{60} = 0.175$

Since the calculated value is greater than the critical value, hence we reject the null hypothesis and conclude that there is a difference among students of different streams

Variables

Variables

- In research, **attribute** is a characteristic of an object (person, thing, etc.). *A* **variable** is a logical set of attributes.
- *A* variable is a variation within a class of objects.
- *A* constant is a characteristic with no variations within a class of objects.
- Variables are classified as either <u>quantitative</u> or <u>categorical</u>
- *A* **quantitative variable** is conceptualized and analyzed in distinct categories, with no continuum implied

Types of Variables

- **A categorical variable/Nominal variable** does not vary in degree, amount or quantity but are qualitatively different.
- **Ordinal variable** used to rank a sample of individuals with respect to some characteristics, but differences (i.e., intervals) and different points of the scale are not necessarily equivalent. Examples: anxiety might be rated on a scale "none," "mild," "moderate," and "severe," with numerical values of 0, 1, 2, 3.
- **The independent variable/Treatment variable/Manipulated variable/Selected Variable** is what the researcher studies to see its relationship or effects.
- The **dependent variable** is what is being influenced or affected by the independent variable.

Types of Variables

- **Control variable/Extraneous variable/Covariate** - When researcher does not intend to examine the relationship but he knows that some of the factor may influence the dependent variable.
- **Exogenous variable** - A variable entering from and determined from outside of the system being studied. A causal system says nothing about its exogenous variables.
- **Endogenous variable** - A variable that is an inherent part of the system being studied and that is determined from within the system. A variable that is caused by other variables in a causal system.
- **Outcome Variable/Criterion Variable -** The presumed effect in a non experimental study.

Types of Variables

- **Intervening variable/mediating variable/ intermediary variable** - A variable that explains a relation or provides a causal link between other variables. Example, Stability indicators mediates the relation between WCM and Profitability.
- **Moderator Variable**
  - Special type of independent variable
  - Selected to determine if it affects (modifies) the basic relationship between the primary independent variable and the dependent variable

Types of Variables

- **Binary variable/Dichotomous variable -** Obsevations (i.e., dependent variables) that occur in one of two possible states, often labelled 0 and 1
- **Polychotomous variables** - Variables that can have more than two possible values. Strictly speaking, this includes all but binary variables. The usual reference is to categorical variables with more than two categories.
- **Latent variable** - An underlying variable that cannot be observed. It is hypothesized to exist in order to explain other variables, such as specific behaviours, that can be observed.

  (say, conservatism or liberalism is a latent variable)
- **Predictor variable** - The presumed "cause" on a non experimental study. Often used in correlational studies.

Types of Variables

- **Dummy Variables** - Created by recoding categorical variables that have more than two categories into a series of binary variables.

  Example, Marital status, if originally labelled 1=married, 2=single, and 3=divorced, widowed, or separated, could be redefined in terms of two variables as follows: var_1: 1=single, 0=otherwise. Var_2: 1=divorced, widowed, or separated, 0=otherwise.

Types of Variables

- Nominal variables are mutually exclusive
- Ordinal are mutually exclusive and ordered
- Interval are mutually exclusive, ordered and have equal increments
- Ratio are mutually exclusive, ordered, equal increments and are proportional.

Collection of Data

Research Guide

- Sources of information are often considered primary, secondary or tertiary depending on their originality and proximity (closeness) of when it was created.
- Primary, Secondary, Tertiary Sources

Primary Sources of Data

- **Primary sources** are first-hand accounts or individual representations and creative works. Primary sources may be published or unpublished works.

- **General examples**: Letters, diaries, speeches, interviews, correspondence
- **History**: Transcript of speech given by Prime Minister; newsreel footage of Kargil War.
- **Literature**: Fiction or novel, story, poetry.
- **Art**: Works by artists such as Pablo Picasso's painting.
- **Social Sciences**: Interview transcripts; raw, analyzed population data; newspaper articles about events.
- **Natural Sciences**: Analyzed results from biological study; analyzed field data collected by environmental org; original experiments or research.
- **Addl. Examples**: court cases, newspaper articles about current events.

Secondary Sources of Data

- **Secondary sources** build off of primary sources with more extensive and in-depth analyses. Secondary sources are published works.

- **General examples**: Textbooks, monographs (books), encyclopaedias, analysis, review articles, dissertations, thesis,
- **History**: Article analyzing Prime Minister's speech; book recounting battle history of Kargil War; biographies
- **Literature**: Literary critiques such as an article that examines Tagores' writing style
- **Art**: Lecture given about Picasso's techniques; Criticism or review of Picasso's painting
- **Social Sciences**: News commentaries; Article analyzing results of theoretical or empirical study; book that discusses population trends over time; evaluations of social and government policy, law and legislation.
- **Natural Sciences**: Review articles that evaluates the theories and works of others; article on the environmental impact of pollution

Tertiary Sources of Data

- **Tertiary sources** are images and collections of primary and secondary sources.
- **General examples**: Encyclopaedias, directories, dictionaries, handbooks, guides, classification, chronology, and other fact books.
- **Addl. Sources**: Internet, online library, etc.

Primary Research

- **New and original** - no one's done it before
- Usually has a **specific purpose**
- **Good primary research** leads to

  -evidence of detailed research involving the selection and evaluation of a wide range of relevant sources.

  -critical analysis and application of the resources with clear links made to appropriate theories and concepts.

Primary Research, Primary Data and Primary Information

- **Primary research** is research that produces data that are only obtainable directly from an original source. In certain types of primary research, the researcher has direct contact with the original source of the data.
- **Primary data** are data that were previously unknown and which have been obtained directly by the researcher for a particular research project.
- **Primary information** is primary data to which meaning has been added; in other words, the data have been analysed, inferences have been drawn from them and, thereby, meaning has been added.

Why Primary data collection is important?

- Primary data collection is obligatory when a researcher cannot find the data needed in secondary sources Or when the data extracted from secondary sources are not reliable or correct.
- Primary data can be considered more trustworthy as in it is more valid than secondary data.
- Primary is conducted with a chosen research method and design and is therefore more credible than secondary data.

Primary Research methods

- **The survey method-** This is a method of primary data collection in which **questionnaire** (Mail) is used as a data collection tool.
- **The interview method** - This is a method of primary data collection in which **questionnaire** (personal or telephone) is used as a data collection tool.
- **The observational method-** This is a method of primary data collection in which researchers collect data based on their personal observations.
- Focus Groups
- Field Notes, Videos, Recordings, etc
- Case Studies

Questionnaire Method

- Provide **Quantitative Data** – data that can be measured
- Tick boxes and mostly, no more than a few words per answer
- Most useful if you want to test an existing theory or take a sample of general thoughts

Interview Method

- Provide **Qualitative Data** – what do people think and why
- Not measurable, but responses are more personal and detailed
- Most useful if you want to speak to subject experts, test an **exploratory** theory or if the meaning can be determined by context, circumstances etc.

Other Methods

- Focus Groups
  - Mostly used in market research.
  - Only useful if you want to canvas opinion on one particular concept.
- Field Notes
  - One classic example is Geography field trips.
  - Best used if you need to observe behaviour, collect samples etc.
- Case Studies
  - An in depth analysis of something bearing close relation to your research.

Primary Research Categories

- Quantitative Research
  - Numerical
  - Statistically reliable
  - Projectable to a broader population
- Quantitative Research provides the best means for understanding large populations in numerical terms
- Quantitative Research data gathered can be analyzed through inferential and descriptive statistical procedures

Primary Research Categories

- Qualitative Research
  - In-depth, insight generating
  - Non-numerical
  - Directional
- Common Techniques
  - Personal interviews (depth, one-on-one)
  - Focus groups (8-12)

- Qualitative Research can lend insight to what direction planning decisions should take

- Qualitative Research can be a useful precursor to quantitative research, questionnaire development

How to Select a Primary Research Tool?

- What kind of information do you need?
- Who can provide this information?
- Will the information be valid, accurate and reliable?
- What primary research method would best help you gather the information you need?
- Will your target audience approve?

Quantitative Research

- Quantitative research uses closed-end or forced choice questions.
- Factual and numerical questions with short responses that have precise and conclusive outcomes.

Scales of Measurement

- Scales of measurement include nominal, ordinal, interval and ratio.
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

Nominal Scale

- Nominal = name categories, count frequencies
- Data are labels or names used to identify an attribute of the element.
- A nonnumeric label or numeric code may be used.
- Variables which have no numerical value.
- Variables which have categories.
- **Example**: Students of a university are classified by the Department in which they are enrolled using a nonnumeric label such as Commerce, Pol. Sc., History, Sociology, Economics, English, Bengali, Education, and so on.
- Besides, Race, Marital status, Religion, Alive or dead

Ordinal Scale

- Ordinal = rank categories in order, but no meaningful distance
- The data have the properties of nominal data and order or rank of the data is meaningful.
- A nonnumeric label or numeric code may be used.
- Example: Students of a university are classified by their class standing using a nonnumeric label such as Sem 1, Sem 2, Sem 3 or Sem 4.
- Variables are in categories, but with an underlying order to their values.
- Rank-order categories from highest to lowest.
- Likert Scale

Interval Scale

- Interval = equal distance between scores; numerical
- You can add & subtract values, but cannot multiply & divide values. No Zero point.
- The data have the properties of ordinal data and the interval between observations is a fixed unit of measure.
- Interval data are always numeric.
- **Example**: Melissa has an SAT score of 1205, while Kevin has an SAT score of 1090. Melissa scored 115 points more than Kevin.

Ratio Scale

- Ratio = equal distance, meaningful zero
- You can add & subtract values, multiply & divide values.
- The data have all the properties of interval data and the ratio of two values is meaningful.
- Distance, height, weight and time use the ratio scale.
- This scale must contain a zero value for which nothing exists for the variable.

Qualitative Data

- Labels or names used to identify an attribute of each element or categorical data.
- Use either the nominal or ordinal scale of measurement.
- It can be either numeric or nonnumeric.
- Appropriate statistical analyses are rather limited.
- Lead to nonparametric statistics.

Quantitative Data

- Quantitative data indicate how many or how much: discrete, if measuring how many or continuous, if measuring how much.
- Quantitative data are always numeric.
- Statistical analysis is meaningful for quantitative data.
- Lead to parametric statistics.

How to Select Appropriate Test?

- Step 1: Know the scale of measurement
- Step 2: Know your goal
  - Is it to compare groups? How many groups do you have?
  - Is it to measure a relationship or association between variables?

How to Select Appropriate Test?

- What is your research question?
- Which is the dependent variable?
- What type of variables are they?
- Which statistical test is most appropriate?
- Should a parametric or non-parametric test be used?

Statistical Analysis

- **Descriptive** - What are the characteristics of the respondents?

- **Inferential** - What are the characteristics of the population? Inferential statistics allow us to decide if our sample results are probably true for the population.
  Inferential statistics allow us to decide if a treatment probably had an effect

- **Differences** - Are two or more groups the same or different?
- **Associative** - Are two or more variables related in a systematic way?

- **Predictive** - Can we predict one variable if we know one or more other variables?
- **Significance** means the use of statistical techniques that are used to determine whether the sample drawn from a population is actually from the population or if by the chance factor.

Some Non-parametric Statistics

- **Spearman rank order correlation coefficient** – nonparametric; determines monotonic symmetric relationship between two ranked variables.
- **Eta correlation coefficient** – nonparametric; determines total linear and nonlinear asymmetric relationship between one nominal and one interval/ratio variable.
- **Kendall's tau** – nonparametric; determines monotonic symmetric relationship between two ordinal variables, used when the number of rows and number of columns are equal, adjusts for tied pairs, based on concordant-discordant pairs.

Some Non-Parametric Regressions

- Kernel Regression
- Spline Regression
- Shrinkage Regression
- Bayesian Regression

Spearman's Rank Correlation

- Spearman's rank correlation

$$r_s = 1 - \left(\frac{6}{n(n^2-1)}\right)\left(\sum d^2\right)$$

- *n* = the number of observations
- *d* = the difference between the ranks for an observation

Non-Parametric Statistical Tests

- Sign Test/Median Test
- The Mann-Whitney test
- The Wilcoxon test
- The Kruskal-Wallis test
- The Friedman test
- Frequency tables, Contingency tables, Charts, Test of proportion, Chi-Square Test, McNemar test

Statistical Tests for Ordinal Data

1. The **Mann-Whitney test** evaluates the difference between two treatments or two populations using data from an independent-measures design; that is, two separate samples.

2. The **Wilcoxon test** evaluates the difference between two treatment conditions using data from a repeated-measures design; that is, the same sample is tested/measured in both treatment conditions.

Statistical Tests for Ordinal Data

3. The **Kruskal-Wallis test** evaluates the differences between three or more treatments (or populations) using a separate sample for each treatment condition.

4. The **Friedman test** evaluates the differences between three or more treatments for studies using the same group of participants in all treatments (a repeated-measures study).

Sign Test

1. **Sign test for a single population median**

- When the sample data have been randomly selected.
- It is not required that the population be normally distributed.
- It is a hypothesis test for the **population median**, not the population *mean*.
- Each of the data values is converted to either a plus sign (+) or a minus sign (-).
- If there is a predominance of plus signs to minus signs, or vice versa (depending on the form of the hypothesis test), then this is evidence against the null hypothesis.

How to perform the sign test?

Step 1: state the hypotheses		
Null	Alternative	Type of Test
H0: *M* =*M0*	*Ha* : *M* > *M0*	Right-tailed test
H0: *M* =*M0*	*Ha* : *M* < *M0*	Left-tailed test
H0: *M* =*M0*	*Ha* : *M is* not equal to *M0*	Two-tailed test
M0 is the value of the population median		

•**Step-2: Find the critical value and state the rejection rule**

*Small-sample case (when sample size *n* is less than or equal to 25)

Choose the column with the appropriate level of significance (*a*) and the applicable one-tailed or two-tailed test. Then select the row with the appropriate sample size *n* = number of pluses and minuses. The number in that row and column is your critical value *Scrit*. The rejection rule is to reject *H0* if *Sdata* is less than or equal to *Scrit*.

How to perform the sign test?

- Large-sample case (sample size *n* **is greater than** 25): The rejection rule is to reject *H0* if *Zdata* is less than or equal to *Zcrit*.
- **Step-3**: **Find the value of the test statistic**. For finding Sdata,

Type of Test	Test Statistics Sdata
Right-tailed test	*Sdata* = number of minus signs
Left-tailed test	*Sdata* = number of plus signs
Two-tailed test	*Sdata* = number of minus signs or plus signs, whichever is smaller

How to perform the sign test?

- Large-sample case ($n > 25$): First use step 3 table to find *S*data, and then calculate the test statistic *Z*data:

  $$Zdata = [(Sdata + 0.5) - n/2] / (\sqrt{n}/2)$$

- **Step 4: state the conclusion and the interpretation**.
  Compare the test statistic with the critical value, using the rejection rule. A generic interpretation is as follows. If *H*0 is rejected, then state, "Evidence exists that [whatever *Ha* says]." If *H*0 is not rejected, then state, "There is insufficient evidence that [whatever *Ha* says]."

How to calculate the sign test results?

- Suppose that we are interested in testing whether the population median M number of dengue related deaths per month is less than 50.
- Change each data value that is less than 50 to a minus sign (2), and change each data value that is greater than 50 to a plus sign (1). Ignore any data values that are equal to 50. The sample size *n* is the total number of plus signs and minus signs.
-

Month	1	2	3	4	5	6	7	8
Deaths	24	11	1	9	19	24	316	13
Sign	-	-	-	-	-	-	+	-

How to calculate the sign test results

- Step 1: **state the hypotheses. the hypotheses are** $H0$: M =50 versus Ha : M < 50
- Step 2: **find the critical value and state the rejection rule.** The total number of plus signs and minus signs is n = 7 + 1 = 8, which is not greater than 25, so we use the small-sample case. We have a one-tailed test, with a = 0.05 and n = 8, which gives us Scrit = 1. The rejection rule is to reject H0 if Sdata is less than or equal to 1.
- Step 3: **find the value of the test statistic.** We have a left-tailed test, and so, our test statistic is S = number of plus signs = 1 data
- Step 4: **state the conclusion and the interpretation.** The value of our test statistic
  is Sdata = 1, which is less than or equal to 1, so we reject H0. Evidence exists that the population median number of dengue-related deaths is less than 50 per month.

2. Sign test for Matched-Pair Data from two Dependent samples

State Accident	2014	2016	Difference (2016-2014)	Sign
1	597	613	+16	+
2	593	611	+18	+
3	508	509	+1	+
4	506	502	-4	-
5	517	505	-12	-
6	461	458	-3	-
7	512	539	+27	+
8	496	505	+9	+
9	397	386	-11	-
10	469	475	+6	+
11	491	484	-7	-
12	474	440	-34	-

Sign test for Matched-Pair Data from two Dependent samples

- The countries represent a random sample of matched-pair data, so the condition for performing the sign test for the population median of the differences is met.

- Step 1: **state the hypotheses.** We have a left-tailed test: *H*0: *Md* = 0 versus *Ha* : *Md* < 0 where *Md* represents the population median of the differences.
- Step 2: **find the critical value and state the rejection rule.**

  The sample size is the sum of the number of plus signs and minus signs: *n* = 6 + 6 = 12. Because *n* < 25, we use the small-sample case. We have a one-tailed test, with *a* = 0.05 and *n* = 12, which gives us *S*crit = 2. The rejection rule is to reject *H*0 if *S*data is less than or equal to 2.
- Step 3 **find the value of the test statistic.**

  We have *S*data = the number of plus signs = 6.
- Step 4 **state the conclusion and the interpretation.** Because *S*data = 6 is not less than or equal to 2, we do not reject *H*0. There is insufficient evidence that the population median score has decreased from 2016 to 2016.

3. Sign Test for Binomial data

- Suppose you have collected the numbers of spam emails and non-spam emails processed by a university spam filter.

- When using the sign test, spam emails are represented by plus (+) signs, and non-spam emails are represented by minus (-) signs.

- We use the same methods for the sign test for binomial data that we used for the sign test for a single population median. However, only the large-sample case is used ($n > 25$), because only when the sample size is large does the Central Limit Theorem apply.

- Note that the hypothesized population proportion is always $p0$ = 0.5.

The Mann-Whitney U Test (Wilcoxon Rank Sum Test)

- The test uses the data from two separate samples to test for a significant difference between two populations.
- However, the Mann-Whitney test only requires that you are able to rank order the individual scores; there is no need to compute means or variances.
- H0: There is no systematic or consistent difference between the two populations being compared.

The Mann-Whitney U Test (Wilcoxon Rank Sum Test)

- **Step 1**. Set up hypotheses and determine level of significance.

  H_0: The two populations are equal

  H_1: The two populations are not equal. α =0.05
- **Step 2**. Select the appropriate test statistic.
- **Step 3**. Set up decision rule.
- **Step 4**. Compute the test statistic.
- **Step 5**. Conclusion

The Mann-Whitney U Test

- The calculation of the Mann-Whitney U statistic requires:

1. Combine the two samples and rank order the individuals in the combined group.
2. Once the complete set is rank ordered, you can compute the Mann-Whitney U by either

   a. Find the sum of the ranks for each of the original samples, and use the formula to compute a U statistic for each sample.

   b. Compute a score for each sample by viewing the ranked data (1st, 2nd, etc.) as if they were finishing positions in a race. Each subject in sample A receives one point for every individual from sample B that he/she beats.

   The total number of points is the U statistic for sample A. In the same way, a U statistic is computed for sample B.
3. The Mann-Whitney U is the smaller of the two U statistics computed for the samples.

The Mann-Whitney U Test

- If there is a consistent difference between the two treatments, the two samples should be distributed at opposite ends of the rank ordering.
- In this case, the final U value should be small. At the extreme, when there is no overlap between the two samples, you will obtain U = 0.
- Thus, a small value for the Mann-Whitney U indicates a difference between the treatments.
- To determine whether the obtained U value is sufficiently small to be significant, you must consult the Mann-Whitney table.
- For large samples, the obtained U statistic can be converted to a z-score and the critical region can be determined using the unit normal table.

The Mann-Whitney U Test

- **U = [N1 * N2] + [NX * (NX + 1)/2 – TX]**

  Where N1 = The number of people in Rank 1, N2 = The number of people in Rank 2, NX = The number of people in the group that gave the larger rank total and TX = Higher of two ranks total.

The Mann-Whitney U Test

Participant	Rating A	Rank 1	Participant	Rating B	Rank 2
A	3	3	A	9	11
B	4	4	B	7	9
C	2	1.5	C	5	5.5
D	6	7.5	D	10	12
E	2	1.5	E	6	7.5
F	5	5.5	F	8	10
	T1	**23**		**T2**	**55**

TX = Higher of T1 or T2 = 55
N1 is the number of people in Rank 1 = 6 and N2 is the number of people in Rank 2 = 6 . NX is the number of people in the group that gave the larger rank total = 6

The Mann-Whitney U Test

U = [N1 * N2] + [NX * (NX + 1)/2 – TX]

U = [6 * 6] + [6 * (6+1)/2 – 55]

U = 36 + 21 – 55

U = 2

Table value of U = 5 at 5% level of significance of (6,6 degrees of freedom) (Calculated value is less than table value).

Therefore, we can conclude that the difference that we have found between the ratings for the two ratings is unlikely to have occurred by chance.

The Wilcoxon T test

- The **Wilcoxon test** uses the data from one sample where each individual has been observed in two different treatment conditions to test for a significant difference between the two treatments.
- However, the Wilcoxon test only requires that you are able to rank order the difference scores; there is no need to measure how much difference exists for each subject, or to compute a mean or variance for the difference scores.
- H0: There is no systematic or consistent difference between the two treatments being compared.

The Wilcoxon T test

The calculation of the Wilcoxon T statistic requires:

1. Observe the difference between treatment 1 and treatment 2 for each subject.
2. Rank order the absolute size of the differences without regard to sign (increases are positive and decreases are negative).
3. Find the sum of the ranks for the positive differences and the sum of the ranks for the negative differences.
4. The Wilcoxon T is the smaller of the two sums.

The Wilcoxon T test

- If there is a consistent difference between the two treatments, the difference scores should be consistently positive (or consistently negative).
- At the extreme, all the differences will be in the same direction and one of the two sums will be zero. (If there are no negative differences then ΣRanks = 0 for the negative differences.)
- Thus, a small value for T indicates a difference between treatments.

The Wilcoxon T test

Participant	RA	RB	Difference	Ranked difference
1	25	32	-7	7.5
2	29	30	-1	1.5
3	10	7	3	4
4	31	36	-5	6
5	27	20	7	7.5
6	24	32	-8	9
7	27	26	1	1.5
8	29	33	-4	5
9	30	32	-2	3
10	32	32	0	ignore
11	20	30	-10	10
12	5	32	-27	11

The Wilcoxon T test

- Add positive sign in ranked difference = 4 + 7.5 + 1.5 = 13
- Add negative sign in ranked difference = 7.5 + 1.5 + 6 + 9 + 5 + 3 + 10 + 11 = 53
- W = Smaller of above two = 13 (Calculated Value)
- Degree of freedom = N – 1 (omitting 0) = 11
- Table Value at 5% level of Significance = 11 at 11 degrees of freedom
- Since calculated value (13) is higher than 11 and then we can conclude that there is no significant difference between the two ratings.

The Wilcoxon T test

- To determine whether the obtained T value is sufficiently small to be significant, you must consult the Wilcoxon table.
- For large samples, the obtained T statistic can be converted to a z-score and the critical region can be determined using the unit normal table.

The Kruskal-Wallis Test

- The **Kruskal-Wallis test** uses data from three or more separate samples to evaluate differences among three or more treatment conditions.
- The test requires that you are able to rank order the individuals but does not require numerical scores.
- H0: There are no systematic or consistent differences among the treatments being compared.

The Kruskal-Wallis Test

The calculation of the Kruskal-Wallis statistic requires:

1. Combine the individuals from all the separate samples and rank order the entire group.
2. Regroup the individuals into the original samples and compute the sum of the ranks (T) for each sample.
3. A formula is used to compute the Kruskal-Wallis statistic which is distributed as a chi-square statistic with degrees of freedom equal to the number of samples minus one.

The obtained value must be greater than the critical value for chi-square to reject H_0 and conclude that there are significant differences among the treatments.

The Kruskal-Wallis Test

- $H = [12/N(N+1) * \sum Tc2/Nc] - 3*(N+1)$

Where N = Total number of participants (combined)

Tc is the rank total for each group.

Nc is the number of participants in each group.

The Kruskal-Wallis Test

Participant	RA	RB	RC	Rank 1	Rank 2	Rank 3
1	23	22	59	2	1	20
2	26	27	66	3	4	24
3	51	39	38	16	9	8
4	49	29	49	14	5.5	14
5	58	46	56	19	11	17.5
6	37	48	60	7	12	21
7	29	49	56	5.5	14	17.5
8	44	65	62	10	23	22
Total				**76.5**	**79.5**	**144**

The Kruskal-Wallis Test

- H = [12/N(N+1) * ∑Tc2/Nc] – 3*(N+1)

  H = [12/24(24+1) * (76.5)2/8 + (79.5)2/8 + (144)2/8] – 3 * (24+1)

  H = [12/600 * 4113.5625] – 75

  H = [0.02 * 4113.5625] – 75

  H = 7.27 (calculated value)

Degrees of freedom = 3 group – 1 = 2

Table value of Chi-square at 5% level of significance = 5.99

So we would conclude that there is a difference of some kind between our three groups.

The Kruskal-Wallis Test

- Assessing the significance of H depends on the number of participants and the number of groups.
- If you have three groups, with five or fewer participants in each group, then you need to use the special table for small sample sizes.
- If you have more than five participants per group, then treat *H* as Chi-Square. H is statistically significant if it is equal to or larger than the critical value of Chi-Square for your particular d.f.

The Friedman Test

- The **Friedman test** uses data from one sample to evaluate differences among three or more treatment conditions.
- The test requires that you are able to rank order the individuals across treatments but does not require numerical scores.
- H0: There are no systematic or consistent differences among the treatments being compared.

The Friedman Test

The calculation of the Friedman statistic requires:

1. Each individual (or the individual's scores) must be ranked across the treatment conditions.
2. The sum of the ranks (R) is computed for each treatment.
3. A formula is used to compute the Friedman test statistic which is distributed as chi-square with degrees of freedom equal to the number of treatments minus one.

The obtained value must be greater than the critical value for chi-square to reject H_0 and conclude that there are significant differences among the treatments.

The Friedman Test

- Friedman's F
  = [12/[N*K*(K+1)]] * ∑R2 – [3*N*(K+1)]
  Where N = the number of subjects, K = the number of conditions, R = the sum of the ranks for each condition

The Friedman Test

Score 1	Score 2	Score 3	Rank 1	Rank 2	Rank 3
12	32	34	1	2	3
14	41	38	1	3	2
15	31	45	1	2	3
12	38	32	1	3	2
7	21	12	1	3	2
4	13	11	1	3	2
10	17	22	1	2	3
4	22	9	1	3	2
14	24	20	1	3	2
4	11	8	1	3	2
5	17	19	1	2	3
10	20	8	2	3	1
Total			**13**	**32**	**27**

The Friedman Test

- Friedman's F
  = [12/[N*K*(K+1)]] * ∑R2 – [3*N*(K+1)]
- F = [12/12*3*(3+1)]]*[(13)2 + (32)2 + (27)2] – [3*12*(3+1)]
- F = [12/144]*[169+1024+729] – 144
- F = [0.083*1922] – 144
- F = 15.526

Table value = 8.67 at 5% level of significance with (k-1 = 2) and (3,12) degrees of freedom

Since the calculated value of F is more than the critical value, we reject the null hypothesis.

The Friedman Test

- IF you reject the null hypothesis, determine whether the pattern of the data completely supports, partially supports or does not support the research hypothesis.
  -- IF you reject the null hypothesis, AND if the pattern of data agrees exactly with the research hypothesis, then the research hypothesis is completely supported.
  -- IF you reject the null hypothesis, AND if part of the pattern of the data agrees with the research hypothesis, BUT part
  of the pattern of the data does not, then the research hypothesis is partially supported.
  -- IF you retain the null hypothesis, OR you reject the null BUT NO PART of the pattern of the data agrees with the research hypothesis, then the research hypothesis is not at all supported.

When is Chi-square Test used?

- It is non-parametric hypothesis test.
- You have two categorical variables from a single population.
- The data are frequencies rather than numerical scores.
- Whether there is a significant association between the two variables.

Types

1. **Tests of goodness-of-fit**

   Observed frequencies of one variable are significantly different from the expected frequencies of the same variable.

2. **Chi-Square tests of independence** (or relationship)

Two variables are associated or independent of the other.

Chi-Square - Goodness-of-Fit Test

Step – 1: H0 : There are the same number of accidents each day of the week

H1 : There are not the same number of accidents each day of the week.

Step-2: test statistic $\chi 2 = \sum(O-E)2/E$

Step-3: Degrees of Freedom (v) = (number of categories after pooling) – (number of parameters estimated) – 1

Step – 4: Collect table value at 5% level of significance

Step – 5: Compare the calculated value with table and conclude.

Chi-Square - Goodness-of-Fit Test

Day	No. of Accidents (Observed)	No. of Accidents (Expected)	O-E	(O-E)2/E
Monday	23	20	3	9/20
Tuesday	18	20	-2	4/20
Wednesday	17	20	-3	9/20
Thursday	19	20	-1	1/20
Friday	23	20	3	9/20
				∑(O-E)2/E = 1.6

Table Value at 5% level = 9.49; Degrees of Freedom (v) = 5 – 0 – 1 = 4

Conclusion: Since the calculated value is less than table value at 5% level of significance, there is no evidence to reject the null hypothesis. Then we conclude that there is insufficient evidence to suggest that accidents do not occur uniformly, that is, we conclude that the number of accidents is, on the whole, the same for each day of the week

Chi-Square tests of independence

- If there are two categorical variables and our interest is to examine whether these two variables are associated with each other, the chi-square ($\chi 2$) test of independence is the correct tool to use.
- **Step-1: H0**: There is no association between the brand preference and income level (These two attributes are independent).

  H1: There is association between brand preference and income level (These two attributes are dependent).

  Step-2: test statistic $\chi 2 = \sum(O-E)2/E$

  Step-3: Degrees of Freedom (v) = (the number of rows - 1) * (the number of columns - 1)

  Step – 4: Collect table value at 5% level of significance

  Step – 5: Compare the calculated value with table and conclude.

Chi-Square tests of independence

	Brands (Observed)				
	Brand 1	Brand 2	Brand 3	Brand 4	Total
Lower Income	25	15	55	65	**160**
Middle Income	30	25	35	30	**120**
Upper Middle	50	55	20	22	**147**
Upper Income	60	80	15	18	**173**
Total	**165**	**175**	**125**	**135**	600

Expected Frequency = (Row Total * Column Total)/Grand Total

Chi-Square tests of independence

	Brands (Expected)				
	Brand 1	Brand 2	Brand 3	Brand 4	Total
Lower Income	44.000	46.667	33.333	36.000	**160**
Middle Income	33.000	35.000	25.000	27.000	**120**
Upper Middle	40.425	42.875	30.625	33.075	**147**
Upper Income	47.575	50.458	36.042	38.925	**173**
Total	165.000	175.000	125.000	135.000	**600**

Chi-Square tests of independence

	Brand 1	Brand 2	Brand 3	Brand 4
Lower Income	8.20	21.49	14.08	23.36
Middle Income	0.27	2.86	4.00	0.33
Upper Middle	2.27	3.43	3.69	3.71
Upper Income	3.24	17.30	12.28	11.25

$\chi 2$ = 131.76; Degrees of freedom = (4-1)*(4-1) = 9
Table Value at 5% level of significance = 16.92

Conclusion: Since the calculated value is more than table value at 5% level of significance, there is enough evidence to reject the null hypothesis. Then we conclude that there is sufficient evidence to suggest that There is association between brand preference and income level (These two attributes are dependent).

When to Use what type of data

- Quality/different – Nominal data
- Difference/which is bigger – Ordinal data
- different? Which is bigger? how much larger? – Interval data

Reliability = *consistency* in procedures and in reactions of participants

- Validity = *truth* - Does it measure what it intended to measure?
- When reliability and validity are achieved, data are free from systematic errors

Odds Ratio

- Odds of an event is the probability it occurs divided by the probability it does not occur
- Odds ratio is the odds of the event for group 1 divided by the odds of the event for group 2
- Sample odds of the outcome for each group:

  Odds1 = n11/n1 / n12/n1 = n11/n12

  Odds2 = n21/n22

  Estimated Odds Ratio: Odds1/Odds2 = n11n22/n12n21

Odds Ratio

- Interpretation
  - Conclude that the probability that the outcome is present is higher (in the population) for group 1 if the entire interval is above 1
  - Conclude that the probability that the outcome is present is lower (in the population) for group 1 if the entire interval is below 1
  - Do not conclude that the probability of the outcome differs for the two groups if the interval contains 1

Thank You All