
# Unit 3: Lecture 3
# Hypothesis testing

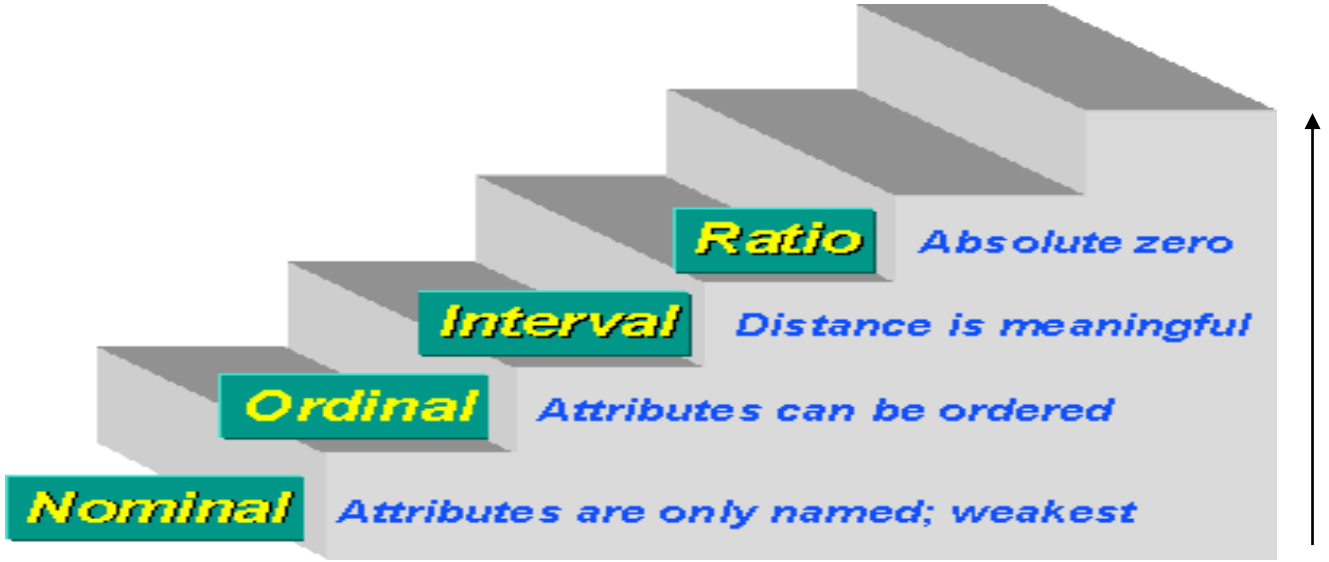
# Types of Variables

Independent variable—the presumed cause (of a dependent variable)

Dependent variable—the presumed effect (of an independent variable)

Example: Smoking (IV) → Lung cancer (DV)

# Levels of Measurement



# What Type of Data To collect?

- The goal of the researcher is to use the highest level of measurement possible.
  - Example: Two ways of asking about Smoking behavior. Which is better, A or B?

    (A)Do you smoke? ▫Yes ▫No

    (B)How many cigarettes did you smoke in the last 3 days (72 hours)?

  (A) Is nominal, so the best we can get from this data are frequencies. (B) is ratio, so we can compute: mean, median, mode, frequencies.

# Inferential Statistics

- Inferential statistics are used to test hypothesis (prediction) about relationship between variables in the population. A relationship is an association between variables.
- It consists of a set of statistical techniques that provide prediction about population characteristics based on information in a sample from population.
- An important aspect of statistical inference involves reporting the likely accuracy, or of confidence of the sample statistic that predicts the value of the population parameter.

# Research Hypothesis

- In research, a **hypothesis** is a clear, testable statement predicting the relationship between variables or the outcome of a study.
- Hypothesis form the foundation of scientific inquiry, providing a direction for investigation and guiding the data collection and analysis process.
- In qualitative research, there is NO hypothesis

# Research Hypothesis

- A tentative prediction or explanation of the relationship between two or more variables
- It's a translation of research question into a precise prediction of the expected outcomes
- Must always involve at least two variables
- Must suggest a predicted relationship between the independent variable and the dependent variable
- Must contain terms that indicate a relationship (e.g., more than, different from, associated with)

# Hypothesis Criteria

- Written in a declarative form.
- Written in present tense.
- Contain the population
- Contain variables
- Reflects problem statement or purpose statement
- Empirically testable

## Hypothesis Testing

- A hypothesis is made about the value of a parameter, but the only facts available to estimate the true parameter are those provided by the sample.
- If the statistic differs (and of course it will) from the hypothesis stated about the parameter, a decision must be made as to whether or not this difference is *significant*. If it is, the null hypothesis is rejected. If not, it cannot be rejected.

# Hypothesis Testing

- $H_0$: The null hypothesis. This contains the hypothesized parameter value which will be compared with the sample value.
- $H_1$: The alternative hypothesis. This will be "accepted" only if $H_0$ is rejected.

**Technically speaking, we never accept $H_0$ What we actually say is that we do not have the evidence to reject it.**

# Two Types of Errors: Alpha and Beta

- Two types of errors may occur: α (alpha) and β (beta).
- The α error is often referred to as a Type I error and β error as a Type II error.
  - You are guilty of an alpha error if you reject $H_0$ when it really is true.
  - You commit a beta error if you "accept" $H_0$ when it is false.

| | | STATE OF NATURE | |
|---|---|---|---|
| | | $H_0$ Is True | $H_0$ Is False |
| DECISION | Do Not Reject $H_0$ | GOOD | β Error (Type II Error) |
| | Reject $H_0$ | α Error (Type I Error) | GOOD |

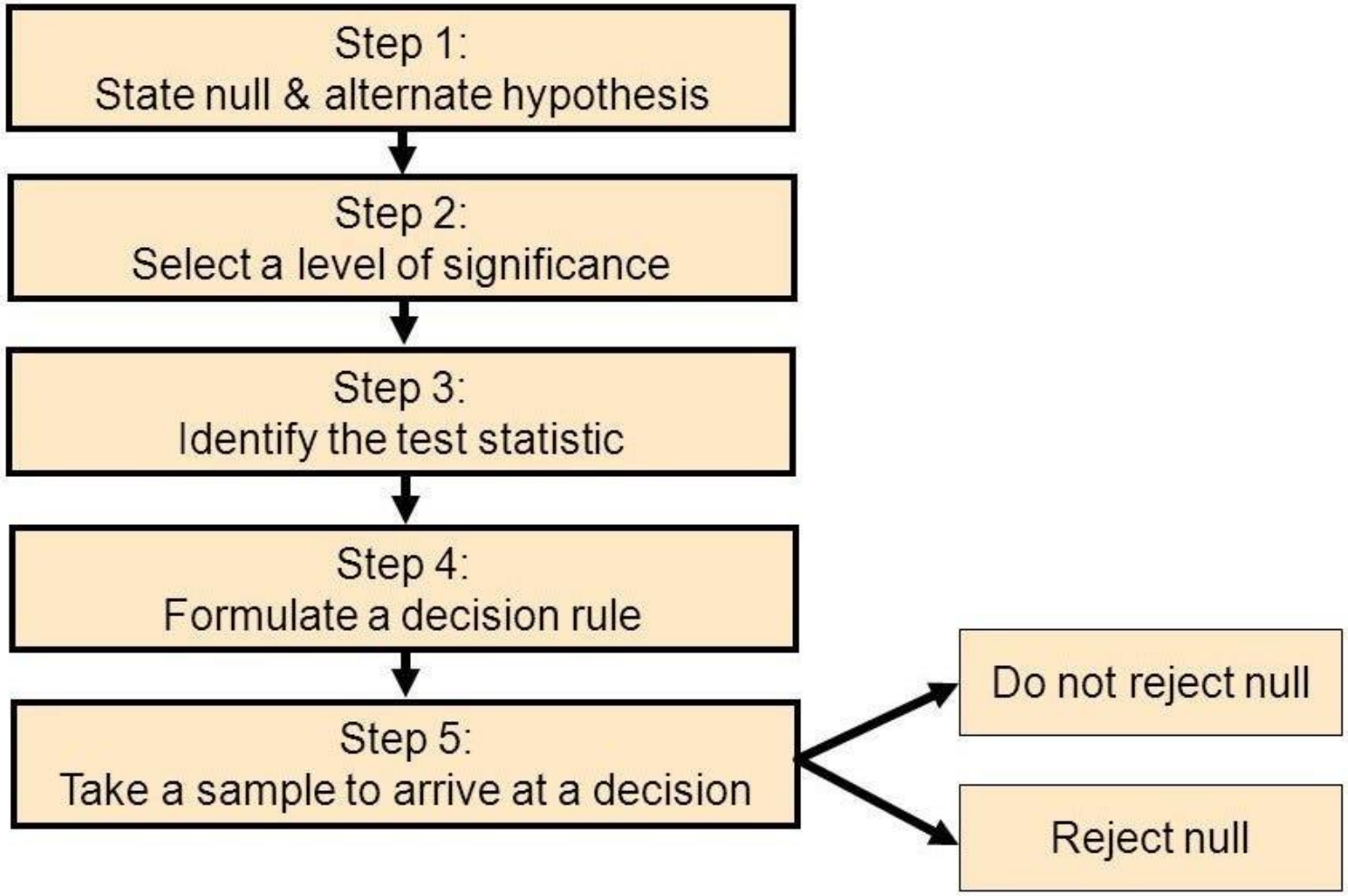
# Types of Errors

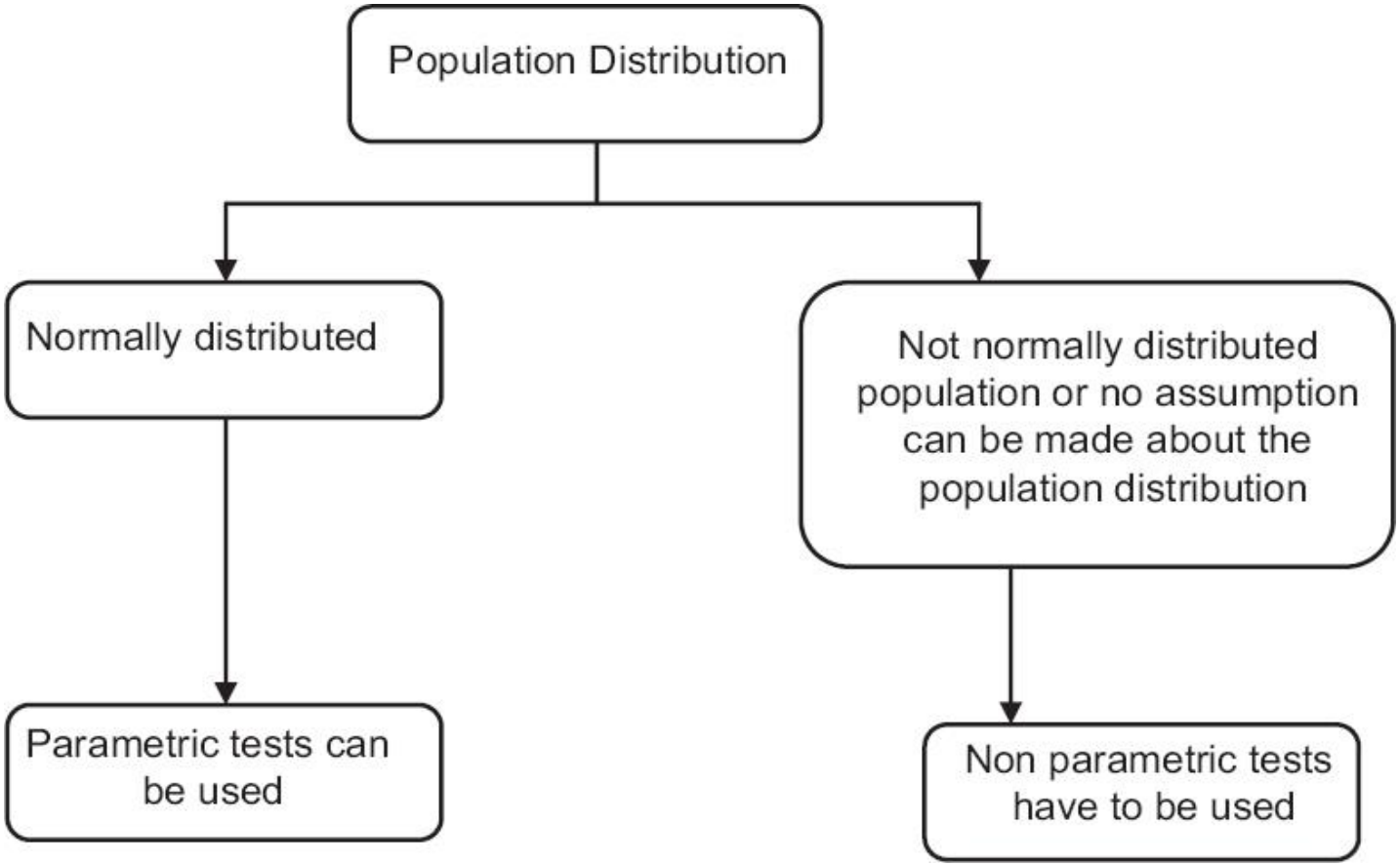
| If You…… | When the Null Hypothesis is… | Then You Have……. |
|---|---|---|
| Reject the null hypothesis | True (there really are no difference) | Made a Type I Error |
| Reject the null hypothesis | False (there really are difference) |  |
| Accept the null hypothesis | False (there really are difference) | Made Type II Error |
| Accept the null hypothesis | True (there really are no difference) |  |

- **Key Characteristics of a Hypothesis**:

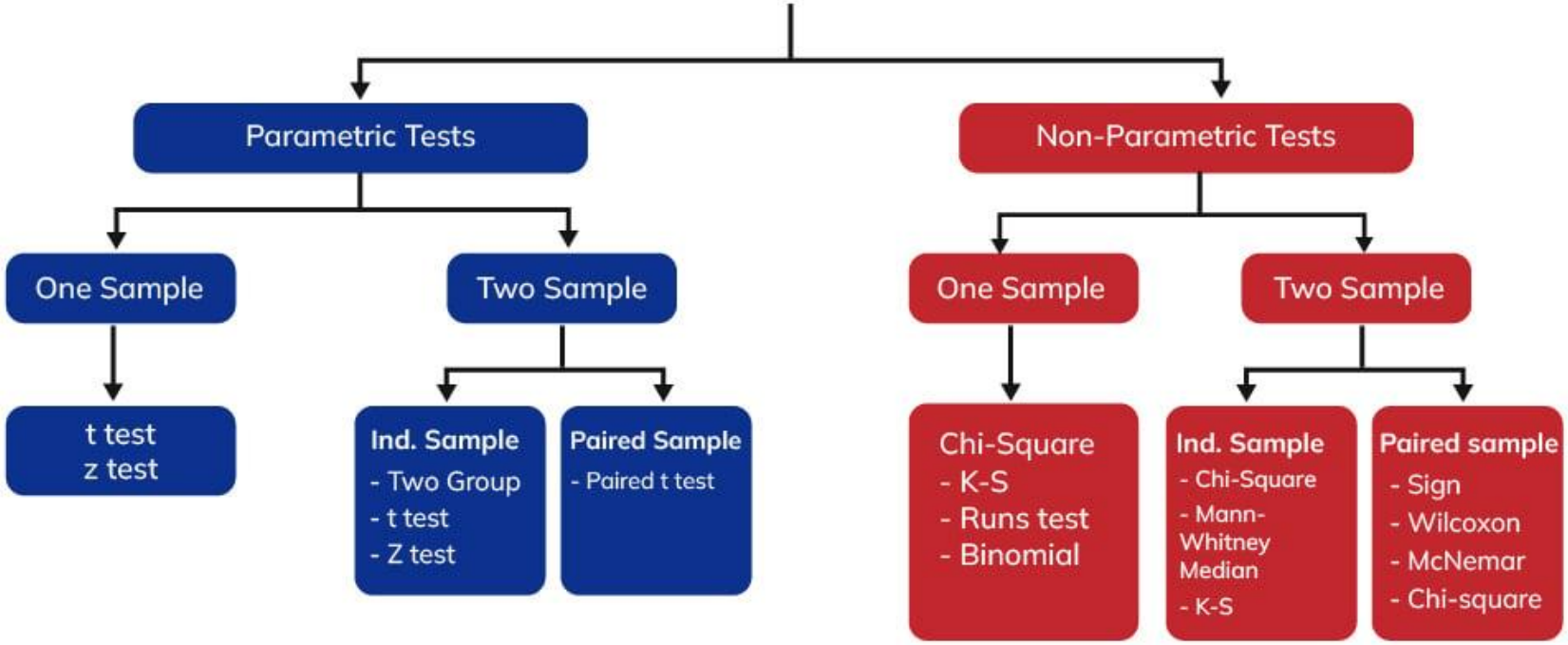
- **Testable**: Must be possible to evaluate or observe the outcome through experimentation or analysis.
- **Specific**: Clearly defines variables and the expected relationship or outcome.
- **Predictive**: States an anticipated effect or association that can be confirmed or refuted.
- **Example**: “Increasing the amount of daily physical exercise will lead to a reduction in stress levels among college students.”

# Five step procedure for testing a hypothesis



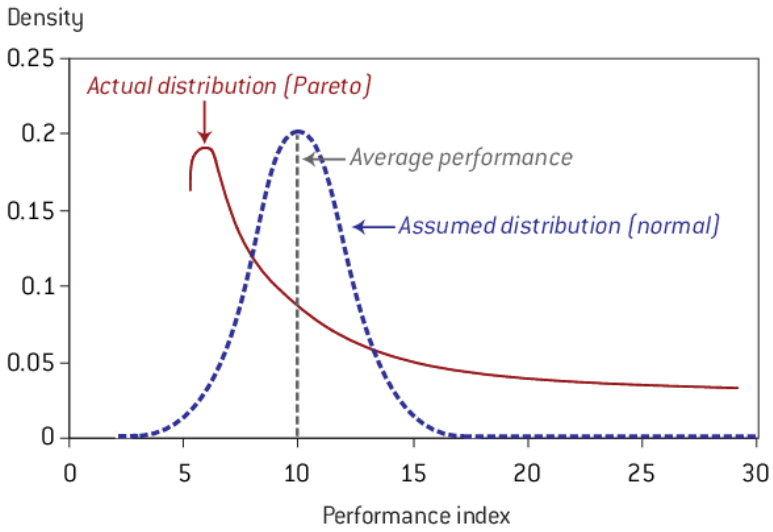

Population Distribution
Normally distributed
Not normally distributed population or no assumption can be made about the population distribution
Parametric tests can be used
Non parametric tests have to be used

# Parametric & Non-Parametric Test



# Parametric Assumptions

✓The observations must be independent.
✓Dependent variable should be continuous (I/R)
✓The observations must be drawn from norm distributed populations
✓These populations must have the same variances. Equal variance (homogeneity of variance)
✓The groups should be randomly drawn from normally distributed and independent popul

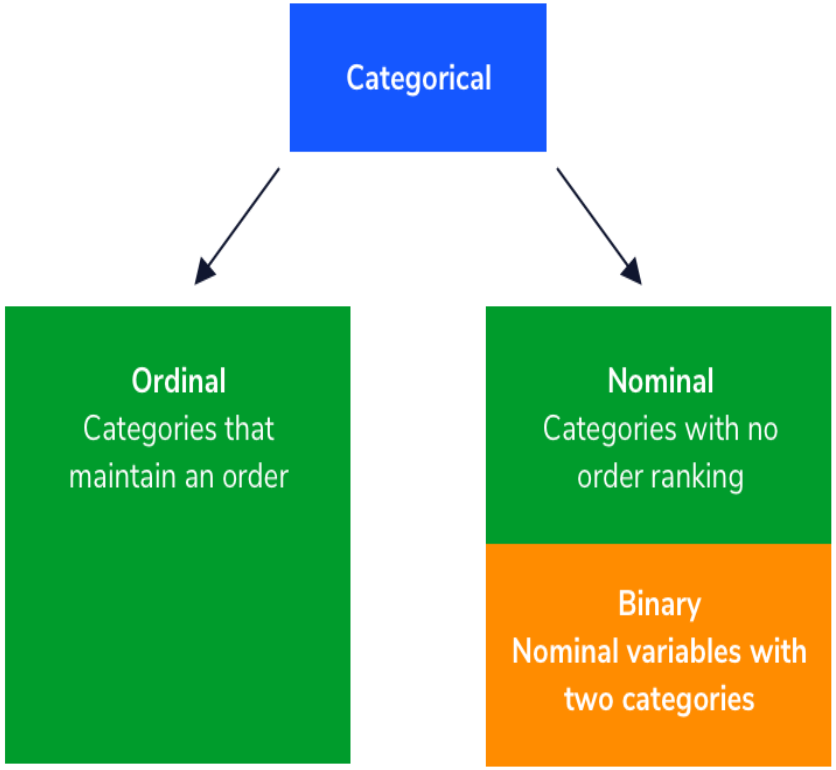



e.g.
Male X Female
Pharmacist X Physician
Manager X Staff

✓ **NO OVER LAP**

# Parametric Assumptions

- The independent variable is categorical with two or more levels.
- Distribution for the **two or more independent** variables is normal.



# Advantages of Parametric Techniques

- They are more powerful and more flexible than nonparametric techniques.

- They not only allow the researcher to study the effect of many independent variables on the dependent variable, but they also make possible the study of their interaction.

# Non-parametric tests

**Nonparametric tests**

- Make no assumptions about the data's characteristics.

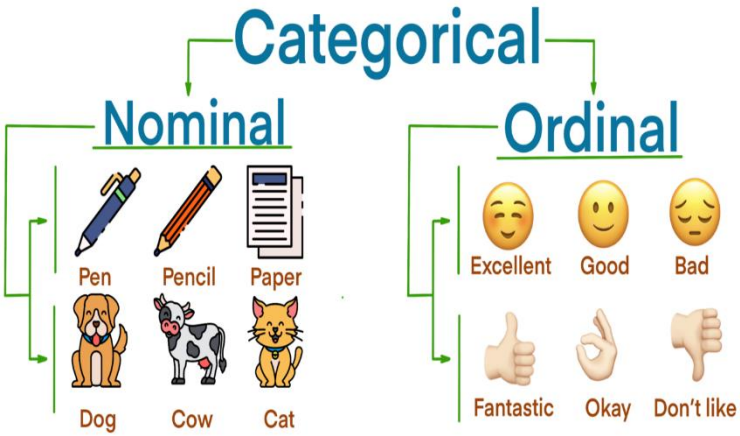
- Use if *any* of the three properties below are true:

- (a) the data are **not normally distributed** (e.g. skewed);
- (b) the data show **in-homogeneity of variance**;
- (c) the data are measurements on an **ordinal scale** (**ranks**).

- Non-parametric tests are used when we do not have ratio/interval data, or when the assumptions of parametric tests are broken

# Nonparametric Methods

- Nonparametric methods are often the only way to analyze nominal or ordinal data and draw statistical conclusions.
- Nonparametric methods require no assumptions about the population probability distributions.
- Nonparametric methods are often called distribution-free methods.
- Nonparametric methods can be used with small samples



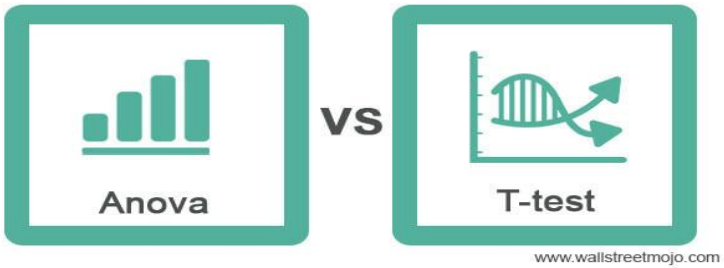
# Difference between Parametric & Non-parametric test

| | Parametric test | Non parametric test |
|---|---|---|
| 1. | Used for ratio or interval data | For ordinal or nominal data |
| 2. | Used for Normal distribution | Any distribution |
| 3. | Mean is usual central measure | Median is usual central measure |
| 4. | Information about population is completely known | No information available |
| 5. | Specific assumptions made regarding population | Assumption free test |

# Types of Statistical Tests

When running a *t test* and *ANOVA*

- We compare:
  - Mean differences between groups
- We assume
  - random sampling
  - the groups are homogeneous
  - distribution is normal
  - samples are large enough to represent population (>30)
  - Data represented on an **interval or ratio** scale
- These are Parametric tests



# Types of Tests

**When the assumptions are violated:**

- Subjects were not randomly sampled

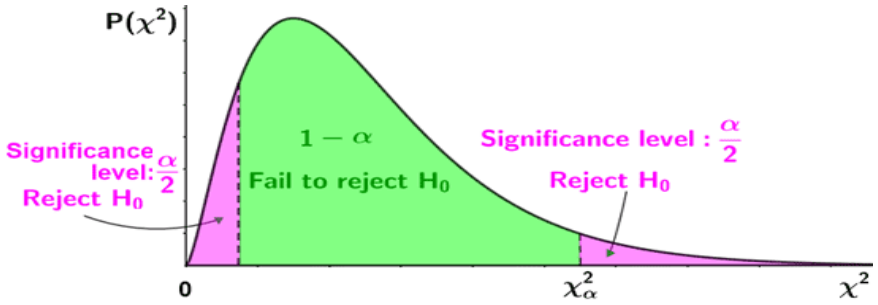


  - Ordinal (ranked)
  - Nominal (categorized: t styles)
  - The scores are greatly skewed or we have no knowledge of the distribution

**We use *Non-Parametric Tests***

# Chi-Square test

- Must be a random sample from population
- Data must be in **raw frequencies**
- Variables must be independent
- Observations must be independent.
- Does not prove causality.




| Count of Sex | | Health | | | |
|---|---|---|---|---|---|
| Age | Sex | Average | Good | Poor | Grand Total |
| 0-15 | Female | 8 | 7 | 4 | 19 |
| 0-15 | Male | 5 | 9 | 3 | 17 |
| 16-29 | Female | 9 | 10 | 5 | 24 |
| 16-29 | Male | 4 | 5 | 10 | 19 |
| 30-44 | Female | 2 | 5 | 2 | 9 |
| 30-44 | Male | 5 | 4 | 5 | 14 |
| 45-64 | Female | 4 | 6 | 7 | 17 |
| 45-64 | Male | 4 | 8 | 4 | 16 |
| 65+ | Female | 10 | 5 | 10 | 25 |
| 65+ | Male | 3 | 8 | 6 | 17 |

# Chi-Square test

- The chi square test can only be used on data that has the following characteristics:

The data must be in the form of frequencies

The frequency data must have a precise numerical value and must be organised into categories or groups.

The expected frequency in any one cell of the table must be greater than 5.

The total number of observations must be greater than 20.

| Degree | Frequency | Relative Frequency | Percentage |
|---|---|---|---|
| High School | 2 | 0.050 | 5.0 |
| Bachelor's | 7 | 0.175 | 17.5 |
| MBA | 20 | 0.500 | 50.0 |
| Master's | 3 | 0.075 | 7.5 |
| Law | 4 | 0.100 | 10.0 |
| PhD | 4 | 0.100 | 10.0 |
| | 40 | | |

# Formula

$$\chi^2 = \frac{\Sigma (O - E)^2}{E}$$

$\chi^2$ = The value of chi square
$O$ = The observed value
$E$ = The expected value
$\Sigma (O - E)^2$ = all the values of (O – E) squared then added together

# Chi Square Test of Independence

- Hypotheses
  - The null hypothesis is that the two variables are independent.
    - **$H_0$: X and Y are independent**
    - **$H_1$: X and Y are dependent**
  - Some examples where one might use the chi-squared test of independence are:
    - Is level of education related to level of income?
    - Is the level of price related to the level of quality in production?

# 2. Establish Level of Significance

- $\alpha$ is a predetermined value
- The convention
  - $\alpha = .05$
  - $\alpha = .01$
  - $\alpha = .001$

# 3. Determine The Hypothesis: Whether There is an Association or Not

- **$H_o$** : The two variables are independent

- **$H_a$** : The two variables are associated

# Decision and Interpretation

- The decision rule is: if the p-value for the test is less than 0.05, we reject the null hypothesis, but if it is greater than or equal to 0.05, we fail to reject the null hypothesis.
- If the probability of the test statistic is less than or equal to the probability of the alpha error rate, we reject the null hypothesis and conclude that there is a relationship between the variables.
- If the probability of the test statistic is greater than the probability of the alpha error rate, we fail to reject the null hypothesis. We conclude that there is no relationship between the variables, i.e. they are independent.

**Community medicine: Biostatistics Attendance Barcode**

- Register your attendance with your university number
- Make sure that the settings of your phone allow tracking location. The location should be turned on to complete this process. Go to settings > privacy> location> services> make sure that location services is ON
- If the student's name doesn't pop up, he should stop by the examination office to add it manually.