# Statistics for Data Science

- Definitions with examples
- Types

# Parametric test

Definition: A parametric test is a statistical test which makes certain assumptions about the distribution of the unknown parameter of interest and thus the test statistic is valid under these assumptions. These tests are those which assume that the sample data comes from a population that follows a probability distribution, (The normal distribution-with fixed set of parameters)

```
Types of parametric tests:
1. Two sample t-test
2. Paired t-test
3. Analysis of Variance (ANOVA)
4. Pearson coefficient of correlation
```

# 1. Two sample or Independent sample t-test

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not. We can use the test when our data values are independent, are randomly sampled from two normal populations and the two independent groups have equal variances.

For the two-sample t-test, we need two variables. One variable defines the two groups. The second variable is the measurement of interest.

Examples:

- Local Male students in class and International male students in class
- Students in a class who speak english as their first language and students who do not.

# 2. Paired Samples t-test

The Paired Samples t-test compares the means of two measurements taken from the same individual, object, or related units.Data is in the form of matched pairs.You can use the test when your data values are paired measurements. The paired t-test is also known as the dependent samples t-test, the paired-difference t-test, the matched pairs t-test and the repeated-samples t-test.

Examples:

- comparison of height of female students in one class
- comparison of water stress on plant length

# 3. Analysis of Variance (ANOVA)

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. It is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. ANOVA is helpful for testing three or more variables.

Examples:

- test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges.
- Effect of three host plant leaves on the oviposition rate of insects

## There are different types of ANOVA which are as follows:

- One-way ANOVA: The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. One-way ANOVA is typically used when you have a single independent variable, or factor, and your goal is to investigate if variations, or different levels of that factor have a measurable effect on a dependent variable. It involves one factor or independent variable. In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. Example: As a crop researcher, you want to test the effect of three different fertilizer mixtures on crop yield.

- Two-way ANOVA: A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. There are two independent variables in a two-way ANOVA. A two-way ANOVA compares multiple groups of two factors. Example: To check which type of fertilizer and planting density produces the greatest crop yield in a field experiment.

- Repeated measures ANOVA: The repeated measures ANOVA compares means across one or more variables that are based on repeated observations. A repeated measures ANOVA model can also include zero or more independent variables. Again, a repeated measures ANOVA has at least 1 dependent variable that has more than one observation. In other words, a repeated measures ANOVA is used to determine whether or not there is a statistically significant difference between the means of three or more groups in which the same subjects show up in each group. Example: To measure the mean scores of subjects during three or more time points (i.e. 1st semester, mid semester, after mid semester)

- ANCOVA (analysis of covariance): ANCOVA is a technique that remove the impact of one or more metric-scaled undesirable variable from dependent variable before undertaking research. It is the midpoint between ANOVA and regression analysis, wherein one variable in two or more population can be compared while considering the variability of other variables. When there is a set of both factors (categorical independent variables) and (covariate metric independent variable) then it is known as ANCOVA. It only use linear model and includes categorical and interval variables. Example: To test the effects of antiinflammatory drugs on blood pressure in participants of varying age. The change in BP after treatment (a continuous variable) is the dependent variable, and the independent variables might be age (a continuous variable) and treatment (a categoric variable).

- MANOVA (multivariate analysis of variance): Multivariate analysis of variance is a procedure for comparing multivariate sample means. As a multivariate procedure, it is used when there are two or more dependent variables, and is often followed by significance tests involving individual dependent variables separately. IN other words, it is used to determine whether multiple levels of independent variables on their own or in combination with one another have an effect on the dependent variables. MANOVA requires that the dependent variables meet parametric requirements. Example: you can use a one-way MANOVA to understand whether there were differences in the perceptions of attractiveness and intelligence of insect feeders in cage (i.e., the two dependent variables are "perceptions of attractiveness" and "perceptions of intelligence", whilst the independent variable is "insect feeders".

- MANCOVA (multivariate analysis of covariance): Multivariate analysis of covariance (MANCOVA) is the extension of analysis of covariance (ANCOVA). Basically, it is the multivariate analysis of variance (MANOVA) with a covariate(s). In MANCOVA, we assess for statistical differences on multiple continuous dependent variables by an independent grouping variable, while controlling for a third variable called the covariate; multiple covariates can be used, depending on the sample size.

# 4. Pearson coefficient of correlation

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship or association, between two continuous variables. It is known to be the best method of measuring the association between variables of interest because it is based on the method of covariance. When one variable changes, the other variable changes in the same direction.

Examples:

- To a certain age a child's height will keep increasing as his/her age increases.
- the income of a person increases as his/her age increases.

# Non-Parametric test

Definition: In statistics, nonparametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed). Due to this reason, they are sometimes referred to as distribution-free tests. The non-parametric tests are used as an alternative method to parametric tests, not as their substitutes. In other words, if the data meets the required assumptions for performing the parametric tests, the relevant parametric test must be applied. These tests are used if the underlying data do not meet the assumptions about the population sample or population sample size is too small or if the analyzed data is ordinal or nominal.

```
Types of non-parametric tests:
1. Mann Whitney U test
2. the sign test
3. the Wilcoxon signed-rank test
4. Kruskal Wallis test
```

# 1. Mann Whitney U test

The Mann-Whitney U Test is a nonparametric version of the independent samples t-test. The test primarily deals with two independent samples that contain ordinal data. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). If your data is following non-normal distribution, then you must go for Mann whitney U test instead of independent t test. It depends on what kind of hypothesis you want to test. If you want to test the mean difference, then use the t-test; if you want to test stochastic equivalence, then use the U-test.

Examples:

- to check the effectiveness of advertising for two competitive brands of clothes
- to investigate the effectiveness of a new drug to reduce symptoms of chicken pox in children.

# 2. Wilcoxon Signed Rank test

The Wilcoxon Signed Rank Test is a type of nonparametric test which in actual is counterpart of the paired samples t-test. The test compares two dependent samples with ordinal data. The Wilcoxon signed rank test should be used if the differences between pairs of data are non-normally distributed. It is generally used as a non-parametric alternative to the one-sample t test or paired t test and for ordered (ranked) categorical variables without a numerical scale. In general, it is used to compare two related samples, matched samples, or to conduct a paired

difference test of repeated measurements on a single sample to assess whether their population mean ranks differ.

Examples:

- comparison of two sets of scores from same participants in class
- to check any change in scores from one time point to another

# 3. Kruskal Wallis test

The Kruskal-Wallis Test is a nonparametric alternative to the one-way ANOVA. It is used to compare more than two independent groups with ordinal data.The procedure is used to compare three or more groups on a dependent variable that is measured on at least an ordinal level. It is used to determine whether or not there is a statistically significant difference between the medians of three or more independent groups.

Examples:

- wether or not six drugs have different effect on abdomen pain
- wether or not five host plants have effect on feeding preference of insect

# Normality tests

A normality test is used to determine whether sample data has been drawn from a normally distributed population or not. The normality tests are actually supplementary to the graphical assessment of normality. Testing for normality is often a first step in analyzing your data. But, what does that mean? Normality refers to a specific statistical distribution called a normal distribution, or sometimes the Gaussian distribution or bell-shaped curve. The normal distribution is a symmetrical continuous distribution defined by the mean and standard deviation of the data. There are many types of normality tests but here we will discuss the two most commonly used in statistics.

```
Types of parametric tests:
1. Shapiro-Wilk test
2. Kolmogorov-Smirnov (K-S) test
```

# 1. Shapiro-Wilk test

The Shapiro-Wilk test is a statistical test used to check if a continuous variable follows a normal distribution. The null hypothesis (H0) states that the variable is normally distributed, and the alternative hypothesis (H1) states that the variable is NOT normally distributed. The null hypothesis for Shapiro-Wilk test is that your data is normal, and if the p-value of the test if less than 0.05, then you reject the null hypothesis at 5% significance and conclude that your data is

non-normal. In order to be considered a normal distribution, a data set (when graphed) must follow a bell-shaped symmetrical curve centered around the mean. The Shapiro–Wilk test is more appropriate method for small sample sizes (<50 samples) although it can also be handling on larger sample sizes.

# 2. Kolmogorov-Smirnov (K-S) test

In statistics, the Kolmogorov–Smirnov test is a non-parametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples. In other words, the Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a normal distribution. The Kolmogorov Smirnov test produces test statistics that are used (along with a degrees of freedom parameter) to test for normality. Kolmogorov–Smirnov test is used for sample size ($n \geq 50$). [$P > 0.05$] is the probability that the null hypothesis is true. 1 minus the P value is the probability that the alternative hypothesis is true. A statistically significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected. A P value greater than 0.05 means that no effect was observed.

# Homogeneity tests

This test determines if two or more populations (or subgroups of a population) have the same distribution of a single categorical variable. A test of homogeneity compares the proportions of responses from two or more populations with regards to a dichotomous variable (e. g., male/female, yes/no) or variable with more than two outcome categories.

- Levene's test: In statistics, Levene's test is used to assess the equality of variances for a variable calculated for two or more groups. Some common statistical procedures assume that variances of the populations from which different samples are drawn are equal. It is used to check that variances are equal for all samples when your data comes from a non normal distribution. You can use Levene's test to check the assumption of equal variances before running a test like One-Way ANOVA.

# Correlation Tests

Correlation test is used to evaluate the association between two or more variables. For instance, if we are interested to know whether there is a relationship between the heights of fathers and sons, a correlation coefficient can be calculated to answer this question.It determines how closely two variables fluctuate.

## Pearson correlation coefficient

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

- If the Pearson correlation coefficient (r) is between 0 and 1 then it is positive correlation (if one variable will change then other one will also change) e.g., insect length and weight (the longer the insect, the heavier their weight).
- If Pearson correlation coefficient (r) is 0 then there is no correlation and there will be no relationship between variables e.g., the car price and width of surface mats (the price of car is not related to the width of inside surface mats).
- If Pearson correlation coefficient (r) is between 0 and -1 then it is negative correlation (if one variable will change then other one will also change but in opposite direction) e.g., elevation and air pressure (the higher the elevation, the lower the air pressure).

## Spearman correlation coefficient

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity (A monotonic function is one that either never increases or never decreases as its independent variable changes) of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function. It takes values between +1 to -1 where

- a value of +1 means a perfect positive association of rank
- a value of 0 means that there is no association between ranks
- a value of -1 means a perfect negative association of rank.

Example: relation of scores of 10 students in Maths and Science in a class.

# Regression

The parameter (the regression coefficient) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x. In this way it represents the degree to which the line slopes upwards or downwards. In other words, it is a measure of the closeness of association of the points in a scatter plot to a linear regression line based on those points. The relationship can be represented by a simple equation called the regression equation. In this context "regression" simply means that the average value of y is a "function" of x, that is, it changes with x.

## Kendall's Tau Regression

Kendall's tau correlation coefficient (Kendall's tau-b, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. The Kendall Tau, or Kendall rank correlation coefficient, measures the monotony of the slope. Its values varies between -1 and 1; it is positive when the trend

increases and negative when the trend decreases. The trend is statistically significant when the p-value is less than 0.05.