What are the different Statistical Tests ?

Normality Tests

- Shapiro-Wilk Test
- D'Agostino's K^2 Test
- Anderson-Darling Test

Correlation Tests

- Pearson's Correlation Coefficient
- Spearman's Rank Correlation
- Kendall's Rank Correlation
- Chi-Squared Test
- Fisher's Exact Test

Stationary Tests

- Augmented Dickey-Fuller
- Kwiatkowski-Phillips-Schmidt-Shin

Parametric Statistical Hypothesis Tests

- Student's t-test(unpaired/independent)
- Welch Test
- Paired Student's t-test
- Analysis of Variance Test (ANOVA)
- Repeated Measures ANOVA Test

Nonparametric Statistical Hypothesis Tests

- Mann-Whitney U Test
- Wilcoxon Signed-Rank Test
- Kruskal-Wallis H Test
- Friedman Test

Assumptions For Normality Tests

Shapiro-Wilk Test: Tests whether a data sample has a Gaussian distribution.

Assumption

- Observations in each sample are independent and distributed identically.

Hypothesis

- H0: the sample has a Gaussian distribution.
- H1: the sample does not have a Gaussian distribution.

D'Agostino's K^2 Test :Tests whether a data sample has a Gaussian distribution.

Assumption

- Observations in each sample are independent and distributed identically.

Hypothesis

- H0: the sample has a Gaussian distribution.
- H1: the sample does not have a Gaussian distribution.

####**Anderson-Darling Test**: Tests whether a data sample has a Gaussian distribution.

Assumption

- Observations in each sample are independent and distributed identically.

Hypothesis

- H0: the sample has a Gaussian distribution.
- H1: the sample does not have a Gaussian distribution.

Assumptions For Correlation Tests

Correlation Tests: are used to check the correlation between two independent features or variables.

Pearson's Correlation Coefficient: Tests whether a data sample is linearly separable.

Assumption

- Observations in each sample are independent and distributed identically.
- Observations are normally distributed.
- Similar variance between independent variables

Hypothesis

- H0: the samples are correlated.
- H1: the sample does not have any correlation.

Spearman Correlation Test: Tests whether a data sample is montonically separable.

Assumption

- Observations in each sample are independent and distributed identically.
- Observations in each sample are ranked .

Hypothesis

- H0: the samples are correlated.
- H1: the sample does not have any correlation.

Kendall's Rank Correlation: Tests whether a data sample is montonically separable.

Assumption

- Observations in each sample are independent and distributed identically.
- Observations in each sample are ranked .

Hypothesis

- H0: the samples are correlated.
- H1: the sample does not have any correlation.

Chi-Squared Test: Tests whether two categorical variables are related to each other.

Assumption

- Observations in used in contengency table are Independent.
- There are more than 25 examples in contengency table.

Hypothesis

- H0: the samples are correlated.
- H1: the sample does not have any correlation.

STEPS

- Build a contingency table (crosstab) of frequencies across all combinations of categories
- If frequencies are above 5:
  - Pearson's Chi-Square Test for goodness of fit
- Else:
  - Fisher's Exact Test

Assumptions For Parametric Tests

Students t-test: Average between two data samples are significantly different.

Assumption

- Each data sample's observation are independent and distributed.
- Observations are normally distributed.
- Observations have same variance between each other.

Hypothesis

- H0: the mean between two samples are equal .
- H1: the mean between two samples are not equal.

Paired Student's t-test: Average between two data samples are significantly different.

Assumption

- Each data sample's observation are independent and distributed.
- Observations are normally distributed.
- Observations have same variance between each other.
- Observations are paired.

Hypothesis

- H0: the mean between two samples are equal .
- H1: the mean between two samples are not equal.

Welch Test- The Welch t-test assumes the following characteristics about the data:

Assumptions:

- Independence of the observations. Each subject should belong to only one group.
- No significant outliers in the two groups
- Normality. the data for each group should be approximately normally distributed.

Analysis of Variance Test (ANOVA): Average between two data samples are significantly independent and different.

Assumption

- Each data sample's observation are independent and distributed.
- Observations are normally distributed.
- Observations have same variance between each other.

Hypothesis

- H0: the mean between two samples are equal .
- H1: the mean between two samples are not equal.

Repeated Measures ANOVA Test: Average between two or more paired samples are significantly different.

Assumption

- Each data sample's observation are independent and distributed.
- Observations are normally distributed.
- Observations have same variance between each other.
- Observation can be paired.

Hypothesis

- H0: the mean between two samples are equal .
- H1: the mean between two samples are not equal.

Assumptions For Non-Parametric Tests

Mann-Whitney U Test: Distribution of two data samples are equal or not.

Assumption

- Each data sample's observation are independent and distributed.
- Observations in each data samples can be ranked.

Hypothesis

- H0: the distribution of two samples are equal .
- H1: the distribution of two samples are not equal.

Wilcoxon Signed-Rank Test: Distribution between two paired samples are significantly equal or not.

Assumption

- Each data sample's observation are independent and distributed.
- Observations can be ranked.
- Observations are paired.

Hypothesis

- H0: the distribution of two samples are equal .
- H1: the distribution of two samples are not equal.

Kruskal-Wallis H Test: Distribution between two independent samples are significantly equal or not.

Assumption

- Each data sample's observation are independent and distributed.
- Observations can be ranked.

Hypothesis

- H0: the distribution of samples are equal .
- H1: the distribution of samples are not equal.

Friedman Test: Distribution between two paired samples are significantly equal or not.

Assumption

- Each data sample's observation are independent and distributed.
- Observations can be ranked.
- Observations can be paired.

Hypothesis

- H0: the distribution of all samples are equal .
- H1: the distribution of one or more samples are not equal.

When to apply Statistical Test?

- **Start by checking the data types of the columns**
- **Decide whether to apply parametric or non-parametric tests**

.Quantitative variables can be further classified into:

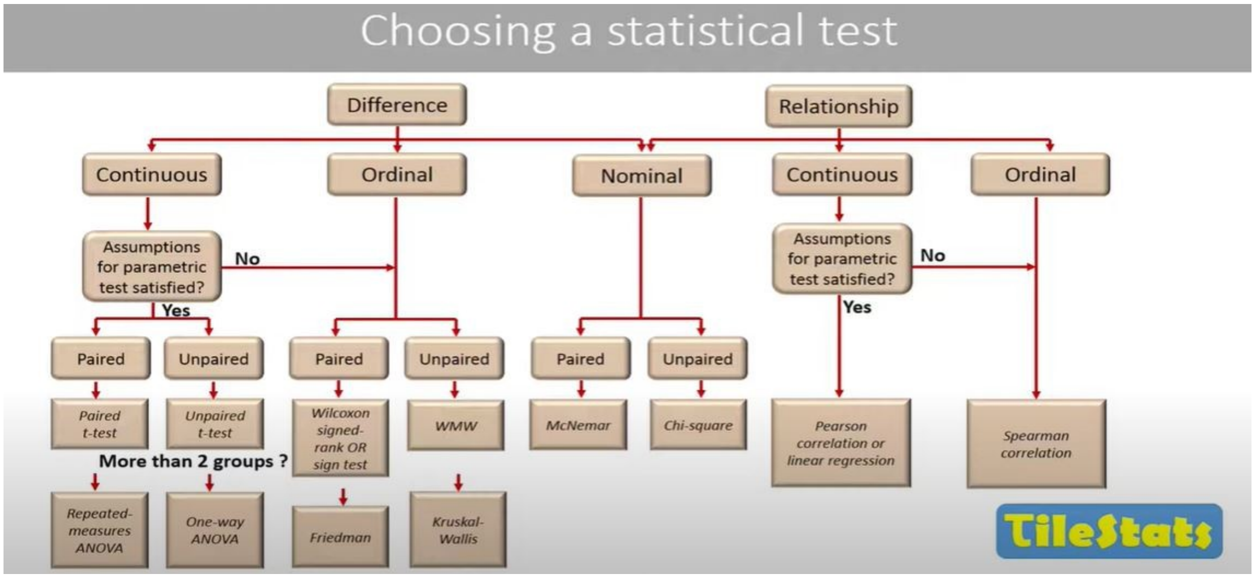
- ***Continuous variables:*** Also known as ratio variables. These are units of measurement that can be represented in quantities less than 1. For instance, 0.8 km.
- ***Discrete or integer variables:*** Units that cannot be divided, such as 1 car or 1 tree.
- **Interval:**Variables that do not have a true zero (or it's arbitrary), or that zero value of that variable doesn't make sense. For example - IQ. Zero IQ doesn't mean anything. At the same time, person with IQ of 100 does not mean being twice as sharp to person with IQ of 50. These variables have equal distance between values.
- **Ratio/Count:**Variables where zero value makes sense. For example - number of apples, 0 apples means nothing, while having 10 apples mean having twice as many as having 5 apples. These variables have an intrinsic order with equal distance between values, howsoever small.

Categorical variables are divided into the following variables:

- ***Binary:*** Data that has one of two outcomes, such as yes/no or pass/fail.
- ***Nominal***: Used to describe data with no intrinsic order, such as brands, families, species.
- **Ordinal**: *italicized text* For ordered data types with observable scales of hierarchy, such as user ratings.

Informal Rules:

- Ordinal: Nominal + Order
- Interval: Order + Equal distance between values
- Ratio: Interval + True Zero



Choosing a statistical test

Continous Variables: Whether Parametric tests or Non-Parametric tests?

Assumptions for Parametric Tests:

- **Independence of observations:** The separate observations (each variable entry) are unrelated to one another (for instance, repeating the same test on a single patient generates non-independent measurements, that is, repeated measurements).
- **Normality of data:** The data follows a normal distribution. This assumption is required only for quantitative data. (For more details, see also here)
- **Homogeneity of variance:** Each group being compared has a similar variance (i.e., the distribution, or “spread,” of scores around the mean). The “power” of the test to detect differences will be diminished if one group has significantly greater variation than the others.

####Normality Tests- validate the Gaussian distribution of data.

Tests of equality of distributions:

- Shapiro-Wilk Test for normality
- Kolmogorov-Smirnov Test (KS-Test) of goodness of fit
- Q-Q plots

Please refer to section 1 to check which tests can be used to check the three conditions: Normality, Homogeneity and Independence

How to apply the statistical tests

Here are a few tests for various data types:

1. Categorical(Input) v/s Numerical(Output):

- Check for Homogeneity of variances: F-Test (Two groups) / Levene Test (Two or more groups)
  - Two categories:
    - If samples coming from populations with similar variances:
      - Pooled T-Test
    - Else:
      - Welch T-Test
  - More than two categories:
    - If samples coming from populations with similar variances:
      - One-Way ANOVA (assumes that variances are equal)
    - Else:
      - Welch ANOVA
      - Kruskal-Wallis Test: Tests whether medians are equal
      - One-way ANOVA:
        - If max_variance / min_variance < 4 (thumb rule)
        - Sample sizes are equal

2. Categorical(Input) v/s Categorical(Output):

- Build a contingency table (crosstab) of frequencies across all combinations of categories
  - If frequencies are above 5:
    - Pearson's Chi-Square Test for goodness of fit
  - Else:
    - Fisher's Exact Test

3. Numerical(Input) v/s Numerical(Output):

- Pearson Correlation
- R2, adjusted R2
- Hypothesis Testing:
- Types:
- Test of equality of proportions:
  - Z-Test
- Test of equality of means:
  - If standard deviation is known:
    - Z-Test
  - If standard deviation is unknown:
    - If samples coming from populations with similar variances:
      -
        1. Pooled T-Test

        - Else:
            - –
                a. Welch T-Test
- Test of equality of variances:
    - F-Test

4. Tests of equality of distributions:

- Kolmogorov-Smirnov Test (KS-Test) of goodness of fit
- Q-Q plots
- Shapiro-Wilk Test for normality

5. Other tests:

- Paired t-test
- Wilcoxon signed-rank test
- Sign test