Inferential Statistics

Statistics is all about the data. Data alone is not interesting. It is the interpretation of the data that we are interested in. Using Hypothesis Testing, we try to interpret or draw conclusions about the population using sample data. A Hypothesis Test evaluates two mutually exclusive statements about a population to determine which statements about a population to determine which statement is best supported by the sample data. Whenever we want to make claims about the distribution of the data or whether one set of results are different from another set of results in applied machine learning, we must rely on statistical hypothesis tests.

There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If result is contrary to the hypothesis, then you've made a discovery.

1) **What is inferential Statistics?**
   The branch of statistics that analyzes sample data to draw conclusion about a population.
2) **What is Hypothesis Testing?**
   A statistical technique to test some hypothesis about the parent population from which the sample is actually drawn.
3) **Null Hypothesis:** The Statement that a population parameter is equal to a specific value, or that the population parameters from two or more groups are equal.
4) **Alternative Hypothesis:** The Statement paired with a null hypothesis that is mutually exclusive to the null hypothesis.
5) **Critical area/Region:** The critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.
6) **One-Tailed Test:** A one tailed test is a statistical hypothesis test in which the critical region of a distribution is one sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.
7) **Two-tail Test :** A two tailed test is a method in which the critical area of a distribution is two sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested fall into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.
8) **Type I Error:** The error that occurs if the null hypothesis is rejected when in fact it is true and should not be rejected.
9) **Type II Error:** The error that occurs if the null hypothesis is not rejected when in fact it is false and should be rejected.
10) **Level of Significance:** The probability of making a Type-I error and it is denoted by **alpha.** Alpha is the maximum probability that we have a Type-I error. For 95% confidence level, the value of alpha is 0.05. This means that there is a 5% probability that we will reject a true null hypothesis.

11) **P-value:** The probability of obtaining a test statistic equal to or more extreme than the result obtained from the sample data, given that the null hypothesis H_0 is true.
12) **High P value:** your data are likely with a true null.
13) **Low p value:** Your data are unlikely with a true null.
14) **Steps involved in Hypothesis Testing**
    a. **Formulate the Hypothesis Testing**
    b. **Select an appropriate Test**
    c. **Choose level of significance**
    d. **Collect data and calculate test statistics**
    e. **Determine the probability (or critical value)**
15) **Test Statistic:** The test statistic measures how close the sample has come to the null hypothesis. It's observed value changes randomly from one random sample to a different sample. A test statistic contains information about the data that is relevant for deciding whether to reject the null hypothesis or not.
16) **Parametric Test:**
    a. **Z-test**
    b. **T-test**
    c. **Chi-square test**
    d. **F-test**
    e. **ANOVA**
17) **What is Z-test?**
    Z-test is a statistical test to determine whether two population means are different when the variance are known and sample size is large.
18) **What is t-test?**
    A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups which may be related in certain features.
19) **What is chi-square test?**
    Chi-square test in hypothesis testing is used to test the hypothesis about the distribution of observations in different categories.
20) **What is F-test?**
    F-test is used to test if the variances of two populations are equal. This test can be two tailed or one tailed test.
21) **What is ANOVA?**
    ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

22) Non Parametric Test:

a. **Sign Test**
b. **Wilcoxon Signed Rank test**
c. **Mann-Whitney Wilcoxon test**
d. **Kruskal Wallis Test**

23) What is Sign Test?

The sign test is a non-parametric test that is used to test whether or not two groups are equally sized. The sign test is used when dependent samples are ordered in pairs, where the bivariate random variables are mutually independent it is based on the direction of the plus and minus sign of the observation, and not their numerical magnitude.

24) What is Wilcoxon Signed Rank Test?

The Wilcoxon signed Rank test is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ.

25) What is Mann-Whitney Wilcoxon Test? A.K.A Wilcoxon Rank-sum Test

The Mann-Whitney U test is a non-parametric test that allows two groups or conditions or treatments to be compared without making the assumption that values are normally distributed.

26) What is Kruskal Wallis Test?

The Kruskal-Wallis test is a non-parametric test, and is used when the assumptions of one-way ANOVA are not met.

27) What is Shapiro Wilk test?

Shapiro Wilk test is used to test whether the data is normally distributed or not.

P value should be greater than alpha 0.05, if it is not then there is non-normality in the data.

```
data("trees")
shapiro.test(trees$Height)
barplot(trees$Height)
lapply(trees,shapiro.test)
```