

**Statistics Assignment 5**

1. How are you going to figure out the average heights of all the trees in Karnataka?

**Answer:**

It is highly not practical to take every trees height in Karnataka state economically not affordable as well time consuming.

Better we use sampling technique to find this information, below are four major sampling techniques.

* Simple Random Sampling
* Systematic Sampling
* Stratified Sampling
* Cluster Sampling

To figure out the average height of all the trees in Karnataka systematic sampling technique finds more suitable. As per my assumption botanically every trees grows up to specific height limit. So in sampling better if we create groups for each trees botanical category and then select regional area i.e. district to choose samples of trees in sufficient numbers.

1. What is hypothesis testing and how does it work?

**Answer:** This is statistical tool to use for testing experiments and find its result appropriate and under favor in a given random data sample. It is an educated guess about specific parameter or population, once it is defined one can collect random sample data to determine whether it provides enough evidence that educated guess true or not. In simple word it is a claim, until the claim is proven to be true it is call the hypothesis, once the claim is proved it become new truth.

In hypothesis testing two important terminology used

Null Hypothesis (H0) – This is a statistical theory that suggest there is no statistical significance exists between the population parameter. H0 always maintain status quo, must always contain equality (=)

Alternative Hypothesis (Ha) – This is statistical theory that suggest there is statistical significance exist between the population parameter. Ha always contains difference (≠,>, ˂)

Below are the step required while doing hypothesis test-

1. The hypothesis could either be the statement that is assumed to be true or the claim which is made to be true.
2. Identify and set the terminology Null Hypothesis and Alternate Hypothesis
3. Identify level of alpha (significance level)
4. Identify test statistics that could be used to assess the Null Hypothesis i.e. T-test, Z-Test, F-Test etc.
5. Compute the Test statistics and find score
6. Find P-value with Test statistics score
7. Compare P-value with significant level and decide if to accept or reject H0.
8. Explain the differences between Alpha and Beta errors. Which inaccuracy is the most hazardous?

**Answer:**

Sometimes statistical test yield erroneous outcome. Specifically two errors may occur in hypothesis test.

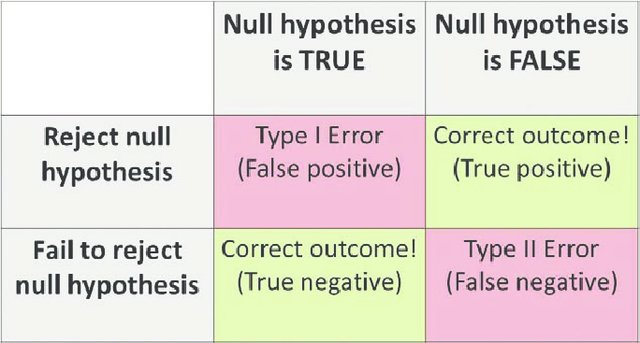
Alpha Error – This error occur when null hypothesis rejected but in actual it is true. This is also known as Type I error of FALSE POSITIVE

Beta Error - This error occur when null hypothesis retained but in actual it is False. This is also known as Type II error of FALSE NEGATIVE

Based on your data and result of statistical test you defined if the result is statistically significant.

* If your results show statistical significance, that means they are very unlikely to occur if the null hypothesis is true. In this case, you would reject your null hypothesis. But sometimes, this may actually be a Type I error.
* If your findings do not show statistical significance, they have a high chance of occurring if the null hypothesis is true. Therefore, you fail to reject your null hypothesis. But sometimes, this may be a Type II error.

The above classification can form below matrix known as “Confusion Matrix”



It is very difficult to determine which error is more hazardous, it is completely depend upon your experiments and status quo you defined for H0 & Ha.

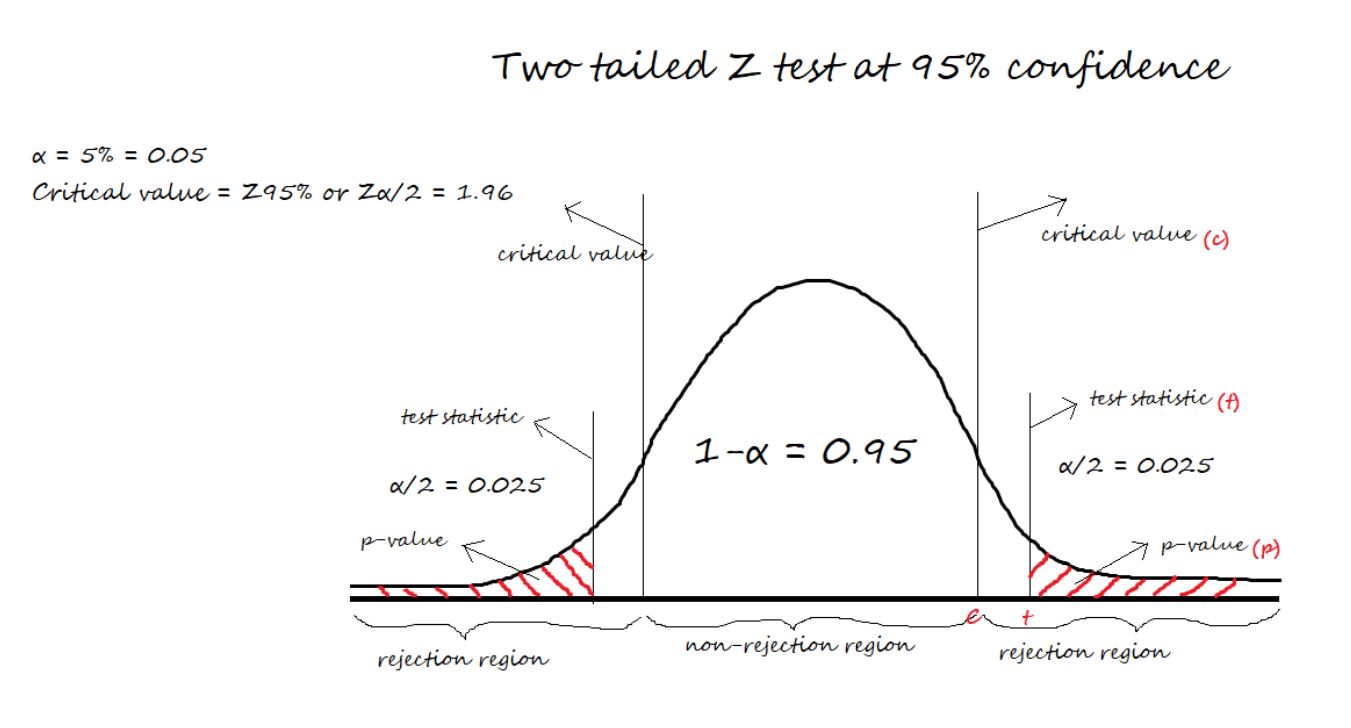
4. What is the significance of the p-value?

**Answer:**

In statistics, p-value is a statistical measure used to validate a hypothesis test against observed data. Lower the p-value, greater the statistical significance of the observed difference and stronger the evidence to reject the null hypothesis. A P-value of 0.05 or lower is generally considered statistically significant.

When you run the hypothesis test, you compare the p-value from your test to the alpha level (significance level) you selected you ran the test.

Lets understand from below diagram



As denoted in above diagram the area under red mark that is both side tail area of probability distribution. If any value which falls in this region which will be treated as less than p-value i.e. α<0.025 shall be consider to reject the null hypothesis. Therefore P-value is significant value to determine if our result falls within an acceptable region to accept or reject. In commercial words this value mostly determined by industrial experts and minimum value generally consider 5% i.e. 0.05

5. What is the Probability Distribution Function and how does it work?

**Answer:**

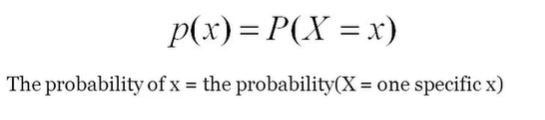
The Probability distribution is a statistical function which describe how the possible outcome values from random experiment distributed.

Suppose you draw a random sample and measure the heights of the subjects. As you measure heights, you create a distribution of heights. This type of distribution is useful when you need to know which outcomes are most likely, the spread of potential values, and the likelihood of different results.

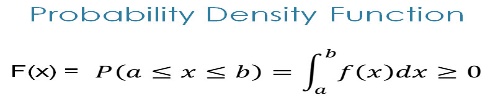
Two important functions that are used to describe a probability distribution.

* Probability Density Function or probability mass function (PDF or PMF)
* Cumulative Distribution Function (CDF).

For a discrete random variable, x, the probability distribution is defined by a Probability Mass Function, denoted by PMF.



For a continuous random variable, x, the probability distribution is defined by a Probability Density Function, denoted by PDF.



PDF is f(x) which describe the shape of the distribution.

The probability distribution function of a random variable always lies between 0 and 1.