What is Statistics?

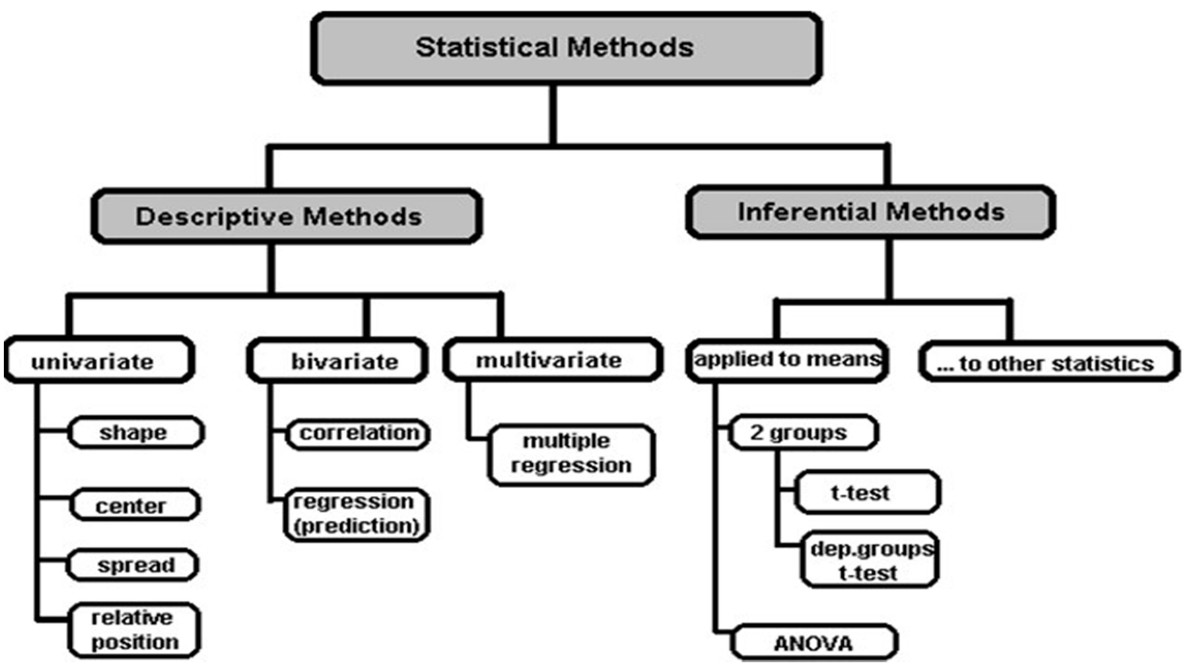
Statistics is neither really a science nor a branch of Mathematics. It is perhaps best considered as a meta-science (or meta-language) for dealing with data collection, analysis, and interpretation. As such its scope is enormous and it provides much guiding insight in many branches of science, business.

- *Statistics* is the science of collecting, organizing, analyzing, interpreting, and presenting data. (*Old Definition*).
- A *statistic* is a single measure (number) used to summarize a sample data set. For example, the average height of students in this class.
- A *statistician* is an expert with at least a master's degree in mathematics or statistics or a trained professional in a related field.

Statistics is a *tool* for creating *new understanding* from a set of numbers.

Statistics is a science of getting informed decisions.

A Taxonomy of Statistics



Statistical Inference: is to draw conclusions about the Population on the basis of information available in the sample which has been drawn from the population by a random sampling technique/ procedure. There are two branches Statistical Inference namely ESTIMATION & TESTING OF HYPOTHESIS.

In ESTIMATION, we try to find an estimate of any population characteristic while in TESTING OF HYPOTHESIS, we try to test the statement about any population characteristic. *Here, our main concern is with "TESTING OF HYPOTHESIS".*

Some Basic Definitions:

Population: Any collection of individuals under study is said to be Population (Universe). The individuals are often called the members or the units of the population may be may be physical objects or measurements expressed numerically or otherwise.

Sample: A part or small section selected from the population is called a sample and process of such selection is called sampling.

(The fundamental object of sampling is to get as much information as possible of the whole universe by examining only a part of it. An attempt is thus made through sampling to give the maximum information about parent universe with the minimum effort).

Parameters: Statistical measurements such as Mean, Variance etc. of the population are called *parameters.*

Statistic: It a statistical measure computed from sample observations alone. The theoretical distribution of a statistics is called its sampling distribution. *Standard deviation of the sampling distribution of a statistic is called Standard Error.*

Hypothesis: is a statement given by an individual. Usually it is required to make decisions about populations on the basis of sample information. Such decisions are called *Statistical Decisions*. In attempting to reach decisions it is often necessary to make assumption about population involved. Such assumptions, which are not necessarily true, are called *statistical hypothesis*.

Parametric Hypothesis: A statistical hypothesis which refers only to values of unknown parameters of population is usually called *a parametric hypothesis.*

Null Hypothesis and Alternative Hypothesis: A hypothesis which is tested under the assumption that it is true is called a *null hypothesis* and is denoted by H_0. Thus a hypothesis which is tested for possible rejection under the assumption that it is true is known as *Null Hypothesis*. The hypothesis which differs from the given Null Hypothesis H_0 and is accepted when H_0 is rejected is called an *alternative hypothesis and is denoted by* H_1 *(The hypothesis against which we test the null hypothesis, is an alternative hypothesis)*.

Simple and Composite Hypothesis: A parametric hypothesis which describes a distribution completely is called a *simple hypothesis* otherwise it is called *composite hypothesis*. For example; In case of Normal Distribution N (μ, σ^2), $\mu = 5$, $\sigma = 3$ is simple hypothesis whereas $\mu = 5$ is a composite hypothesis as nothing have been said about σ.
Similarly, $\mu < 5$, $\sigma = 3$ is a composite hypothesis.
Let H_0: $\mu = 5$ be the null hypothesis, then

H_1: $\mu \neq 5$ is two sided composite alternative hypothesis.

H_1: $\mu < 5$ is one sided (Left) composite alternative hypothesis.

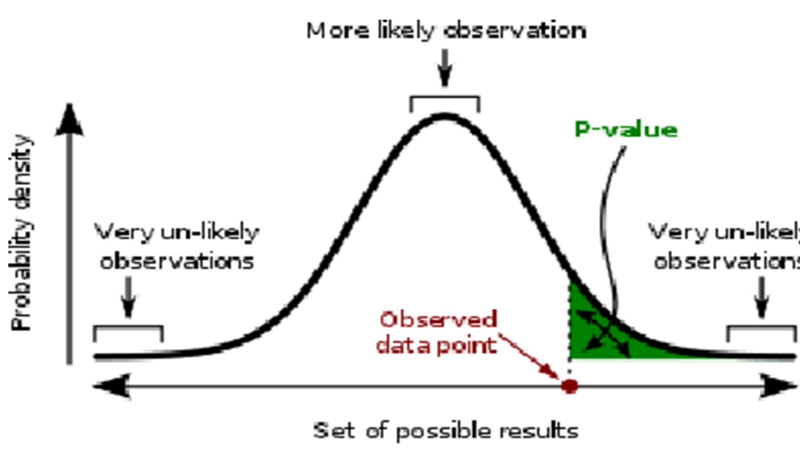
H_1: $\mu > 5$ is one sided (Right) composite alternative hypothesis.

Test: *Test is a rule through we test the null hypothesis against the given alternative hypothesis.*

Tests of Significance: Procedure which enables us to decide, on the basis of sample information whether to accept or reject the hypothesis or to determine whether observed sampling results differ significantly from expected results are called *tests of significance, rules of decisions or tests of hypothesis.*

Level of Significance: The probability level below which we reject the hypothesis is called *level of significance.* The levels of significance usually employed in testing of hypothesis are 5% and 1%.

P-Value: The **p-value** is the level of marginal significance within a **statistical** hypothesis test representing the probability of the occurrence of a given event. The **p-value** is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence **against** a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That's pretty tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and *not* due to anything in your experiment. Therefore, the smaller the p-value, the more important ("significant") your results.

Critical Region and Acceptance Region: A region (corresponding to a statistic t) is called the sample space. The part of sample space which amounts to rejection of null hypothesis H_0, is called *critical region or region of rejection*.

If $X = (x_1, x_2, \dots, x_n)$ is the random vector observed and W_c is the critical region (which corresponds the rejection of the hypothesis according to a prescribed test procedure) of the sample space W, then $W_a = W - W_c$ of the sample space is called the *acceptance region*.

Two Types of Errors in Testing of a Hypothesis: While testing a hypothesis H_0, the following four situations may arise:

(a) The test statistic may fall in the critical region even if H_0 is true then we shall be led to reject H_0 when it is true.

(b) The test statistic may fall in the acceptance region when H_0 is true we shall be led to accept H_0.

(c) The test statistic may fall in the critical region when H_0 is not true i.e. H_1 is true, then we shall be led to reject H_0 when H_1 is true.

(d) The test statistic may fall in the acceptance region even if H_0 is not true, then we shall be led to accept H_0 when it is not true.

It is quite obvious that the decisions taken in (b) and (c) are correct while the decisions taken in (a) and (d) are incorrect.

The wrong decision of rejecting a null hypothesis H_0, when it is true is called the ***Type I Error*** i.e. we reject H_0 when it is true. Similarly, the wrong decision of accepting the null hypothesis H_0 when it is not true is called the ***Type II Error*** i.e. we accept H_0 when H_1 is true.

Probability Forms:

$$P(reject\ H_0\ when\ it\ is\ true) = P(reject\ \ H_0/H_0) = \alpha$$

and

$$P(accept\ H_0\ when\ it\ is\ wrong) = P(reject\ \ H_0/H_1) = \beta$$

The α and β are called the size of Type I Error and size of Type II Error respectively.

Rules or Procedure for Testing of Hypothesis:
A test is a statistical procedure or a rule for deciding whether to accept or reject the hypothesis on the basis of sample values obtained.
Following is the Procedure for testing of Hypothesis:-

(a) Mention the null hypothesis H_0 to be tested along with an alternative hypothesis H_1.
(b) Make some assumptions such as the sample is random, the population is normal, the variance of two different populations are equal or unknown etc.
(c) Then find the most *appropriate test statistic* together with its sampling distribution. A statistic whose primary role is that of providing a test of some hypothesis is called a test statistic.
(d) On the basis of the sampling distribution make a decision to either accept or reject the null hypothesis H_0.
(e) Take a random sample and compute the test statistic. If the calculated value of the test statistic falls in the acceptance region, then accept the null hypothesis H_0. If it falls in the region of rejection (or Critical Region), reject the null hypothesis H_0.

Power Function of a Test: The Power Function of a test of a statistical hypothesis H_0: $\theta = \theta_0$, say, against alternative hypothesis H_1: $\theta > \theta_0$, $\theta < \theta_0$, $\theta \neq \theta_0$ is a function of the parameter, under consideration, which gives the probability that the test statistic will fall in the critical region when θ is the true value of the parameter.

Deduction: $P(\theta) = P$ (rejecting H_0 when H_1 is true)
$= P$ (W belong to W_c / H_1) $= 1 - P$ (accept H_0/ H_1)
$= 1 - \beta(\theta)$

The value of the power function at a particular value of the parameter is called the power of the test

Best Critical Region: In testing the hypothesis H_0: $\theta = \theta_0$ against the given alternative H_1: $\theta = \theta_1$, the critical region is best if the type II error is minimum or the power is maximum when compared to every other possible critical region of size α. *A test defined by this critical region is called* ***most powerful test.***