Sampling

An engineering student working part-time at a mobile store is given a task to find out whether the college population prefers Blackberry phones over Android phones. Since she has access to the potential customers i.e. her fellow college mates, she prepares a survey to study their preferences. The number of students in her college is 2431. Obviously, it is not possible or her to send the survey to all of them. She has about 270 students from her college in her Facebook friends list. She sends out a mass message them requesting them to fill out her questionnaire. Only 142 of them respond and actually take the survey.

- What is the population in the study? (What group she wants information about?)
- What is the sample? (From what group she actually obtains information?)

Very often we would like to find out some characteristics or qualities of a large group (which we call **population**). For example, we would like to find out (i) how the Mechanical engineering students of the Mumbai University are doing in Mathematics in the fourth semester or (ii) the proportion of defective items in the machinery parts sent by a contractor.

A cosmetic company would like to find out how well its present perfume product is liked by consumers and whether it has to change the fragrance of the perfume.

The Mumbai municipal corporation would like to find out whether building a new foot over bridge along the eastern express highway near Chheda Nagar will benefit pedestrians.

In the above cases, since the **population** is very large, we would like to draw conclusions based on a **part** of the population (which we call **sample**).

Such instance of drawing conclusions about some characteristic of the population based on a sample is known as **sampling**.

Sampling is used largely by governments and industries in cases where it is not possible to study the entire population due to time, cost and manpower constraints.

Parameter: Population measures like mean and variance are known as parameters and are denoted by μ and σ^2

Statistic: Sample measures like mean and variance are known as statistic and are denoted by $\bar{x}$ and s^2

In sampling, some statements, known as **hypotheses**, are made about a population parameter. For example, the statement might be that the expected value of the height of ten year old boys in Mumbai is not different from that of ten year old girls. A hypothesis might also be a statement about the distributional form of a characteristic of interest, for example that the height of ten year old boys is normally distributed within the Mumbai population.
Before we detail the different tests, we define the various terms and concepts used.

Null Hypothesis: - A definite statement about a population parameter which is tested for possible rejection under the assumption that it is true. It is usually a hypothesis of **no difference** and is denoted by H_o.

In other words, the null hypothesis, H_o, represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on an average, than the current drug. We would write

H_o: there is no difference between the two drugs on average.

Alternative hypothesis: - Any hypothesis that is complementary to the null hypothesis is called an alternative hypothesis and is denoted by H_1. For instance, if

H_o: $\mu = \mu_o$, then H_1 can be $\mu \neq \mu_o$ or $\mu < \mu_o$ or $\mu > \mu_o$

We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if or when the null hypothesis is rejected.

Check out this page and their brilliant explanation of a Null Hypothesis and Alternative Hypothesis with an example about how and alien is stealing their socks.
http://www.null-hypothesis.co.uk/science/item/what_is_a_null_hypothesis/

Type I error & Type II errors

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected.

A type II error occurs when the null hypothesis is accepted when it is in fact false; that is, H_0 is wrongly accepted.

That is, Type I error is rejecting the null hypothesis H_o when it is true and Type II error is accepting H_o when it is false.
Now if P (rejecting H_o when it is true)
= P (rejecting H_o/H_o) = α
Then α is called the **size** of Type I error.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.

H_o: there is no difference between the two drugs on average.

A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

Similarly, if

P (accepting H_o when it is false)

= P (accepting H_o/H_1) = β

then β is called the **size** of Type II error.

α & β are usually referred to as **Producer's risk** and **consumer's risk** respectively.

These terms can be understood very easily if you imagine yourself the manufacturer of a certain type of goods, let's say mobile phones. Suppose your company produces mobile phones and the packages go out to the retailers in boxes of 1000 phones. Let us say there are 500 such boxes. You want to check these boxes for the defective pieces in them. You will not have the time to check each and every one of 500,000 mobile phones. So you decide to take samples. We can say the Null hypothesis is that there are no defective phones. Naturally, the Alternative Hypothesis will be that there are defective pieces. Let's say in each box of 1000, you have employed executives to check 50 of the phones. If too many phones in those 50 are defective, you reject the entire box of 1000 phones, thinking that most of them will be defective. There are two kind of errors that could occur in such a situation.

Suppose you checked 50 phones in a box and found 35 of them to be defective, and decided that it's too large a number of bad phones and rejected the entire box. It may not be too probably, but it is possible, that over 900 phones in the box were good. As a producer, you will suffer a loss as the good phones were rejected. This is a **Type I error** or the **Producer's risk**. Here we have rejected the null hypothesis H_0 though it was true.

On the other hand, you could check 50 phones in a box and find only 5 of them defective. You decide that's an acceptable number and sanction the box forward to the retailer for selling. It could so happen that there are over 400 defective pieces in that box, they just weren't detected because the sample you chose had good phones. These phones will now go out in the market and a lot of consumers will buy defective phones. This is a **Type II error** or the **Consumer's risk**. Here we have accepted the null hypothesis H_0 though it is false. That is, we have accepted H_0 when H_1 is true.

		Decision	
		Reject H_0	Accept H_0
Truth	H_0	Type I Error	Right decision
	H_1	Right decision	Type II Error

A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. This probability of a type I error can be precisely computed as

P(type I error) = significance level = α

Critical Region: The region of rejection of the null hypothesis H_o is known as the critical region.

Critical value: The value (of the test statistic) that separates the region of rejection and acceptance is known as the critical value.

Level of Significance :- The probability (say α) that a random value of the statistic t belongs to the critical region (region of rejection of H_o) is known as the level of significance. L.O.S is the size of Type I error. The L.O.S is **always** fixed in advance before collecting the sample information. It is usually fixed at 5%. (Which means in a problem, if you are not given the l.o.s. explicitly, you assume it to be 5%.)

Confidence interval: The interval in which a population parameter is supposed to be with a given probability p, is said to be the 100xp% confidence interval for that parameter.

For eg: in large samples, 95% confidence limits (or interval) for the population mean is given by $\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}$, where $\bar{x}$ is the sample mean, σ, the population standard deviation & n, the size of the sample. This means that if you know the mean of the sample that is $\bar{x}$, you can say with 95% confidence or certainty that the mean of the population will lie in the interval

$$\left(\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right)$$