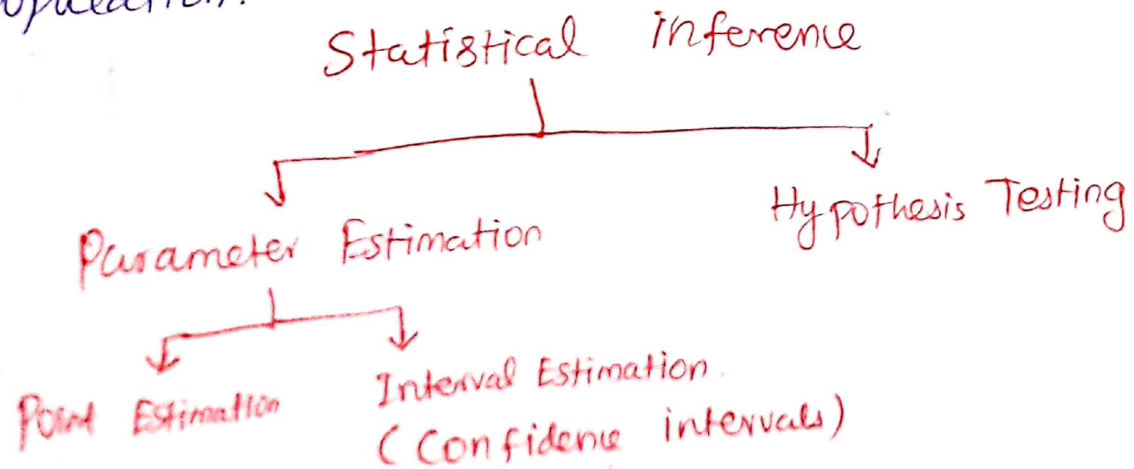
Statistical Inference:

Often in Practice we are interested in drawing valid Conclusions about a large group of individuals or Objects. Instead of examining the entire group, Called, the Population, which may be difficult or impossible to do, we may examine only a small part of this Population, which is called a Sample.

We do this with the aim of inferring Certain facts about the population from results found in the sample. This aspect of statistics is called Statistical inference.

The process of obtaining Samples is called Sampling.

Example: We may wish to draw Conclusions about the height of 10,000 Students (the population) by examining only 100 students (a Sample) selected randomly from this population.



Defn: A population consists of the totality of the observations with which we are concerned.

The number of Observations in the population is defined to be the Size of the population.

If there are 600 Students in the school whom we Classified according to blood type, we say that we have a Population of size 600.

Defn: A sample is a subset of a population.

It is desirable to choose a random sample in the sense that the observations are made independently and at random.

·Consider a random experiment. Let X be the r.v associated with this experiment. Let $f_X(x)$ be the PDF of X. Let us repeat this experiment 'n' times. Let X_k be the r.v associated with the k^{th} repetition. Then the collection of the r.vs $\{X_1, X_2, .. X_n\}$ is a random sample of size n. with numerical value $\{x_1, x_2, \ldots, x_n\}$ $[X_i$'s are (iids)$]$

Defn: Let $X_1, X_2, \ldots, X_n$ be n independent r.vs, each having the Same Probability distribution $f_X(x)$. Define $\{X_1, X_2 .. X_n\}$ to be a random sample of size n from the population $f_X(x)$. and write its joint PDF as $\boxed{f(x_1, x_2, \ldots, x_n) = f(x_1).f(x_2) \cdots f(x_n)}$

The primary purpose in taking a random sample is to Obtain information about the unknown population Parameters. Suppose, for example, that we wish to reach a Conclusion about the Proportion of people in India who prefer a particular brand of Soft drink. It would be impossible to question every indian in order to compute the value of the parameter 'p' representing the population proportion. Instead, a large random Sample is selected and the proportion $\hat{p}$ of people in this sample favouring the brand of soft drink in question is calculated. The value $\hat{p}$ is now used to make an inference concerning the true proportion p. Now $\hat{p}$ is a function of the observed values in the random sample. Because many random samples are possible from a population, the value of $\hat{p}$ will vary from sample to sample. That, is $\hat{p}$ is a r.v [Such a r.v is Statistic]

Defn: A statistic is any function of the r.vs constituting a random sample.

Some Examples of Statistic. $T = h(X_1, X_2, \cdots X_n)$

If $X_1, X_2, \ldots, X_n$ is a random sample of size 'n'.

(a) Sample mean :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

When X_1 assumes the value x_1, X_2 assumes the value x_2 and so forth. The term sample mean is applied to both the statistic $\bar{X}$ and its computed values $\bar{x}$

(b). Sample Median: [It is used for measuring the middle value of sample, arranged in order of magnitude]

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even.} \end{cases}$$

(c) Sample Mode: is the value of the sample that occurs most often.

Variability Measures of a Sample:

The variability in a sample displays how the observations spread out from the average. It is possible to have two sets of observations with the same mean that differ considerably in the variability of their measurements about the average. Consider the following measurements, in liters, for two Samples of orange juice bottled by Companies A & B:

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

Both Samples have the Same mean, 1.00 liter. It is Obvious that Company A bottles Orange juice with a more uniform Content than Company B.

We say that the variability, or the dispersion, Of the Observations from the average is Less for Sample A than for Sample B.

Therefore, in buying Orange juice, we would feel more Confident that the bottle we select will be Close to the advertised average if we buy from company A.

a) Sample Variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The computed value of S^2 for a given sample is denoted by s^2.

$$\boxed{s^2 = \frac{1}{\boxed{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Alternative Formula:

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]$$

If s^2 is small, there is relatively little variability in the data, but if s^2 is large, the variability is relatively large.

Sampe Standard deviation:

$\boxed{S = \sqrt{S^2}}$ where S^2 is the Sample Variance.

Samplo range:

$\boxed{R = X_{max} - X_{min}}$, where X_{max} denote the largest of the X_i values and X_{min} the Smallest.

Example: Find the variance of the data 3, 4, 5, 6, 6 and 7.

Soln: Here $n = 6$, $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$

Hence $$s^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]$$

$$= \frac{1}{(6)(5)}\left[(6)(171) - (31)^2\right] = \frac{13}{6}.$$

Thus, the sample standard deviation $s = \sqrt{\frac{13}{6}} = 1.47$

and the Sample range is $7 - 3 = 4$.

Defn: The Probability distribution of a Statistic is called a Sampling Distribution.

For example, the probability distribution of $\bar{X}$ is called the sampling distribution of the mean

Remark : " Degrees of freedom"

let us consider the Sample variance s^2 as being based on 'n-1' degrees of freedom. The term degrees of freedom results from the fact that the sum of the 'n' deviations $x_1-\bar{x},\ x_2-\bar{x},\ \dots,\ x_n-\bar{x}$ always to zero.

$$\left[\ \sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}x_i - n\bar{x} = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}x_i = 0 .\right.$$

$$\left.\text{Since } \bar{x} = \frac{\sum_{i=1}^{n}x_i}{n} \Rightarrow n\bar{x} = \sum_{i=1}^{n}x_i\ \right]$$

and so Specifying the Values of any n-1 of these quantities automatically determines the remaining One. We may think of the number of degrees of freedom as the number of independent pieces of information in the data..

Sampling Distribution of the Sample Mean $\bar{X}$:

Suppose that a random sample of size 'n' is taken from a normal population with mean μ & Variance σ^2.

Now each observation in this sample, say, $X_1, X_2, \ldots, X_n$, is a normally and independently distributed r.v with mean μ and variance σ^2.

Then because linear functions of Independent, normally distributed r.Vs are also normally distributed.

$\Rightarrow$ $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ has a normal distribution

with mean
$$\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$
$$= \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n}.$$
$$= \frac{\mu + \mu + \cdots + \mu}{n}$$
$$= \frac{n.\mu}{n} = \mu$$

Variance
$$\sigma^2_{\bar{X}} = Var(\bar{X}) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n^2}\left(Var(X_1) + Var(X_2) + \cdots + Var(X_n)\right)$$
$$= \frac{1}{n^2}(n.\sigma^2) = \frac{\sigma^2}{n}$$

$$\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

Remark:

If we are sampling from a population that has an unknown Probability distribution, either finite or infinite size, the Sampling distribution of $\bar{x}$ will still be approximately normal with mean μ & Variance $\frac{\sigma^2}{n}$, Provided the Sample size is large.

Central Limit Theorem:

If $X_1, X_2, \dots X_n$ is a random sample of size 'n' taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$\boxed{Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}, \quad \text{as } n \to \infty,$$

is the Standard normal distribution. $N(0, 1)$

- The normal approximation for $\bar{X}$ will generally be good if $\boxed{n \geq 30}$.

(P1). An electrical firm manufactures light bulbs that have a length of life that is approximately Normally distributed, with mean equal to 800 hours and a Standard deviation of 40 hours. Find the Probability that a random sample of 16 bulbs will have an average life of Less than 775 hours.

Solution:

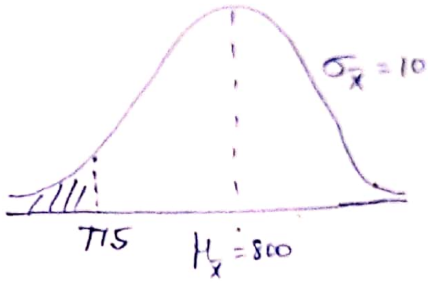
Since random sample of size 16 is taken from Normal Population.

$\therefore$ The sampling distribution of $\bar{x}$ will be approximately normal, with $\mu_{\bar{x}} = \mu = 800$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$

Corresponding to $\bar{x} = 775$,

We find that $Z = \frac{775 - 800}{10}$

$= -2.5$



$\therefore P(\bar{X} < 775) = P(Z < -2.5) = \Phi(-2.5) = 0.0062.$

Ex.1 Suppose that a r.v X has a continuous uniform distribution $f_X(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$

Find the distribution of the sample mean of a random sample of size $n = 40$. [Use Central Limit Theorem].

Draw the distributions of X & $\bar{X}$.

Ex.2 Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume that the mean time is measured to the nearest minute.