ESTIMATION THEORY

Estimation theory is a framework that uses a combination of effect sizes, confidence intervals and precision to plan experiments, analyses data and interpret results originated from samples as a means of getting the true estimate of population parameter.
There are two forms of estimation theory namely:

1. Point estimation
2. Interval estimation

Basic Concepts in Estimation Theory

Sampling Distribution: This is the distribution of values of a statistic. Since the values of statistic are the results of several simple random samples, therefore, they are random variables.

Population distribution: is the distribution of values of its members and has a mean denoted by μ and variance σ^2 when it is normally distributed.

Standard Error: is the standard deviation of the sampling distribution of a statistic. It is a measure of reasonable difference between a particular sample statistic and the population parameter. It is usually denoted by $\sigma_{\bar{x}}$.

The important uses of estimation theory are:

1. It is used in tests of whether a particular sample could have been drawn from a given parent population.
2. Also used in working out confidence limits and confidence intervals.

Point Estimate: is a single numerical value obtained from measurement of samples, for the estimation of corresponding population parameters.

Sampling Distribution of Mean: refers to distribution of all the possible sample means that could be obtained if we select all possible samples of a given size.

It depends on the distribution of the population from which the sample is drawn. If a population is normally distributed, then it is also normally distributed regardless of the sample size. Even if the population is not normal, the sampling distribution of the sample mean tends to be normally distributed as the sample size becomes sufficiently large.

Point Estimation

This is the procedure that allow us to calculate a single sample statistic (such as $\bar{x}$, $\hat{p}$ or s) from selected sample to provide a best estimate of the true value of the corresponding population parameters (such as μ, P or σ).

A Point Estimator is a single relevant statistic obtained from sample observations to provide a best estimate of the true value of the corresponding population parameter while Point Estimate is the value of the statistic obtained from sample observations using a particular estimator.

The shortcoming of point estimates is that they do not tell us how close we can expect them to be to the true value of quantities they are supposed to estimate. In determining the accuracy of estimator, it should always be accompanied by some information known as its properties, (i.e its unbiasedness, sufficiency, efficiency and consistency), which makes it possible to judge their merits.

Properties of Point Estimator

- **Unbiasedness:** An estimator is said to be unbiased if the expected value or mean of such statistic obtained from different samples drawn from the same population is equal to the population parameter being estimated. For instance, $E(\bar{X}) = \mu$
- **Consistency:** A point estimator is said to be consistent if its value tends to become closer to the population parameter as the sample size increases. For instance, $\bar{X}$is a consistent estimator of the population mean μ because the standard error (standard deviation of mean) $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ tends to become smaller as sample size 'n' increases.
- **Efficiency:** Given two unbiased estimator say $\hat{\theta}_1$ & $\hat{\theta}_2$, then $\hat{\theta}_1$is more efficient than $\hat{\theta}_2$ if and only if $V(\hat{\theta}_1) < V(\hat{\theta}_2)$.
- **Sufficiency:** A point estimator $\hat{\theta}$ is said to be sufficient if it contains all the information in the observation given for the estimation of θ.

Interval Estimation

It is the procedure of constructing intervals (i.e. Lower bound and Upper bound) in terms of point estimate and margin of error in which the population parameter being estimated is expected to lie within. The margin of error for a normally distributed population is given as $Z_{1-\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}}$ or $Z_{\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}}$ depending on the type of statistical table in used. Thus, the C.I estimate of a population parameter is obtained by the formula:

Point estimate ± $Z_{1-\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}}$ (11.2)

or Point estimate ± $Z_{\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}}$ (11.3)

For instance, a 95% C.I estimate implies that, if all possible samples of the same size were drawn, then 95% of them would include the true population mean somewhere within the interval around their sample mean and only 5% of them would not.

Generally, we expressed confidence interval as $100(1-\alpha)\%$ and for the determination of level of significance (α), the value of Student-t distribution at infinity (∞) is always equal to that of Normal distribution.

Confidence interval for Population Mean (μ)

This shall be treated under two different cases of small and large sample.

Case 1: When σ is known and n ≥ 30 (large sample)

The $100(1-\alpha)\%$ confidence interval for μ is given as

$$\bar{X} \pm Z_{\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}} \qquad (11.4)$$

Case 2: When σ is unknown and n < 30 (Small sample)

An $100(1-\alpha)\%$ confidence interval for μ is given as

$$\bar{X} \pm t_{\frac{\alpha}{2}}^{(n-1)} {}^{S}/_{\sqrt{n}} \qquad (11.5)$$

$$where\ S = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}} \text{ or } \sqrt{\frac{\sum x_i^2-\frac{(\sum x_i)^2}{n}}{n-1}} \text{ or } \sqrt{\frac{\frac{\sum x_i^2}{n}-\left(\frac{\sum x_i}{n}\right)^2}{n-1}}$$

$$where\ n-1\ is\ the\ degree\ of\ freedom\ and\ t_{\frac{\alpha}{2}}\ is\ the\ t-distribution\ table\ value$$

The shape of the t-distribution is very much like that of the normal distribution curve. i.e. it is symmetrical with zero mean, but there is slightly higher probability of getting values falling into the two tails. This shape depends on the sample size or on degrees of freedom (n – 1). Thus, as sample size increases, t-distribution gradually reaches the normal distribution and 'S' becomes a better estimate of σ.

Example 1 According to a voters register in Lagos state, it was found that the mean and standard deviation of ages of a random sample of 40 registered voters in a particular ward were 25years and 5 years respectively. Construct a 95% confidence interval for the true mean age of the entire registered voters.

Solution: Given $\bar{X} = 25,$ σ = 5, n = 40, α = 0.05, $Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$.The $100(1-\alpha)\%$ confidence interval for μ is given as

$$\bar{X} \pm Z_{\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}} = \bar{X} - Z_{\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} {}^{\sigma}/_{\sqrt{n}}$$

Thus, for a 95% confidence interval for μ is

$$25 - 1.96({}^{5}/_{\sqrt{40}}) < \mu < 25 + 1.96({}^{5}/_{\sqrt{40}})$$

$$=25 - 1.5495160535 < \mu < 25 + 1.5495160535$$

$$=23.450483947 < \mu < 26.5495160535$$

Thus, the average age of all registered voters in the ward lie between 24years and 27years.

Example 2 The following represent the scores of 10 randomly selected students in Statistics methods exams: 59, 40, 38, 67, 78, 45, 56, 60, 45, 80. Construct a 95% confidence interval for the average score of the entire students who registered for the exams.

Solution: Given $\mathrm{n} = 10,\quad \alpha = 0.05$

The $100(1-\alpha)\%$ confidence interval for μ is given as

$$\bar{X} \pm t_{\frac{\alpha}{2}}^{(n-1)} S/\sqrt{n} = \bar{X} - t_{\frac{\alpha}{2}}^{(n-1)} S/\sqrt{n} < \mu < \bar{X} + t_{\frac{\alpha}{2}}^{(n-1)} S/\sqrt{n}$$

$$where\ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{59 + 40 + 38 + 67 + \cdots + 80}{10} = \frac{568}{10} = 56.8$$

$$S = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{59^2+40^2+38^2+\cdots+80^2 - \frac{(568)^2}{10}}{9}} = \sqrt{\frac{34284-32262.4}{9}}$$

$= \sqrt{224.622222}$ =14.9874

$$t_{\frac{\alpha}{2}}^{(n-1)} = t_{\frac{0.05}{2}}^{(9)} = t_{0.025}^{(9)} = 2.26$$

Thus, 95% confidence interval is given as

$$56.8 - 2.26\left(14.9874/\sqrt{10}\right) < \mu < \bar{X} + 56.8 + 2.26(14.9874/\sqrt{10})$$

$$56.8 - 10.711117874 < \mu < 56.8 + 10.711117874$$

$$46.088882126 < \mu < 67.511117874$$

Thus, the average students' score lies between 46 and 68.

Confidence Interval for Difference of two Population Means

Case 1: when σ_1^2 and σ_2^2 are known and $n_1, n_2 \geq 30$

The $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is given as

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sigma_{\bar{X}_1 - \bar{X}_2} \qquad (11.6)$$

where $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$

$\bar{X}_1$ *is the sample mean of the first populaion*

$\bar{X}_2$ *is the sample mean of the second populaion*

$\sigma_{\bar{X}_1-\bar{X}_2}$ *is the pooled standard deviation*

Case 2: when σ_1^2 and σ_2^2 are unknown and $n_1, n_2 < 30$

The $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_2)$ is given as

$$(\bar{X}_1-\bar{X}_2) \pm t_{\frac{\alpha}{2}}^{(v)} S_{\bar{x}_1-\bar{x}_2} \qquad (11.7)$$

Where: $S_{\bar{x}_1-\bar{x}_2} = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ and $S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

$$S_1^2 = \frac{\frac{\sum x_{1i}^2}{n_1}-\left(\frac{\sum x_{1i}}{n_1}\right)^2}{n_1-1} \quad \text{and} \quad S_2^2 = \frac{\frac{\sum x_{2i}^2}{n_2}-\left(\frac{\sum x_{2i}}{n_2}\right)^2}{n_2-1}$$

$v = n_1 + n_2 - 2$ is the degree of freedom

$\bar{X}_1$ *is the sample mean of the first populaion*
$\bar{X}_2$ *is the sample mean of the second populaion*
$S_{\bar{x}_1-\bar{x}_2}$ *is the pooled sample standard deviation*

Example 3: XYZ College keeps records on seminar costs for different number of their employees. If a random sample of 50 staffs is selected from their non-academic staff and the average cost of conducting seminars for them in the year 2016 were found to be #2 million and standard deviation of #10. If another sample of 40 academic staffs are selected for the same year with the average cost and standard deviation of #1million and #5 respectively. Construct 90% confidence interval for the difference in the average cost of conducting training for the entire staffs of the college.

Solution: Given$\boldsymbol{n_1} = \mathbf{50}, \boldsymbol{n_2} = 40,\ \sigma_1 = 10; \sigma_2 = 5; \bar{X}_1 \ = 2{,}000{,}000; \bar{X}_2 = \ 1{,}000{,}000;\ \alpha = 0.10\ Z_{\frac{\alpha}{2}} = Z_{\frac{0.10}{2}} = Z_{0.05} = 1.645$

The $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_2)$ is given as

$$(\bar{X}_1-\bar{X}_2) \pm Z_{\frac{\alpha}{2}}\sigma_{\bar{X}_1-\bar{X}_2}$$

where $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}} = \sqrt{\frac{100}{50}+\frac{25}{40}}$

$$= \sqrt{2+0.625} = \sqrt{2.625} = 1.6202$$

Thus, for a 90% confidence interval for μ is

$(2{,}000{,}000 - 1{,}000{,}000\ ) \pm 1.645(1.6202)$

$1{,}000{,}000 - 2.665229 < \ \mu_1 - \mu_2 < 1{,}000{,}000 + 2.665229$

$\#999{,}997.33 < \ \mu_1 - \mu_2 < \#1{,}000{,}002.67$

Example 4: In a soap making factory, there are two production lines. A sample of 10 bars of soap was taken at random from the first line while sample of 8 bars was taken from the second line and subjected to measurements. The following are the weights in gram recorded for each sample.

1stLine	9.1	9.2	9.6	10	10.2	9.8	9.9	9.9	10.1	10
2ndLine	10.1	9.8	9.9	10.1	10	9.9	10	9.7		

Construct a 90% confidence interval for the difference in the average weights of all soaps manufactured in the factory.

Solution: Given $n_1 = 10, n_2 = 8$, α $= 0.10$

The $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is given as

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}}^{(v)} S_{\bar{x}_1 - \bar{x}_2}$$

$$where\ \bar{X}_1 = \frac{\sum_{i=1}^{n} X_{1i}}{n_1} = \frac{9.1 + 9.2 + 9.6 + \cdots + 10}{10} = \frac{97.8}{10} = 9.78$$

$$\bar{X}_2 = \frac{\sum_{i=1}^{n} X_{2i}}{n_2} = \frac{10.1 + 9.8 + 9.9 + \cdots + 9.7}{8} = \frac{79.5}{8} = 9.9375$$

$v = 10 + 8 - 2 = 16$

$t_{\frac{\alpha}{2}}^{(v)} = t_{\frac{0.05}{2}}^{(16)} = t_{0.025}^{(16)} = 2.120$

$$S_1^2 = \frac{\frac{\sum x_{1i}^2}{n_1} - \left(\frac{\sum x_{1i}}{n_1}\right)^2}{n_1 - 1} = \frac{9.1^2 + 9.2^2 + 9.6^2 + \cdots + 10^2 \ - \frac{(9.78)^2}{10}}{9} = \frac{957.72 \ - \frac{(97.8)^2}{10}}{9}$$

$= 0.137333333$

$$S_2^2 = \frac{\frac{\sum x_{2i}^2}{n_2} - \left(\frac{\sum x_{2i}}{n_2}\right)^2}{n_2 - 1} = \frac{10.1^2 + 9.8^2 + 9.9^2 + \cdots + 9.9^2 \ - \frac{(7.95)^2}{8}}{7}$$

$$= \frac{790.17 \ -\frac{(79.5)^2}{8}}{7} \quad = 0.0198214286$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(10-1)0.137333333+(8-1)0.0198214286}{10+8-2}} \qquad = \sqrt{\frac{1.235999997+0.1387500002}{16}} = 0.0859218748$$

$$S_{\bar{x}_1-\bar{x}_2} = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad = 0.0859218748\sqrt{\frac{1}{10}+\frac{1}{8}}$$

$$= 0.0859218748\sqrt{0.225} = 0.0407563238$$

Thus, the 90% confidence interval is given as

$(9.78 - 9.9375) \pm 1.645(0.0407563238) = -0.1575 \pm 0.0670441527$

$$-0.224544152 < \ \mu_1 - \mu_2 < -0.090455847$$

$$-0.22gram < \ \mu_1 - \mu_2 < -0.09gram$$

Confidence interval for Proportion

The $100(1-\alpha)\%$ confidence interval for population proportion 'P' is given as

$$\hat{P} \pm Z_{\frac{\alpha}{2}}\sigma_P \ \Rightarrow \hat{P} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \qquad (11.8)$$

where $\hat{P} = \frac{x}{n} \qquad \boldsymbol{x\ is\ the\ number\ of\ success\ in\ 'n'\ trials}$

Example 5

Out of a random sample of 800 people who casted their votes at Lagos polling post, it was discovered that only 400 people voted for APC. Construct a 95% confidence interval for the actual proportion of the entire voters who casted their votes for APC at the polling station.
Solution: Given $n = 800, x = 400, \alpha = 0.05$

The $100(1-\alpha)\%$ confidence interval for population proportion 'P' is given as

$$\hat{P} \pm Z_{\frac{\alpha}{2}}\sigma_P \ \Rightarrow \hat{P} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

where $\hat{P} = \frac{x}{n} = \frac{400}{800} = 0.5$

$$\sigma_P \quad = \sqrt{\frac{0.5(1-0.5)}{400}} \ = 0.025$$

$0.5 \pm 1.96(0.025) = 0.5 \mp 0.049$

$0.451 < P < 0.549$

Confidence Interval for Difference of two Population Proportions

The $100(1-\alpha)\%$ confidence interval for difference of 2 proportions P_1 and P_2 is given as

$$(\hat{P}_1 - \hat{P}_2) \pm Z_{\frac{\alpha}{2}} \sigma_{\hat{P}_1 - \hat{P}_2} \qquad (11.8)$$

Where $\hat{P}_1 is\ the\ sample\ proportion\ for\ the\ first\ population$
$\hat{P}_2\ is\ the\ sample\ proportion\ for\ the\ first\ population$

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\sigma^2_{\hat{P}_1} + \sigma^2_{\hat{P}_2}} = \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

Example 6: In a random sample of 50 Business administration students who sat for Business Statistics exams, only 40 passed while 20 passed from a random sample of 40 Accounting students. Construct a 95% confidence interval for the difference in proportion of all students who passed Business statistics in Business administration and Accounting department.

Solution: Given $\boldsymbol{n_1 = 50},\ \boldsymbol{x_1} = 40, \boldsymbol{n_2} = 40,\ x_2 = 20$

The $100(1-\alpha)\%$ confidence interval for difference of 2 proportions P_1 and P_2 is given as

$$(\hat{P}_1 - \hat{P}_2) \pm Z_{\frac{\alpha}{2}} \sigma_{\hat{P}_1 - \hat{P}_2}$$

Where $\hat{P}_1 = \frac{x_1}{n_1} = \frac{40}{50} = 0.8$ and $\hat{P}_2 = \frac{x_2}{n_2} = \frac{20}{40} = 0.5$

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{0.8(0.2)}{50} + \frac{0.5(0.5)}{40}} = 0.0972111105$$

Thus, 95% confidence interval for P_1 - P_2 is given as

$$(0.8 - 0.5) \pm 1.96(0.0972111105) = 0.3 \mp 0.1905337766$$

$$0.1094662234 < P_1 - P_2 < 0.4905337766$$

$$0.1095 < P_1 - P_2 < 0.4905$$