Note: Greek letter will be used for population parameters, whereas sample statistics will be represented using specific alphabets.

Sample mean of *n* random observations drawn from a population is defined as,

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

Sample variance is defined as,

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

The sample standard deviation or standard deviation is defined as,

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

Quantitative variables

Let the variable of interest is measured on a ratio or an interval scale, and its population mean, and variance are given by μ and σ^2, respectively. We draw a random sample $X_1, X_2, X_3, \ldots X_n$ of size n from the population. The sample mean is:

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

Qualitative variables (two categories)

Let the variable of interest is a qualitative (nominal) variable, say with two categories. For example, $\{male, female\}$. The population proportion of the variable is given by π.

We could use a Bernoulli random variable to represent this categorical variable in the following way:

$$X = \left\{ \begin{array}{c} 1 \; with \; probability \; \pi \\ 0 \; with \; probability \; (1-\pi) \end{array} \right\}$$

$$\mu = E(X) = \pi.1 + (1-\pi).0 = \pi$$

$$\sigma^2 = E(X-\mu)^2 = E(X^2) - \mu^2 = \pi.1 - \pi^2 = \pi(1-\pi)$$

Therefore, the qualitative variables could be treated similar to quantitative variable with mean $\mu = \pi$ and variance $\sigma^2 = \pi(1-\pi)$.

Sample Mean

We draw a random sample $X_1, X_2, X_3, \ldots X_n$ of size n from the population. The sample mean for a quantitative variable is:

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

In the case of a qualitative variable, we define X_i as follows:

$$X_i = \begin{Bmatrix} 1\ if\ observation\ i\ belongs\ to\ the\ category\ of\ interest \\ 0\ otherwise \end{Bmatrix}$$

Then the sample mean (of a categorical variable) represents proportion for the category of interest:

$$p = \frac{\sum_{i=1}^{n} X_i}{n};$$

Central Limit Theorem (CLT)

Let $X_1, X_2, X_3, \ldots X_n$ be a random sample from a population with mean μ and variance σ^2. Then for large n, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ even if the underlying distribution of the individual observations in the population is not normal. (The symbol ~ is used to represent "approximately distributed".)

Example 1 (Quantitative variable):

For the interval or ratio scale variable, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, we directly apply CLT, which gives us the following result.

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ which can also be expressed in terms of standardized variable, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Example 2: (Qualitative variable):

Applying CLT on a categorical variable gives the following result:

$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Which can also be expressed as, $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$

Variance of sample mean and standard error

The population mean (μ) for a quantitative variable is estimated using sample mean $(\bar{X})$. Further,

$$\sigma_{\bar{X}}^2 = Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Therefore, *standard error of mean* (sem) or simply *standard error* is defined as,

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

For a given random sample, the population variance σ^2 is estimated using s^2.

The standard error $\sigma_{\bar{X}}$ is estimated by $s_{\bar{X}} = \frac{s}{\sqrt{n}}$.

The standard error represents the estimated standard deviation from a set of sample means from repeated samples of size *n* from a population with underlying variance σ^2.

For a *qualitative variable* the population proportion is estimated using *p*, and we calculate standard error of proportion in the following way,

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

Confidence interval

The interval (L, U) is a (1-α)100% confidence interval for a parameter θ (μ, π *etc.*) if, the probability that θ lies within the interval (L, U) is (1-α).

$$Probability(L \leq \theta \leq U) = 1 - \alpha$$

For the sake of brevity, we will use $P(.)$ instead of $Probability(.)$.

The standard normal distribution (Z), has symmetric distribution with respect to origin. Hence, 95% confidence interval is defined as,

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P(|Z| \leq z_{\alpha/2}) = 1 - \alpha;\ \alpha = 0.05 \qquad (1)$$

Since $z_{\alpha/2} = 1.96$, therefore the 95% confidence interval for the standard normal distribution is (-1.96, 1.96).

Determining sample size given precision

Example 3: Quantitative variable (mean)

Suppose the researcher wants to estimate the mean household monthly expense for department store shopping so that the estimate is within $\pm$₹ 5.00 of the true population. Past studies indicate that the population standard deviation σ can be assumed to be ₹ 55.

We are given, $D = |\bar{X} - \mu| = 5.00$, and $\sigma = 55$

From CLT, we also know $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z$

For 95% confidence interval, $z_{\alpha/2} = 1.96$

$$\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| = z_{\alpha/2}, therefore\ \sqrt{n}\frac{D}{\sigma} = z_{\alpha/2}$$

Hence, $n = \frac{{z_{\alpha/2}}^2.\sigma^2}{D^2} = \frac{1.96^2.55^2}{5^2} \approx 465$

Confidence interval from sample

Suppose a sample of n = 465 is drawn, and these observations generate a mean $\bar{x} = 180$. Further, the sample standard deviation $s = 50$. Then the confidence interval for μ, with revised estimates can be calculated in the following way using equation 1.

$$|Z| \leq z_{\alpha/2} \implies \left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}$$

Which on expansion results in $\left(\bar{X} - z_{\alpha/2}\,\sigma/\sqrt{n}\right) \leq \mu \leq \left(\bar{X} + z_{\alpha/2}\,\sigma/\sqrt{n}\right)$

For large sample, $n \geq 200$, if σ is unknown then we can estimate it using sample standard deviation. Therefore, the revised confidence interval using realized sample mean ($\bar{x}$), and sample standard deviation (s) is $\bar{x} - z_{\alpha/2}\,s/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2}\,s/\sqrt{n}$.

Note: Strictly speaking, when the population standard deviation is unknown, we use *t*-distribution with $n - 1$ degrees of freedom to get the confidence interval. But, for large sample size, *t*-distribution resembles with z-distribution. We will discuss it in detail in Chapter 15.

Substituting the relevant values in the above expression, we get

$$180 - 1.96\left(\frac{50}{\sqrt{465}}\right) \leq \mu \leq 180 + 1.96\left(\frac{50}{\sqrt{465}}\right)$$

Hence, (175.45, 184.55) is the confidence interval for μ which is narrower than the planned because population standard deviation was overestimated.

In some case, precision is expressed in relative rather than absolute terms. For example, it could be specified that the estimate be within R percentage points of the mean. Mathematically,

$$D = R\mu$$

The sample size is determined by, $n = \frac{{z_{\alpha/2}}^2}{R^2}\left(\frac{\sigma}{\mu}\right)^2$

Example 4: Qualitative variable (proportion)

Suppose that the researcher is interested in estimating the proportion of households possessing the department store credit card. Determine the sample size if the desired precision and confidence level is 5% and 95%, respectively. Based on secondary data, the researcher estimates that 64% of the households in the target population possess a department store credit card.

We have

$$\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim Z$$

Therefore,

$$n\left(\frac{D^2}{\pi(1-\pi)}\right) = {z_{\alpha/2}}^2$$

$$n = {z_{\alpha/2}}^2 \frac{\pi(1-\pi)}{D^2} = \frac{1.96^2(.64)(1-.64)}{0.05^2} = 355$$

Suppose that after the sample has been taken, the sample proportion p turns out to be 0.55. We re-estimate the confidence interval by using s_p.

$$\left(p - z_{\alpha/2}s_p\right) \leq \pi \leq \left(p + z_{\alpha/2}s_p\right) \Longrightarrow 0.55 - 1.96\sqrt{\frac{0.55x0.45}{355}} \leq \pi \leq 0.55 + 1.96\sqrt{\frac{0.55x0.45}{355}}$$

Therefore, the confidence interval is 0.55 ± 0.052 which is wider than that specified by the researcher. This is because standard deviation based on $p = 0.55$ is larger than the estimate of the population standard deviation based on $\pi = 0.64$.

If this wider interval is unacceptable, then the sample size can be determined to reflect maximum possible variation in the population. This happens when $\pi = 0.5$, and the corresponding sample size is

$$n = {z_{\alpha/2}}^2 \frac{\pi(1-\pi)}{D^2} = \frac{1.96^2(.5)(1-.5)}{0.05^2} = 385$$

If the precision is specified in relative terms then, $D = R\pi$

$$n = {z_{\alpha/2}}^2 \frac{(1-\pi)}{R^2\pi}$$

Finite population correction

If the resulting sample size represents 10 percent or more of the population, then we apply finite population correction (fpc).

Without fpc:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z \qquad (2)$$

With fpc (we are producing the result below without the mathematical proof)

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n_f}}\sqrt{\frac{N-n_f}{N-1}}} = Z \quad (3)$$

In the above two expressions, n corresponds to the sample size without fpc, whereas n_f, corresponds to the sample size with fpc.

Comparing (2) and (3) gives us,

$$\sigma/\sqrt{n} = \frac{\sigma}{\sqrt{n_f}}\sqrt{\frac{N-n_f}{N-1}}$$

$$\frac{1}{n} = \frac{1}{n_f}\left(\frac{N-n_f}{N-1}\right)$$

Hence, $n_f = n\left(1-\frac{n-1}{N+n-1}\right)$

Similarly, fpc for proportion results in the same expression.