Introduction to Sampling Theory

Lecture 8
Simple Random Sampling

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Estimation of Confidence limits for the population mean: σ^2 known

Assume that the population is normally distributed $N(\bar{Y},\sigma^2)$ with mean ($\bar{Y}$) and variance (σ^2).

When σ^2 is known, then the $100(1-\alpha)\%$ confidence interval is

$$\left[\bar{y} - Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})} \le \bar{Y} \le \bar{y} + Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})}\right]$$

where $Z_{\alpha/2}$ denotes the upper ($\alpha/2$)% points of $N(\mu,\sigma^2)$.

Estimation of Confidence limits for the population mean: σ^2 unknown

When $\boldsymbol{\sigma^2}$ is unknown, then the $\mathbf{100(1-\alpha)\%}$ confidence interval is

$$\left[\bar{y} - t_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})} \le \bar{Y} \le \bar{y} + t_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})}\right]$$

where $t_{\alpha/2}$ denotes the upper (α/2)% points of *t* distribution with (*n* – 1) degrees of freedom.

Proof: Useful Result when σ^2 is known and unknown

Assume that the population is normally distributed $N(\bar{Y}, \sigma^2)$

with mean $\bar{Y}$ and variance σ^2 .

Then

$$\frac{\bar{y} - \bar{Y}}{\sqrt{Var(\bar{y})}} \sim N(0,1)$$ **when σ^2 is known.**

$$\frac{\bar{y} - \bar{Y}}{\sqrt{\widehat{Var}(\bar{y})}} \sim t_{(n-1)}$$ **when σ^2 is unknown.**

Proof: Estimation of Confidence limits for the population mean: σ^2 known

When σ^2 is known, then the 100(1 – α)% confidence interval is

given by

$$P\left[-Z_{\frac{\alpha}{2}} \le \frac{\bar{y}-\bar{Y}}{\sqrt{Var(\bar{y})}} \le Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

or $$P\left[\bar{y}-Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})} \le \bar{Y} \le \bar{y}+Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})}\right] = 1-\alpha$$

the confidence limits are

$$\left(\bar{y}-Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})},\ \bar{y}+Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})}\right)$$

where $Z_{\alpha/2}$ denotes the upper (α/2)% points on *N*(0, 1) distribution.

Proof: Estimation of Confidence limits for the population mean: σ^2 unknown

When σ^2 is unknown, then the 100(1 – α)% confidence interval is given by

$$P\left[-t_{\frac{\alpha}{2}} \le \frac{\bar{y}-\bar{Y}}{\sqrt{\widehat{Var}(\bar{y})}} \le t_{\frac{\alpha}{2}}\right] = 1-\alpha$$

or $$P\left[\bar{y}-t_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})} \le \bar{Y} \le \bar{y}+t_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})}\right] = 1-\alpha$$

the confidence limits are

$$\left[\bar{y}-t_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})} \le \bar{Y} \le \bar{y}+t_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\bar{y})}\right]$$

where $t_{\alpha/2}$ denotes the upper (α/2)% points on *t* distribution with (*n* – 1) degrees of freedom.

Estimation of Population Total

Sometimes, it is also of interest to estimate the population total, e.g. total household income, total expenditures etc.

Let Y_T denotes the population total defined as

$$Y_T = \sum_{i=1}^{N} Y_i = N\overline{Y}$$

Y_T can be estimated by

$$\hat{Y}_T = N\hat{\overline{Y}} = N\overline{y}.$$

Estimation of Population Total

Then $E(\hat{Y}_T) = NE(\bar{y}) = N\bar{Y} = Y_T.$

Variance of $\hat{Y}_T$ **is**

$$Var(\hat{Y}_T) = N^2\ Var(\bar{y})$$

$$= \begin{cases} N^2\left(\dfrac{N-n}{Nn}\right)S^2 = \dfrac{N(N-n)}{n}S^2 & \textbf{for SRSWOR} \\ N^2\left(\dfrac{N-1}{Nn}\right)S^2 = \dfrac{N(N-1)}{n}S^2 & \textbf{for SRSWR.} \end{cases}$$

Estimate of variance of $\hat{Y}_T$ **is**

$$\widehat{Var}(\hat{Y}_T) = N^2\ \widehat{Var}(\bar{y})$$

$$= \begin{cases} \dfrac{N(N-n)}{n}s^2 & \textbf{for SRSWOR} \\ \dfrac{N^2}{n}s^2 & \textbf{for SRSWR.} \end{cases}$$

Determination of Sample Size

The size of the sample is needed before the survey starts and goes into operation.

When the sample size increases, the variance of estimators decreases but the cost of survey increases and vice versa.

So there has to be a balance between the two aspects.

Determination of Sample Size

The sample size can be determined on the basis of prescribed values of the

- **standard error of the sample mean,**
- **error of estimation,**
- **width of the confidence interval,**
- **coefficient of variation of the sample mean,**
- **relative error of sample mean or total cost among several others.**

Determination of Sample Size

An important constraint or need to determine the sample size is that the information regarding the population standard derivation *S* should be known for these criteria.

The reason and need for this will be clear when we derive the sample size.

A question arises about how to have information about *S* beforehand?

Determination of Sample Size

A possible solution to this issue is to conduct a pilot survey and collect a preliminary sample of small size, estimate *S* and use it as a known value of *S* it.

Alternatively, such information can also be collected from past data, past experience, the long association of experimenter with the experiment, prior information etc.

Now we find the sample size under different criteria assuming that the samples have been drawn using SRSWOR. The case for SRSWR can be derived similarly.

1. Prespecified Variance

The sample size is to be determined such that the variance of $\bar{y}$ should not exceed a given value, say *V*. In this case, find n such that

$$Var(\bar{y}) \leq V$$

or $$\frac{N-n}{Nn} S^2 \leq V$$

or $$\frac{1}{n} - \frac{1}{N} \leq \frac{V}{S^2}$$

or $$\frac{1}{n} - \frac{1}{N} \leq \frac{1}{n_e}$$ **where** $$n_e = \frac{S^2}{V}.$$

or $$n \geq \frac{n_e}{1 + \frac{n_e}{N}}$$

1. Prespecified Variance

It may be noted here that n_e can be known only when S^2 is known.

This reason compels to assume that *S* should be known.

The same reason will also be seen in other cases.

The smallest sample size needed in this case is

$$n_{smallest} = \frac{n_e}{1 + \frac{n_e}{N}}$$

It *N* is large, then the required *n* is $n \geq n_e$ **and** $n_{smallest} = n_e$ **.**

2. Pre-specified Estimation Error

It may be possible to have some prior knowledge of population mean $\bar{Y}$ and it may be required that the sample mean $\bar{y}$ should not differ from it by more than a specified amount of absolute estimation error, i.e., which is a small quantity.

Such a requirement can be satisfied by associating a probability (1 – α) with it and can be expressed as

$$P\left[\left|\bar{y}-\bar{Y}\right| \leq e\right] = (1-\alpha).$$

2. Pre-specified Estimation Error

Assuming the normal distribution for the population,

$$\bar{y} \sim N\left(\bar{Y}, \frac{N-n}{Nn} S^2\right)$$

we can write

$$P\left[\frac{\left|\bar{y}-\bar{Y}\right|}{\sqrt{Var(\bar{y})}} \leq \frac{e}{\sqrt{Var(\bar{y})}}\right] = 1-\alpha$$

which implies that

$$\frac{e}{\sqrt{Var(\bar{y})}} = Z_{\frac{\alpha}{2}}$$

or $$Z_{\frac{\alpha}{2}}^2 Var(\bar{y}) = e^2$$

2. Pre-specified Estimation Error

Now

$$Z_{\frac{\alpha}{2}}^{2}\frac{N-n}{Nn}S^{2}=e^{2}$$

or

$$n=\left(\frac{\left(Z_{\frac{\alpha}{2}}S/e\right)^{2}}{1+\frac{1}{N}\left(Z_{\frac{\alpha}{2}}S/e\right)^{2}}\right)$$

which is the required sample size. If *N* is large then

$$n=\left(Z_{\frac{\alpha}{2}}S/e\right)^{2}.$$

3. Pre-specified Width of the Confidence Interval

If the requirement is that the width of the confidence interval of $\bar{y}$ with confidence coefficient (1 – α) should not exceed a prespecified amount *W*, then the sample size *n* is determined such that

$$2Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})} \leq W$$

assuming σ^2 is known and population is normally distributed.

3. Pre-specified Width of the Confidence Interval

This can be expressed as

$$2Z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{Nn}}S \le W$$

or

$$4Z_{\frac{\alpha}{2}}^{2}\left(\frac{1}{n}-\frac{1}{N}\right)S^{2} \le W^{2}$$

or

$$\frac{1}{n} \le \frac{1}{N} + \frac{W^{2}}{4Z_{\frac{\alpha}{2}}^{2}S^{2}}$$

or

$$n \ge \frac{\dfrac{4Z_{\frac{\alpha}{2}}^{2}S^{2}}{W^{2}}}{1+\dfrac{4Z_{\frac{\alpha}{2}}^{2}S^{2}}{NW^{2}}}.$$

3. Pre-specified Width of the Confidence Interval

The minimum sample size required is $$n_{smallest} = \frac{\dfrac{4Z_{\frac{\alpha}{2}}^2 S^2}{W^2}}{1 + \dfrac{4Z_{\frac{\alpha}{2}}^2 S^2}{NW^2}}$$

If *N* is large then

$$n \geq \frac{4Z_{\frac{\alpha}{2}}^2 S^2}{W^2}$$

and the minimum sample size needed is

$$n_{smallest} = \frac{4Z_{\frac{\alpha}{2}}^2 S^2}{W^2}$$

4. Pre-specified Coefficient of Variation

The coefficient of variation (CV) is defined as the ratio of standard error (or standard deviation) and mean.

The knowledge of coefficient of variation has played an important role in the sampling theory as this information has helped in deriving efficient estimators.

If it is desired that the CV of $\bar{y}$ should not exceed a given or pre-specified value of CV, say C_0.

4. Pre-specified Coefficient of Variation

$$CV(\bar{y}) \le C_0$$

or $$\frac{\sqrt{Var(\bar{y})}}{\bar{Y}} \le C_0$$

or $$\frac{\frac{N-n}{Nn} S^2}{\bar{Y}^2} \le C_0^2$$

or $$\frac{1}{n} - \frac{1}{N} \le \frac{C_0^2}{C^2}$$ **or** $$n \ge \frac{\frac{C^2}{C_o^2}}{1 + \frac{C^2}{NC_0^2}}$$

is the required sample size where $C = \frac{S}{\bar{Y}}$ **is the population coefficient of variation.**

4. Pre-specified coefficient of Variation

The smallest sample size needed in this case is

$$n_{smallest} = \frac{\frac{C^2}{C_0^2}}{1 + \frac{C^2}{NC_0^2}}$$

If *N* is large, then

$$n \geq \frac{C^2}{C_0^2}$$

and

$$n_{smalest} = \frac{C^2}{C_0^2}$$

5. Pre-specified Relative Error

When $\bar{y}$ is used for estimating the population mean $\bar{Y}$, then the relative estimation error is defined as $\dfrac{\bar{y}-\bar{Y}}{\bar{Y}}$.

If it is required that such relative estimation error should not exceed a pre-specified value *R* with probability (1 – α) , then such requirement can be satisfied by expressing it like such requirement can be satisfied by expressing it like

$$P\left[\frac{\left|\bar{y}-\bar{Y}\right|}{\sqrt{Var(\bar{y})}} \le \frac{R\bar{Y}}{\sqrt{Var(\bar{y})}}\right] = 1-\alpha.$$

5. Pre-specified Relative Error

Assuming the population to be normally distributed,

$$\bar{y} \sim N\left(\bar{Y}, \frac{N-n}{Nn} S^2\right).$$

So it can be written that

$$\frac{R\bar{Y}}{\sqrt{Var(\bar{y})}} = Z_{\frac{\alpha}{2}}$$

or

$$Z_{\frac{\alpha}{2}}^2 \left(\frac{N-n}{Nn}\right) S^2 = R^2 \bar{Y}^2$$

or

$$\left(\frac{1}{n} - \frac{1}{N}\right) = \frac{R^2}{C^2 Z_{\frac{\alpha}{2}}^2}$$

5. Pre-specified Relative Error

or

$$n = \frac{\left(\frac{Z_{\frac{\alpha}{2}} C}{R}\right)^2}{1 + \frac{1}{N}\left(\frac{Z_{\frac{\alpha}{2}} C}{R}\right)^2}$$

where $C = \frac{S}{\bar{Y}}$ **is the population coefficient of variation and should be known.**

If N **is large, then**

$$n = \left(\frac{z_{\frac{\alpha}{2}} C}{R}\right)^2 .$$

6. Pre-specified Cost

Let an amount of money *C* is being designated for sample survey to collect *n* observations,

C_0 be the overhead cost and

C_1 be the cost of collection of one unit in the sample.

Then the total cost *C* can be expressed as

$$C = C_0 + nC_1$$

or

$$n = \frac{C - C_0}{C_1}$$

is the required sample size.