Introduction to Sampling Theory

Lecture 10
Simple Random Sampling for Proportions and Percentages

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

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

Sampling for Proportions and Percentages

Notations and relationships

Population	Sample
$P = \frac{A}{N} = \overline{Y}$	$p = \frac{a}{n} = \overline{y}$
$Q = 1 - P$	$q = 1 - p$
$S^2 = \frac{N}{N-1} PQ$	$s^2 = \frac{n}{n-1} pq$

Estimation of Population Proportion and Percentage

Estimate population proportion by sample mean

$$\bar{y} = p = \sum_{i=1}^{n} y_i/n \, .$$

The variance of *p* under SRSWOR and SRSWR are

$$Var_{WOR}(p) = \frac{N-n}{N-1} \cdot \frac{PQ}{n}$$ **in case of SRSWOR.**

$$Var_{WR}(p) = \frac{PQ}{n}$$ **in case of SRSWR.**

Estimation of Population Proportion and Percentage

The estimate of variance of *p* under SRSWOR and SRSWR are

$$\widehat{Var}_{WOR}(p) = \frac{N-n}{N(n-1)}pq$$ **in case of SRSWOR.**

$$\widehat{Var}_{WR}(p) = \frac{pq}{n-1}$$ **in case of SRSWR.**

The standard error of *p* is found by

$$+\sqrt{\widehat{Var}(p)}$$

Proof: Sample proportion *p* is an unbiased estimator of population proportion

Since sample mean $\bar{y}$ an unbiased estimator of population mean $\bar{Y}$

in case of SRSWOR and SRSWR, so

$$E(\bar{y}) = E(p) = \bar{Y} = P$$

and *p* is an unbiased estimator of *P*.

Proof: Variance and Standard Error of *p* under SRSWOR

Using the expression of $var(\bar{y})$ under SRSWOR, the variance of *p* and its estimate can be derived as

$$Var_{WOR}(p) = Var_{WOR}(\bar{y}) = \frac{N-n}{Nn}S^2$$

$$= \frac{N-n}{Nn}.\frac{N}{N-1}PQ$$

$$= \frac{N-n}{N-1}.\frac{PQ}{n}.$$

$$\widehat{Var}(p)_{WOR} = \widehat{Var}_{WOR}(\bar{y}) = \frac{N-n}{Nn}s^2$$

$$= \frac{N-n}{Nn}\frac{n}{n-1}pq$$

$$= \frac{N-n}{N(n-1)}pq.$$

Proof: Variance and Standard Error of *p* under SRSWR

Using the expression of $var(\bar{y})$ under SRSWR, the variance of *p* and its estimate can be derived as

$$Var_{WR}(p) = Var_{WR}(\bar{y}) = \frac{N-1}{Nn} S^2$$

$$= \frac{N-1}{Nn} \frac{N}{N-1} PQ$$

$$= \frac{PQ}{n}$$

$$\widehat{Var}_{WR}(p) = \frac{n}{n-1} . \frac{pq}{n}$$

$$= \frac{pq}{n-1}.$$

Estimation of Population Total or Total Number of Count

An estimate of population total *A* (or total number of count) is

$$\hat{A} = Np = \frac{Na}{n},$$

its variance is

$$Var(\hat{A}) = N^2 Var(p)$$

and the estimate of variance is

$$\widehat{Var}(\hat{A}) = N^2 \widehat{Var}(p).$$

Confidence Interval Estimation of *P*

If *N* and *n* are large then $\frac{p-P}{\sqrt{Var(p)}}$ **approximately follows *N*(0,1).**

With this approximation, we can write and then the 100(1 – α)%

confidence interval of *P* is

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

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

Confidence Interval Estimation of *P*

It may be noted that in this case, a discrete random variable is being approximated by a continuous random variable, so a continuity correction *1/2n* can be introduced in the confidence limits and the limits become

$$\left(p - Z_{\frac{\alpha}{2}} \sqrt{Var(p)} + \frac{1}{2n},\ p + Z_{\frac{\alpha}{2}} \sqrt{Var(p)} - \frac{1}{2n} \right).$$

Estimation of Proportion for More than Two Classes

We have assumed up to now that there are only two classes in which the population can be divided based on a qualitative characteristic.

There can be situations when the population is to be divided into more than two classes.

Estimation of Proportion for More than Two Classes

For example, the taste of a coffee can be divided into four categories very strong, strong, mild and very mild.

Similarly, in another example, the damage to crop due to the storm can be classified into categories like heavily damaged, damaged, minor damage and no damage etc.

Estimation of Proportion for More than Two Classes

These type of situations can be represented by dividing the population of size into, say k, mutually exclusive classes $C_1, C_2, \ldots, C_k$

Corresponding to these classes, let be the proportions of units in the classes $C_1, C_2, \ldots, C_k$ respectively.

Let a sample of size n is observed such that $c_1, c_2, \ldots, c_k$ number of units have been drawn from $C_1, C_2, \ldots, C_k$ respectively.

Estimation of Proportion for More than Two Classes

Then the probability of observing $c_1, c_2, \ldots, c_k$ is

$$P(c_1, c_2, \ldots, c_k) = \frac{\binom{C_1}{c_1}\binom{C_2}{c_2}\cdots\binom{C_k}{c_k}}{\binom{N}{n}}$$

$$\sum_{i=1}^{k} C_i = N, \sum_{i=1}^{k} c_i = n.$$

The population proportions P_i can be estimated by

$$p_i = \frac{c_i}{n}, i = 1, 2, \ldots, k.$$

Estimation of Proportion for More than Two Classes

It can be shown that

$$E(p_i) = P_i, \quad i = 1, 2, \ldots, k,$$

$$Var(p_i) = \frac{N-n}{N-1}\frac{P_iQ_i}{n}$$

and

$$\widehat{Var}(p_i) = \frac{N-n}{N}\frac{p_iq_i}{n-1}$$

For estimating the number of units in the *i*th class,

$$\hat{C}_i = Np_i$$

$$Var(\hat{C}_i) = N^2 Var(p_i)$$

and

$$\widehat{Var}(\hat{C}_i) = N^2 \widehat{Var}(p_i).$$

Estimation of Proportion for More than Two Classes

The confidence intervals can be obtained based on a single p_i as in the case of two classes.

If N is large, then the probability of observing c_1, c_2,…, c_k can be approximated by multinomial distribution given by

$$P(c_1, c_2, \ldots, c_k) = \frac{n!}{c_1!c_2!\ldots c_k!} P_1^{c_1} P_2^{c_2} \ldots P_k^{c_k}$$

Estimation of Proportion for More than Two Classes

For this distribution

$$E(p_i) = P_i, \quad i = 1, 2, .., k,$$

$$Var(p_i) = \frac{P_i(1 - P_i)}{n}$$

and

$$\widehat{Var}(\hat{p}_i) = \frac{p_i(1 - p_i)}{n}.$$

Use of Hypergeometric distribution:

When SRS is applied for the sampling of a qualitative characteristic, the methodology is to draw the units one-by-one, and so the probability of selection of every unit remains the same at every step.

If n sampling units are selected together from N units, then the probability of selection of units does not remain the same as in the case of SRS.

Use of Hypergeometric Distribution:

Consider a situation in which the sampling units in a population are divided into two mutually exclusive classes.

Let P : Proportions of sampling units in the population belonging to class '1'

Q : Proportions of sampling units in the population belonging to class '2'

NP : Total number of sampling units in the population belonging to class '1'

NQ : Total number of sampling units in the population belonging to class '2' and so

$$NP + NQ = N.$$

Use of Hypergeometric Distribution:

The probability that in a sample of *n* selected units out of *N* units by SRS such that n_1 selected units belong to class '1' and n_2 selected units belong to class '2' is governed by the hypergeometric distribution and

$$P(n_1) = \frac{\binom{NP}{n_1}\binom{NQ}{n_2}}{\binom{N}{n}}$$

Use of Hypergeometric Distribution:

As *N* grows large, the hypergeometric distribution tends to Binomial distribution and $P(n_1)$ is approximated by

$$P(n_1) = \binom{n}{n_1} p^{n_1} (1-p)^{n_2}$$

Inverse Sampling

In general, it is understood in the SRS methodology for a qualitative characteristic that the attribute under study is not a rare attribute.

If the attribute is rare, then the procedure of estimating the population proportion P by sample proportion n/N is not suitable.

Some such situations are, e.g., estimation of the frequency of the rare type of genes, the proportion of some rare type of cancer cells in a biopsy, proportion of the rare type of blood cells affecting the red blood cells etc.

In such cases, the methodology of inverse sampling can be used.

Inverse Sampling

In the methodology of inverse sampling,

the sampling is continued until a predetermined number of units possessing the attribute under study occur in the sampling, which is useful for estimating the population proportion.

The sampling units are drawn one-by-one with equal probability and without replacement.

The sampling is discontinued as soon as the number of units in the sample possessing the characteristic or attribute equals a predetermined number. .

Inverse Sampling

Let *m* denotes the predetermined number indicating the number of units possessing the characteristic.

The sampling is continued <u>till *m* number</u> of units are obtained.

Therefore, the sample size *n* required to attain *m* becomes a random variable.

Probability Distribution Function of *n*

In order to find the probability distribution function of n, consider the stage of drawing of samples t such that at $t = n$, the sample size n completes the m units with attribute.

Thus the first $(t - 1)$ draws would contain $(m - 1)$ units in the sample possessing the characteristic out of NP units.

Equivalently, there are $(t - m)$ units which do not possess the characteristic out of NQ such units in the population.

Note that the last draw must ensure that the units selected possess the characteristic.

Probability Distribution Function of *n*

So the probability distribution function of *n* can be expressed as

$$P(n) = P\begin{pmatrix}\textbf{In a sample of } (n\text{-}1) \textbf{ units} \\ \textbf{drawn from } N,\ (m\text{-}1) \textbf{ units} \\ \textbf{will possess the attribute}\end{pmatrix} \times P\begin{pmatrix}\textbf{The unit drawn at} \\ \textbf{the } n^{th} \textbf{ draw will} \\ \textbf{possess the attribute}\end{pmatrix}$$

$$= \left[\frac{\binom{NP}{m-1}\binom{NQ}{n-m}}{\binom{N}{n-1}}\right]\left(\frac{NP-m+1}{N-n+1}\right), \quad n = m, m+1, \ldots, m+NQ.$$

Probability Distribution Function of *n*

Note that the first term

$$\left[\frac{\binom{NP}{m-1}\binom{NQ}{n-m}}{\binom{N}{n-1}}\right]$$

is derived using hypergeometric distribution.

It is the probability for deriving a sample of size (*n* – 1) in which (*m* – 1) units are from *NP* units and (*n* – *m*) units are from units.

$$\left(\frac{NP-m+1}{N-n+1}\right)$$

: The second term is the probability associated with the last draw, where it is assumed that we get the unit possessing the characteristic.

Note that $\sum_{n=m}^{m+NQ} P(n) = 1.$

Estimate of Population Proportion

Consider the expectation of $\dfrac{m-1}{n-1}$ **.**

$$E\left(\frac{m-1}{n-1}\right)=\sum_{n=m}^{m+NQ}\left(\frac{m-1}{n-1}\right)P(n)=\sum_{n=m}^{m+NQ}\left(\frac{m-1}{n-1}\right)\frac{\binom{NP}{m-1}\binom{NQ}{n-m}}{\binom{N}{n-1}}.\frac{Np-m+1}{N-n+1}$$

$$=\sum_{n=m}^{m+NQ-1}\left(\frac{NP-m+1}{N-n+1}\right)\frac{\binom{NP-1}{m-2}\binom{NQ}{n-m}}{\binom{N-1}{n-2}}$$

which is obtained by replacing *NP* by *NP* – 1, *m* by (*m* – 1) and *n* by (*n* - 1) in the earlier step.

Thus $$E\left(\frac{m-1}{n-1}\right)=P.$$

So $\hat{P}=\dfrac{m-1}{n-1}$ **is an unbiased estimator of *P*.**

Estimate of Variance of $\hat{P}$

Now we derive an estimate of the variance of $\hat{P}$. By definition

$$Var(\hat{P}) = E(\hat{P}^2) - \left[E(\hat{P}) \right]^2$$
$$= E(\hat{P}^2) - P^2.$$

Thus $\widehat{Var}(\hat{P}) = \hat{P}^2 -$ **Estimate of** P^2

In order to obtain an estimate of *P*², consider the expectation of

$$\frac{(m-1)(m-2)}{(n-1)(n-2)}.$$

Estimate of Variance of $\hat{P}$

In order to obtain an estimate of P^2 , consider the expectation of

$$E\left[\frac{(m-1)(m-2)}{(n-1)(n-2)}\right] = \sum_{n\geq m}\left[\frac{(m-1)(m-2)}{(n-1)(n-2)}\right]P(n)$$

$$= \frac{P(NP-1)}{N-1}\sum_{n\geq m}\left(\frac{NP-m+1}{N-n+1}\right)\left[\frac{\binom{NP-2}{m-3}\binom{NQ}{n-m}}{\binom{N-2}{n-3}}\right]$$

where the last term inside the square bracket is obtained by replacing *NP* by (*NP* - 2), *n* by (*n* - 2) and *m* by (*m* - 2) in the probability distribution function of the hypergeometric distribution.

Estimate of Variance of $\hat{P}$

This solves further to

$$E\left[\frac{(m-1)(m-2)}{(n-1)(n-2)}\right]=\frac{NP^2}{N-1}-\frac{P}{N-1}.$$

Thus an unbiased estimate of P^2 is

$$\text{Estimate of } P^2=\left(\frac{N-1}{N}\right)\frac{(m-1)(m-2)}{(n-1)(n-2)}+\frac{\hat{P}}{N}$$

$$=\left(\frac{N-1}{N}\right)\frac{(m-1)(m-2)}{(n-1)(n-2)}+\frac{1}{N}.\frac{m-1}{n-1}.$$

Estimate of variance of $\hat{P}$

Finally, an estimate of the variance of $\hat{P}$ is

$$\widehat{Var}(\hat{P}) = \hat{P}^2 - \textbf{Estimate of } P^2$$

$$= \left(\frac{m-1}{n-1}\right)^2 - \left[\frac{N-1}{N} \cdot \frac{(m-1)(m-2)}{(n-1)(n-2)} + \frac{1}{N}\left(\frac{m-1}{n-1}\right)\right]$$

$$= \left(\frac{m-1}{n-1}\right)\left[\left(\frac{m-1}{n-1}\right) + \frac{1}{N}\left(1 - \frac{(N-1)(m-2)}{n-2}\right)\right].$$

Estimate of variance of $\hat{P}$

For large *N*, the hypergeometric distribution tends to negative Binomial distribution with probability density function

$$\binom{n-1}{m-1} P^{m-1} Q^{n-m}.$$

So $\hat{P} = \dfrac{m-1}{n-1}$

and

$$\widehat{Var}(\hat{P}) = \frac{(m-1)(n-m)}{(n-1)^2(n-2)} = \frac{\hat{P}(1-\hat{P})}{n-2}.$$