Chapter 3

Sampling For Proportions and Percentages

In many situations, the characteristic under study on which the observations are collected are qualitative in nature. For example, the responses of customers in many marketing surveys are based on replies like 'yes' or 'no' , 'agree' or 'disagree' etc. Sometimes the respondents are asked to arrange several options in the order like first choice, second choice etc. Sometimes the objective of the survey is to estimate the proportion or the percentage of brown eyed persons, unemployed persons, graduate persons or persons favoring a proposal, etc. In such situations, the first question arises how to do the sampling and secondly how to estimate the population parameters like population mean, population variance, etc.

Sampling procedure:

The same sampling procedures that are used for drawing a sample in case of quantitative characteristics can also be used for drawing a sample for qualitative characteristic. So, the sampling procedures remain same irrespective of the nature of characteristic under study - either qualitative or quantitative. For example, the SRSWOR and SRSWR procedures for drawing the samples remain the same for qualitative and quantitative characteristics. Similarly, other sampling schemes like stratified sampling, two stage sampling etc. also remain same.

Estimation of population proportion:

The population proportion in case of qualitative characteristic can be estimated in a similar way as the estimation of population mean in case of quantitative characteristic.

Consider a qualitative characteristic based on which the population can be divided into two mutually exclusive classes, say C and C^*. For example, if C is the part of population of persons saying 'yes' or 'agreeing' with the proposal then C^* is the part of population of persons saying 'no' or 'disagreeing' with the proposal. Let A be the number of units in C and $(N - A)$ units in C^* be in a population of size N. Then the proportion of units in C is

$$P = \frac{A}{N}$$

and the proportion of units in C^* is

$$Q = \frac{N - A}{N} = 1 - P.$$

An indicator variable Y can be associated with the characteristic under study and then for i = 1,2,..,N

$$Y_i = \begin{cases} 1 & i^{th} \text{ unit belongs to } C \\ 0 & i^{th} \text{ unit belongs to } C^*. \end{cases}$$

Now the population total is

$$Y_{TOTAL} = \sum_{i=1}^{N} Y_i = A$$

and population mean is

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{A}{N} = P.$$

Suppose a sample of size n is drawn from a population of size N by simple random sampling .

Let a be the number of units in the sample which fall into class C and $(n-a)$ units fall in class C^*, then the sample proportion of units in C is

$$p = \frac{a}{n}.$$

which can be written as

$$p = \frac{a}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}.$$

Since $\sum_{i=1}^{N} Y_i^2 = A = NP,$ so we can write S^2 and s^2 in terms of P and Q as follows:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \\ &= \frac{1}{N-1}(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2) \\ &= \frac{1}{N-1}(NP - NP^2) \\ &= \frac{N}{N-1}PQ. \end{aligned}$$

Similarly, $\sum_{i=1}^{n} y_i^2 = a = np$ and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$= \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)$$
$$= \frac{1}{n-1}(np - np^2)$$
$$= \frac{n}{n-1}pq.$$

Note that the quantities $\bar{y}, \bar{Y}, s^2$ and S^2 have been expressed as functions of sample and population proportions. Since the sample has been drawn by simple random sampling and sample proportion is same as the sample mean, so the properties of sample proportion in SRSWOR and SRSWR can be derived using the properties of sample mean directly.

1. SRSWOR

Since sample mean $\bar{y}$ an unbiased estimator of population mean $\bar{Y}$, *i.e.* $E(\bar{y}) = \bar{Y}$ in case of SRSWOR, so

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

and p is an unbiased estimator of P.

Using the expression of $Var(\bar{y})$, the variance of p can be derived as

$$Var(p) = Var(\bar{y}) = \frac{N-n}{Nn}S^2$$
$$= \frac{N-n}{Nn}.\frac{N}{N-1}PQ$$
$$= \frac{N-n}{N-1}.\frac{PQ}{n}.$$

Similarly, using the estimate of $Var(\bar{y})$, the estimate of variance can be derived as

$$\widehat{Var}(p) = \widehat{Var}(\bar{y}) = \frac{N-n}{Nn}s^2$$
$$= \frac{N-n}{Nn}\frac{n}{n-1}pq$$
$$= \frac{N-n}{N(n-1)}pq.$$

(ii) SRSWR

Since the sample mean $\bar{y}$ is an unbiased estimator of population mean $\bar{Y}$ in case of SRSWR, so the sample proportion,

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

i.e., p is an unbiased estimator of P.

Using the expression of variance of $\bar{y}$ and its estimate in case of SRSWR, the variance of p and its estimate can be derived as follows:

$$\begin{aligned} Var(p) = Var(\bar{y}) &= \frac{N-1}{Nn} S^2 \\ &= \frac{N-1}{Nn} \frac{N}{N-1} PQ \\ &= \frac{PQ}{n} \end{aligned}$$

$$\begin{aligned} \widehat{Var}(p) &= \frac{n}{n-1} . \frac{pq}{n} \\ &= \frac{pq}{n-1}. \end{aligned}$$

Estimation of population total or total number of count

It is easy to see that 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

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

and the $100(1-\alpha)\%$ confidence interval of P is

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

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

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

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 remains the same as in the case of SRS.

Consider a situation in which the sampling units in a population are divided into two mutually exclusive classes. Let P and Q be the proportions of sampling units in the population belonging to classes '1' and '2' respectively. Then NP and NQ are the total number of sampling units in the population belonging to class '1' and '2', respectively and so $NP + NQ = N$. The probability that in a sample of n selected units out of N units by SRS such that n_1 selected units belongs to class '1' and n_2 selected units belongs to class '2' is governed by the hypergeometric distribution and

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

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 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 frequency of rare type of genes, proportion of some rare type

of cancer cells in a biopsy, proportion of rare type of blood cells affecting the red blood cells etc. In such cases, the methodology of inverse sampling can be used.

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.

Let m denotes the predetermined number indicating the number of units possessing the characteristic. The sampling is continued **till *m* number** 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.

So the probability distribution function of n can be expressed as

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

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

Note that the first term (in square brackets) is derived using hypergeometric distribution as 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 NQ units. The second term $\frac{NP-m+1}{N-n+1}$ 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 $\frac{m-1}{n-1}$.

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

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

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

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

Estimate of variance of $\hat{P}$

Now we derive an estimate of 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 $\frac{(m-1)(m-2)}{(n-1)(n-2)}$, *i.e.*,

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

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

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 hypergeometric distribution. 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

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

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

$$\begin{aligned}\widehat{Var}(\hat{P}) &= \hat{P}^2-\text{Estimate of } P^2\\ &= \left(\frac{m-1}{n-1}\right)^2-\left[\frac{N-1}{N}.\frac{(m-1)(m-2)}{(n-1)(n-2)}+\frac{1}{N}\left(\frac{m-1}{n-1}\right)\right]\\ &= \left(\frac{m-1}{n-1}\right)\left[\left(\frac{m-1}{n-1}\right)+\frac{1}{N}\left(1-\frac{(N-1)(m-2)}{n-2}\right)\right].\end{aligned}$$

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

$$\hat{P}=\frac{m-1}{n-1}$$

and

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

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. 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 storm can be classified into categories like heavily damaged, damaged, minor damage and no damage etc.

These type of situations can be represented by dividing the population of size N into, say k, mutually exclusive classes $C_1, C_2, ..., C_k$. Corresponding to these classes, let $P_1 = \frac{C_1}{N}, P_2 = \frac{C_2}{N}, ..., P_k = \frac{C_k}{n},$ be the proportions of units in the classes $C_1, C_2, ..., C_k$ respectively.

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

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

The population proportions P_i can be estimated by $p_i = \frac{c_i}{n}, i = 1, 2, ..., k.$

It can be easily shown that

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

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

and

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

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

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

The confidence intervals can be obtained based on 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

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

For this distribution

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

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

and

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