

**BSc II Year**

**Paper Code: 295**

**Survey Sampling**

**Unit 1 and Unit 2**

**UNIFIED SYLLABUS OF STATISTICS**  
**B.Sc. Part- II**

**Paper II : Survey Sampling**

**UNIT – I**

Sampling Method : Concept of population, sample, parameter and statistic, sampling versus census, advantages of sampling methods, role of sampling theory, sampling and non-sampling errors, bias and its effects, probability sampling.

**UNIT-II**

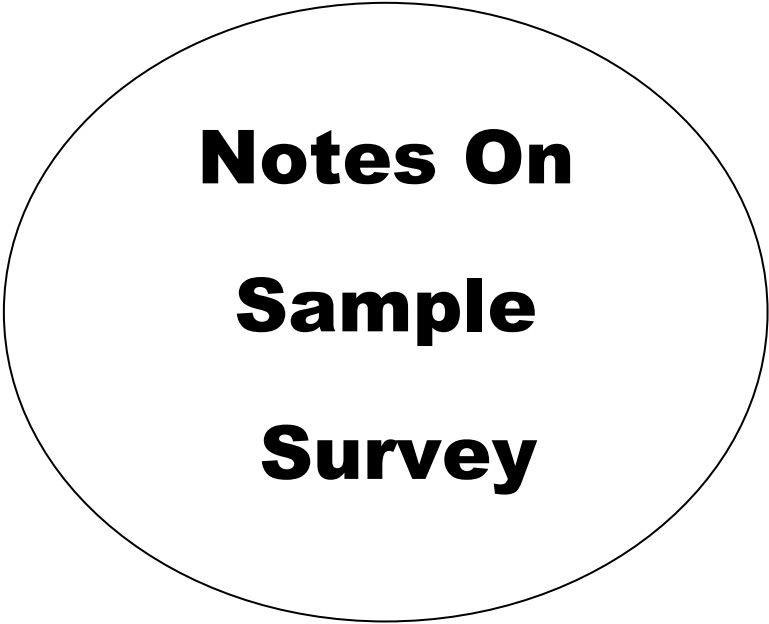
Simple Random sampling with and without replacement, use of random number tables in selection of simple random sample, estimation of population mean and proportion. Derivation of expression for variance of these estimates. Estimates of variance. Sample size determination.

**UNIT-III**

Stratified random sampling. Problem of allocation, proportional allocation, optimum allocation. Derivation of the expression for the standard errors of the usual estimators when these allocation are used. Gain in precision due to stratification.

**UNIT-IV**

Systematic sampling : estimation of population mean and population total, standard errors of these estimators. Cluster sampling with equal clusters. Estimation of population mean and their mean square error.



**Notes On  
Sample  
Survey**

# **Chapter 1**

## **Introduction**

Statistics is the science of data.

Data are the numerical values containing some information.

Statistical tools can be used on a data set to draw statistical inferences. These statistical inferences are in turn used for various purposes. For example, government uses such data for policy formulation for the welfare of the people, marketing companies use the data from consumer surveys to improve the company and to provide better services to the customer, etc. Such data is obtained through sample surveys. Sample surveys are conducted throughout the world by governmental as well as non-governmental agencies. For example, “National Sample Survey Organization (NSSO)” conducts surveys in India, “Statistics Canada” conducts surveys in Canada, agencies of United Nations like “World Health Organization (WHO), “Food and Agricultural Organization (FAO)” etc. conduct surveys in different countries.

Sampling theory provides the tools and techniques for data collection keeping in mind the objectives to be fulfilled and nature of population.

There are two ways of obtaining the information

1. **Sample surveys**
2. **Complete enumeration or census**

Sample surveys collect information on a fraction of total population whereas census collect information on whole population. Some surveys e.g., economic surveys, agricultural surveys etc. are conducted regularly. Some surveys are need based and are conducted when some need arises, e.g., consumer satisfaction surveys at a newly opened shopping mall to see the satisfaction level with the amenities provided in the mall .

**Sampling unit:**

An element or a group of elements on which the observations can be taken is called a sampling unit. The objective of the survey helps in determining the definition of sampling unit.

For example, if the objective is to determine the total income of all the persons in the household, then the sampling unit is household. If the objective is to determine the income of any particular person in the household, then the sampling unit is the income of the particular person in the household. So the definition of sampling unit depends and varies as per the objective of the survey. Similarly, in another example, if the objective is to study the blood sugar level, then the sampling unit is the value of blood sugar level of a person. On the other hand, if the objective is to study the health conditions, then the sampling unit is the person on whom the readings on the blood sugar level, blood pressure and other factors will be obtained. These values will together classify the person as healthy or unhealthy.

**Population:**

Collection of all the sampling units in a given region at a particular point of time or a particular period is called the population. For example, if the medical facilities in a hospital are to be surveyed through the patients, then the total number of patients registered in the hospital during the time period of survey will be the population. Similarly, if the production of wheat in a district is to be studied, then all the fields cultivating wheat in that district will constitute the population. The total number of sampling units in the population is the population size, denoted generally by  $N$ . The population size can be finite or infinite ( $N$  is large).

**Census:**

The complete count of population is called census. The observations on all the sampling units in the population are collected in the census. For example, in India, the census is conducted at every tenth year in which observations on all the persons staying in India is collected.

**Sample:**

One or more sampling units are selected from the population according to some specified procedure. A sample consists only of a portion of the population units. Such a collection of units is called the sample.

In the context of sample surveys, a collection of units like households, people, cities, countries etc. is called a finite population.

A census is a 100% sample and it is a complete count of the population.

### **Representative sample:**

When all the salient features of the population are present in the sample, then it is called a representative sample,

It goes without saying that every sample is considered as a representative sample.

For example, if a population has 30% males and 70% females, then we also expect the sample to have nearly 30% males and 70% females.

In another example, if we take out a handful of wheat from a 100 Kg. bag of wheat, we expect the same quality of wheat in hand as inside the bag. Similarly, it is expected that a drop of blood will give the same information as all the blood in the body.

### **Sampling frame:**

The list of all the units of the population to be surveyed constitutes the sampling frame. All the sampling units in the sampling frame have identification particulars. For example, all the students in a particular university listed along with their roll numbers constitute the sampling frame. Similarly, the list of households with the name of head of family or house address constitutes the sampling frame. In another example, the residents of a city area may be listed in more than one frame - as per automobile registration as well as the listing in the telephone directory.

### **Ways to ensure representativeness:**

There are two possible ways to ensure that the selected sample is representative.

#### **1. Random sample or probability sample:**

The selection of units in the sample from a population is governed by the laws of chance or probability.

The probability of selection of a unit can be equal as well as unequal.

## **2. Non-random sample or purposive sample:**

The selection of units in the sample from population is not governed by the probability laws.

For example, the units are selected on the basis of personal judgment of the surveyor. The persons volunteering to take some medical test or to drink a new type of coffee also constitute the sample on non-random laws.

Another type of sampling is Quota Sampling. The survey in this case is continued until a predetermined number of units with the characteristic under study are picked up.

For example, in order to conduct an experiment for rare type of disease, the survey is continued till the required number of patients with the disease are collected.

### **Advantages of sampling over complete enumeration:**

#### **1. Reduced cost and enlarged scope.**

Sampling involves the collection of data on smaller number of units in comparison to the complete enumeration, so the cost involved in the collection of information is reduced. Further, additional information can be obtained at little cost in comparison to conducting another separate survey. For example, when an interviewer is collecting information on health conditions, then he/she can also ask some questions on health practices. This will provide additional information on health practices and the cost involved will be much less than conducting an entirely new survey on health practices.

#### **2. Organizational work:**

It is easier to manage the organization of collection of smaller number of units than all the units in a census. For example, in order to draw a representative sample from a state, it is easier to manage to draw small samples from every city than drawing the sample from the whole state at a time. This ultimately results in more accuracy in the statistical inferences because better organization provides better data and in turn, improved statistical inferences are obtained.

### **3. Greater accuracy:**

The persons involved in the collection of data are trained personals. They can collect the data more accurately if they have to collect smaller number of units than large number of units.

### **4. Urgent information required:**

The data from a sample can be quickly summarized.

For example, the forecasting of the crop production can be done quickly on the basis of a sample of data than collecting first all the observation.

### **5. Feasibility:**

Conducting the experiment on smaller number of units, particularly when the units are destroyed, is more feasible. For example, in determining the life of bulbs, it is more feasible to fuse minimum number of bulbs. Similarly, in any medical experiment, it is more feasible to use less number of animals.

## **Type of surveys:**

There are various types of surveys which are conducted on the basis of the objectives to be fulfilled.

### **1. Demographic surveys:**

These surveys are conducted to collect the demographic data, e.g., household surveys, family size, number of males in families, etc. Such surveys are useful in the policy formulation for any city, state or country for the welfare of the people.

### **2. Educational surveys:**

These surveys are conducted to collect the educational data, e.g., how many children go to school, how many persons are graduate, etc. Such surveys are conducted to examine the educational programs in schools and colleges. Generally, schools are selected first and then the students from each school constitute the sample.



### **3. Economic surveys:**

These surveys are conducted to collect the economic data, e.g., data related to export and import of goods, industrial production, consumer expenditure etc. Such data is helpful in constructing the indices indicating the growth in a particular sector of economy or even the overall economic growth of the country.

### **4. Employment surveys:**

These surveys are conducted to collect the employment related data, e.g., employment rate, labour conditions, wages, etc. in a city, state or country. Such data helps in constructing various indices to know the employment conditions among the people.

### **5. Health and nutrition surveys:**

These surveys are conducted to collect the data related to health and nutrition issues, e.g., number of visits to doctors, food given to children, nutritional value etc. Such surveys are conducted in cities, states as well as countries by the national and international organizations like UNICEF, WHO etc.

### **6. Agricultural surveys:**

These surveys are conducted to collect the agriculture related data to estimate, e.g., the acreage and production of crops, livestock numbers, use of fertilizers, use of pesticides and other related topics. The government bases its planning related to the food issues for the people based on such surveys.

### **7. Marketing surveys:**

These surveys are conducted to collect the data related to marketing. They are conducted by major companies, manufacturers or those who provide services to consumer etc. Such data is used for knowing the satisfaction and opinion of consumers as well as in developing the sales, purchase and promotional activities etc.

### **8. Election surveys:**

These surveys are conducted to study the outcome of an election or a poll. For example, such polls are conducted in democratic countries to have the opinions of people about any candidate who is contesting the election.

## **9. Public polls and surveys:**

These surveys are conducted to collect the public opinion on any particular issue. Such surveys are generally conducted by the news media and the agencies which conduct polls and surveys on the current topics of interest to public.

## **10. Campus surveys:**

These surveys are conducted on the students of any educational institution to study about the educational programs, living facilities, dining facilities, sports activities, etc.

### **Principal steps in a sample survey:**

The broad steps to conduct any sample surveys are as follows:

#### **1. Objective of the survey:**

The objective of the survey has to be clearly defined and well understood by the person planning to conduct it. It is expected from the statistician to be well versed with the issues to be addressed in consultation with the person who wants to get the survey conducted. In complex surveys, sometimes the objective is forgotten and data is collected on those issues which are far away from the objectives.

#### **2. Population to be sampled:**

Based on the objectives of the survey, decide the population from which the information can be obtained. For example, population of farmers is to be sampled for an agricultural survey whereas the population of patients has to be sampled for determining the medical facilities in a hospital.

#### **3. Data to be collected:**

It is important to decide that which data is relevant for fulfilling the objectives of the survey and to note that no essential data is omitted. Sometimes, too many questions are asked and some of their outcomes are never utilized. This lowers the quality of the responses and in turn results in lower efficiency in the statistical inferences.

#### **4. Degree of precision required:**

The results of any sample survey are always subjected to some uncertainty. Such uncertainty can be reduced by taking larger samples or using superior instruments. This involves more cost and more time. So it is very important to decide about the required degree of precision in the data. This needs to be conveyed to the surveyor also.

#### **5. Method of measurement:**

The choice of measuring instrument and the method to measure the data from the population needs to be specified clearly. For example, the data has to be collected through interview, questionnaire, personal visit, combination of any of these approaches, etc. The forms in which the data is to be recorded so that the data can be transferred to mechanical equipment for easily creating the data summary etc. is also needed to be prepared accordingly.

#### **6. The frame:**

The sampling frame has to be clearly specified. The population is divided into sampling units such that the units cover the whole population and every sampling unit is tagged with identification. The list of all sampling units is called the frame. The frame must cover the whole population and the units must not overlap each other in the sense that every element in the population must belong to one and only one unit. For example, the sampling unit can be an individual member in the family or the whole family.

#### **7. Selection of sample:**

The size of the sample needs to be specified for the given sampling plan. This helps in determining and comparing the relative cost and time of different sampling plans. The method and plan adopted for drawing a representative sample should also be detailed.

#### **8. The Pre-test:**

It is advised to try the questionnaire and field methods on a small scale. This may reveal some troubles and problems beforehand which the surveyor may face in the field in large scale surveys.

## **9. Organization of the field work:**

How to conduct the survey, how to handle business administrative issues, providing proper training to surveyors, procedures, plans for handling the non-response and missing observations etc. are some of the issues which need to be addressed for organizing the survey work in the fields. The procedure for early checking of the quality of return should be prescribed. It should be clarified how to handle the situation when the respondent is not available.

## **10. Summary and analysis of data:**

It is to be noted that based on the objectives of the data, the suitable statistical tool is decided which can answer the relevant questions. In order to use the statistical tool, a valid data set is required and this dictates the choice of responses to be obtained for the questions in the questionnaire, e.g., the data has to be qualitative, quantitative, nominal, ordinal etc. After getting the completed questionnaire back, it needs to be edited to amend the recording errors and delete the erroneous data. The tabulating procedures, methods of estimation and tolerable amount of error in the estimation needs to be decided before the start of survey. Different methods of estimation may be available to get the answer of the same query from the same data set. So the data needs to be collected which is compatible with the chosen estimation procedure.

## **11. Information gained for future surveys:**

The completed surveys work as guide for improved sample surveys in future. Beside this they also supply various types of prior information required to use various statistical tools, e.g., mean, variance, nature of variability, cost involved etc. Any completed sample survey acts as a potential guide for the surveys to be conducted in the future. It is generally seen that the things always do not go in the same way in any complex survey as planned earlier. Such precautions and alerts help in avoiding the mistakes in the execution of future surveys.

### **Variability control in sample surveys:**

The variability control is an important issue in any statistical analysis. A general objective is to draw statistical inferences with minimum variability. There are various types of sampling schemes which are adopted in different conditions. These schemes help in controlling the variability at different stages. Such sampling schemes can be classified in the following way.

#### **1. Before selection of sampling units**

- Stratified sampling
- Cluster sampling
- Two stage sampling
- Double sampling etc.

#### **2. At the time of selection of sampling units**

- Systematic sampling
- Varying probability sampling

#### **3. After the selection of sampling units**

- Ratio method of estimation
- Regression method of estimation

*Note that the ratio and regression methods are the methods of estimation and not the methods of drawing samples.*

### **Methods of data collection**

There are various way of data collection. Some of them are as follows:

#### **1. Physical observations and measurements:**

The surveyor contacts the respondent personally through the meeting. He observes the sampling unit and records the data. The surveyor can always use his prior experience to collect the data in a better way. For example, a young man telling his age as 60 years can easily be observed and corrected by the surveyor.

## **2. Personal interview:**

The surveyor is supplied with a well prepared questionnaire. The surveyor goes to the respondents and asks the same questions mentioned in the questionnaire. The data in the questionnaire is then filled up accordingly based on the responses from the respondents.

## **3. Mail enquiry:**

The well prepared questionnaire is sent to the respondents through postal mail, e-mail, etc. The respondents are requested to fill up the questionnaires and send it back. In case of postal mail, many times the questionnaires are accompanied by a self addressed envelope with postage stamps to avoid any non-response due to the cost of postage.

## **4. Web based enquiry:**

The survey is conducted online through internet based web pages. There are various websites which provide such facility. The questionnaires are to be in their formats and the link is sent to the respondents through email. By clicking on the link, the respondent is brought to the concerned website and the answers are to be given online. These answers are recorded and responses as well as their statistics is sent to the surveyor. The respondents should have internet connection to support the data collection with this procedure.

## **5. Registration:**

The respondent is required to register the data at some designated place. For example, the number of births and deaths along with the details provided by the family members are recorded at city municipal office which are provided by the family members.

## **6. Transcription from records:**

The sample of data is collected from the already recorded information. For example, the details of the number of persons in different families or number of births/deaths in a city can be obtained from the city municipal office directly.

The methods in (1) to (5) provide primary data which means collecting the data directly from the source. The method in (6) provides the secondary data which means getting the data from the primary sources.

## **Chapter -2**

### **Simple Random Sampling**

Simple random sampling (SRS) is a method of selection of a sample comprising of  $n$  number of sampling units out of the population having  $N$  number of sampling units such that every sampling unit has an equal chance of being chosen.

The samples can be drawn in two possible ways.

- The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population .
- The sampling units are chosen with replacement in the sense that the chosen units are placed back in the population.

#### **1. Simple random sampling without replacement (SRSWOR):**

SRSWOR is a method of selection of  $n$  units out of the  $N$  units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected, i.e.  $1/N$ .

#### **2. Simple random sampling with replacement (SRSWR):**

SRSWR is a method of selection of  $n$  units out of the  $N$  units one by one such that at each stage of selection each unit has equal chance of being selected, i.e.,  $1/N$ .

#### **Procedure of selection of a random sample:**

The procedure of selection of a random sample follows the following steps:

1. Identify the  $N$  units in the population with the numbers 1 to  $N$ .
2. Choose any random number arbitrarily in the random number table and start reading numbers.
3. Choose the sampling unit whose serial number corresponds to the random number drawn from the table of random numbers.
4. In case of SRSWR, all the random numbers are accepted even if repeated more than once.

In case of SRSWOR, if any random number is repeated, then it is ignored and more numbers are drawn.

Such process can be implemented through programming and using the discrete uniform distribution. Any number between 1 and  $N$  can be generated from this distribution and corresponding unit can be selected into the sample by associating an index with each sampling unit. Many statistical softwares like R, SAS, etc. have inbuilt functions for drawing a sample using SRSWOR or SRSWR.

## Notations:

The following notations will be used in further notes:

$N$ : Number of sampling units in the population (Population size).

$n$ : Number of sampling units in the sample (sample size)

$Y$ : The characteristic under consideration

$Y_i$ : Value of the characteristic for the  $i^{\text{th}}$  unit of the population

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i : \text{sample mean}$$

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i : \text{population mean}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N-1} (\sum_{i=1}^N Y_i^2 - N\bar{Y}^2)$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N} (\sum_{i=1}^N Y_i^2 - N\bar{Y}^2)$$

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

## Probability of drawing a sample :

### 1.SRSWOR:

If  $n$  units are selected by SRSWOR, the total number of possible samples are  $\binom{N}{n}$ .

So the probability of selecting any one of these samples is  $\frac{1}{\binom{N}{n}}$ .

Note that a unit can be selected at any one of the  $n$  draws. Let  $u_i$  be the  $i^{\text{th}}$  unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, ..., or  $n^{\text{th}}$  draw.



Let  $P_j(i)$  denotes the probability of selection of  $u_i$  at the  $j^{\text{th}}$  draw,  $j = 1, 2, \dots, n$ . Then

$$\begin{aligned} P_j(i) &= P_1(i) + P_2(i) + \dots + P_n(i) \\ &= \frac{1}{N} + \frac{1}{N} + \dots + \frac{1}{N} \quad (n \text{ times}) \\ &= \frac{n}{N} \end{aligned}$$

Now if  $u_1, u_2, \dots, u_n$  are the  $n$  units selected in the sample, then the probability of their selection is

$$P(u_1, u_2, \dots, u_n) = P(u_1) \cdot P(u_2) \cdot \dots \cdot P(u_n)$$

Note that when the second unit is to be selected, then there are  $(n - 1)$  units left to be selected in the sample from the population of  $(N - 1)$  units. Similarly, when the third unit is to be selected, then there are  $(n - 2)$  units left to be selected in the sample from the population of  $(N - 2)$  units and so on.

If  $P(u_1) = \frac{n}{N}$ , then

$$P(u_2) = \frac{n-1}{N-1}, \dots, P(u_n) = \frac{1}{N-n+1}.$$

Thus

$$P(u_1, u_2, \dots, u_n) = \frac{n}{N} \cdot \frac{n-1}{N-1} \cdot \frac{n-2}{N-2} \cdots \frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}.$$

### Alternative approach:

The probability of drawing a sample in SRSWOR can alternatively be found as follows:

Let  $u_{i(k)}$  denotes the  $i^{\text{th}}$  unit drawn at the  $k^{\text{th}}$  draw. Note that the  $i^{\text{th}}$  unit can be any unit out of the  $N$  units. Then  $s_o = (u_{i(1)}, u_{i(2)}, \dots, u_{i(n)})$  is an ordered sample in which the order of the units in which they are drawn, i.e.,  $u_{i(1)}$  drawn at the first draw,  $u_{i(2)}$  drawn at the second draw and so on, is also considered. The probability of selection of such an ordered sample is

$$P(s_o) = P(u_{i(1)})P(u_{i(2)} | u_{i(1)})P(u_{i(3)} | u_{i(1)}u_{i(2)}) \dots P(u_{i(n)} | u_{i(1)}u_{i(2)} \dots u_{i(n-1)}).$$

Here  $P(u_{i(k)} | u_{i(1)}u_{i(2)} \dots u_{i(k-1)})$  is the probability of drawing  $u_{i(k)}$  at the  $k^{\text{th}}$  draw given that  $u_{i(1)}, u_{i(2)}, \dots, u_{i(k-1)}$  have already been drawn in the first  $(k - 1)$  draws.

Such probability is obtained as

$$P(u_{i(k)} | u_{i(1)}u_{i(2)}\dots u_{i(k-1)}) = \frac{1}{N - k + 1}.$$

So

$$P(s_o) = \prod_{k=1}^n \frac{1}{N - k + 1} = \frac{(N - n)!}{N!}.$$

The number of ways in which a sample of size  $n$  can be drawn =  $n!$

Probability of drawing a sample in a given order =  $\frac{(N - n)!}{N!}$

So the probability of drawing a sample in which the order of units in which they are drawn is

$$\text{irrelevant} = n! \frac{(N - n)!}{N!} = \frac{1}{\binom{N}{n}}.$$

## 2. SRSWR

When  $n$  units are selected with SRSWR, the total number of possible samples are  $N^n$ . The

Probability of drawing a sample is  $\frac{1}{N^n}$ .

Alternatively, let  $u_i$  be the  $i^{\text{th}}$  unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, ..., or  $n^{\text{th}}$  draw. At any stage, there are always  $N$  units in the population in case of SRSWR, so the probability of selection of  $u_i$  at any stage is  $1/N$  for all  $i = 1, 2, \dots, n$ . Then the probability of selection of  $n$  units  $u_1, u_2, \dots, u_n$  in the sample is

$$\begin{aligned} P(u_1, u_2, \dots, u_n) &= P(u_1) \cdot P(u_2) \dots P(u_n) \\ &= \frac{1}{N} \cdot \frac{1}{N} \dots \frac{1}{N} \\ &= \frac{1}{N^n} \end{aligned}$$

## Probability of drawing an unit

### 1. SRSWOR

Let  $A_\ell$  denotes an event that a particular unit  $u_j$  is not selected at the  $\ell^{\text{th}}$  draw. The probability of selecting, say,  $j^{\text{th}}$  unit at  $k^{\text{th}}$  draw is

$$\begin{aligned} P(\text{selection of } u_j \text{ at } k^{\text{th}} \text{ draw}) &= P(A_1 \cap A_2 \cap \dots \cap A_{k-1} \cap \bar{A}_k) \\ &= P(A_1)P(A_2 | A_1)P(A_3 | A_1 A_2) \dots P(A_{k-1} | A_1, A_2, \dots, A_{k-2})P(\bar{A}_k | A_1, A_2, \dots, A_{k-1}) \\ &= \left(1 - \frac{1}{N}\right) \left(1 - \frac{1}{N-1}\right) \left(1 - \frac{1}{N-2}\right) \dots \left(1 - \frac{1}{N-k+2}\right) \frac{1}{N-k+1} \\ &= \frac{N-1}{N} \cdot \frac{N-2}{N-1} \dots \frac{N-k+1}{N-k+2} \cdot \frac{1}{N-k+1} \\ &= \frac{1}{N} \end{aligned}$$

### 2. SRSWR

$$P[\text{selection of } u_j \text{ at } k^{\text{th}} \text{ draw}] = \frac{1}{N}.$$

## Estimation of population mean and population variance

One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatterdness of the data around the central value. Among various indicators of central tendency and dispersion, the popular choices are arithmetic mean and variance. So the population mean and population variability are generally measured by the arithmetic mean (or weighted arithmetic mean) and variance, respectively. There are various popular estimators for estimating the population mean and population variance. Among them, sample arithmetic mean and sample variance are more popular than other estimators. One of the reason to use these estimators is that they possess nice statistical properties. Moreover, they are also obtained through well established statistical estimation procedures like maximum likelihood estimation, least squares estimation, method of moments etc. under several standard statistical distributions. One may also consider other indicators like median, mode, geometric mean, harmonic mean for measuring the central tendency and mean deviation, absolute deviation, Pitman nearness etc. for measuring the dispersion. The properties of such estimators can be studied by numerical procedures like bootstrapping.

## 1. Estimation of population mean

Let us consider the sample arithmetic mean  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  as an estimator of population mean

$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$  and verify  $\bar{y}$  is an unbiased estimator of  $\bar{Y}$  under the two cases.

### SRSWOR

Let  $t_i = \sum_{i=1}^n y_i$ . Then

$$\begin{aligned} E(\bar{y}) &= \frac{1}{n} E\left(\sum_{i=1}^n y_i\right) \\ &= \frac{1}{n} E(t_i) \\ &= \frac{1}{n} \left( \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} t_i \right) \\ &= \frac{1}{n} \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \left( \sum_{i=1}^n y_i \right). \end{aligned}$$

When  $n$  units are sampled from  $N$  units by without replacement, then each unit of the population can occur with other units selected out of the remaining  $(N-1)$  units is the population and each unit

occurs in  $\binom{N-1}{n-1}$  of the  $\binom{N}{n}$  possible samples. So

$$\text{So } \sum_{i=1}^{\binom{N}{n}} \left( \sum_{i=1}^n y_i \right) = \binom{N-1}{n-1} \sum_{i=1}^N y_i.$$

Now

$$\begin{aligned} E(\bar{y}) &= \frac{(N-1)!}{(n-1)!(N-n)!} \frac{n!(N-n)!}{nN!} \sum_{i=1}^N y_i \\ &= \frac{1}{N} \sum_{i=1}^N y_i \\ &= \bar{Y}. \end{aligned}$$

Thus  $\bar{y}$  is an unbiased estimator of  $\bar{Y}$ . Alternatively, the following approach can also be adopted to show the unbiasedness property.

$$\begin{aligned}
 E(\bar{y}) &= \frac{1}{n} \sum_{j=1}^n E(y_j) \\
 &= \frac{1}{n} \sum_{j=1}^n \left[ \sum_{i=1}^N Y_i P_j(i) \right] \\
 &= \frac{1}{n} \sum_{j=1}^n \left[ \sum_{i=1}^N Y_i \cdot \frac{1}{N} \right] \\
 &= \frac{1}{n} \sum_{j=1}^n \bar{Y} \\
 &= \bar{Y}
 \end{aligned}$$

where  $P_j(i)$  denotes the probability of selection of  $i^{\text{th}}$  unit at  $j^{\text{th}}$  stage.

### SRSWR

$$\begin{aligned}
 E(\bar{y}) &= \frac{1}{n} E\left(\sum_{i=1}^n y_i\right) \\
 &= \frac{1}{n} \sum_{i=1}^n E(y_i) \\
 &= \frac{1}{n} \sum_{i=1}^n (Y_1 P_i + \dots + Y_N P_i) \\
 &= \frac{1}{n} \sum_{i=1}^n \bar{Y} \\
 &= \bar{Y}.
 \end{aligned}$$

where  $P_i = \frac{1}{N}$  for all  $i = 1, 2, \dots, N$  is the probability of selection of a unit. Thus  $\bar{y}$  is an unbiased estimator of population mean under SRSWR also.

## Variance of the estimate

Assume that each observation has some variance  $\sigma^2$ . Then

$$\begin{aligned}
 V(\bar{y}) &= E(\bar{y} - \bar{Y})^2 \\
 &= E\left[\frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})\right]^2 \\
 &= E\left[\frac{1}{n^2} \sum_{i=1}^n (y_i - \bar{Y})^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i}^n (y_i - \bar{Y})(y_j - \bar{Y})\right] \\
 &= \frac{1}{n^2} \sum_{i=1}^n E(y_i - \bar{Y})^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i}^n E(y_i - \bar{Y})(y_j - \bar{Y}) \\
 &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 + \frac{K}{n^2} \\
 &= \frac{N-1}{Nn} S^2 + \frac{K}{n^2}
 \end{aligned}$$

where  $K = \sum_{i=1}^n \sum_{j \neq i}^n E(y_i - \bar{Y})(y_j - \bar{Y})$  assuming that each observation has variance  $\sigma^2$ . Now we find

$K$  under the setups of SRSWR and SRSWOR.

## SRSWOR

$$K = \sum_{i=1}^n \sum_{j \neq i}^n E(y_i - \bar{Y})(y_j - \bar{Y}).$$

Consider

$$E(y_i - \bar{Y})(y_j - \bar{Y}) = \frac{1}{N(N-1)} \sum_{k=1}^N \sum_{\ell \neq k}^N (y_k - \bar{Y})(y_\ell - \bar{Y})$$

Since

$$\begin{aligned}
 \left[ \sum_{k=1}^N (y_k - \bar{Y}) \right]^2 &= \sum_{i=1}^N (y_i - \bar{Y})^2 + \sum_{k=1}^N \sum_{\ell \neq k}^N (y_k - \bar{Y})(y_\ell - \bar{Y}) \\
 0 &= (N-1)S^2 + \sum_{k=1}^N \sum_{\ell \neq k}^N (y_k - \bar{Y})(y_\ell - \bar{Y}) \\
 \sum_{k=1}^N \sum_{\ell \neq k}^N (y_k - \bar{Y})(y_\ell - \bar{Y}) &= \frac{1}{N(N-1)} [-(N-1)S^2] \\
 &= -\frac{S^2}{N}.
 \end{aligned}$$

Thus  $K = -n(n-1)\frac{S^2}{N}$  and so substituting the value of  $K$ , the variance of  $\bar{y}$  under SRSWOR is

$$\begin{aligned} V(\bar{y}_{WOR}) &= \frac{N-1}{Nn} S^2 - \frac{1}{n^2} n(n-1) \frac{S^2}{N} \\ &= \frac{N-n}{Nn} S^2. \end{aligned}$$

## SRSWR

$$\begin{aligned} K &= \sum_{i \neq j}^N \sum_{j \neq i}^N E(y_i - \bar{Y})(y_j - \bar{Y}) \\ &= \sum_{i \neq j}^N \sum_{j \neq i}^N E(y_i - \bar{Y})E(y_j - \bar{Y}) \\ &= 0 \end{aligned}$$

because the  $i$ th and  $j$ th draws ( $i \neq j$ ) are independent.

Thus the variance of  $\bar{y}$  under SRSWR is

$$V(\bar{y}_{WR}) = \frac{N-1}{Nn} S^2.$$

It is to be noted that if  $N$  is infinite (large enough), then

$$V(\bar{y}) = \frac{S^2}{n}$$

is both the cases of SRSWOR and SRSWR. So the factor  $\frac{N-n}{N}$  is responsible for changing the variance of  $\bar{y}$  when the sample is drawn from a finite population in comparison to an infinite population. This is why  $\frac{N-n}{N}$  is called a finite population correction (fpc). It may be noted that

$\frac{N-n}{N} = 1 - \frac{n}{N}$ , so  $\frac{N-n}{N}$  is close to 1 if the ratio of sample size to population  $\frac{n}{N}$ , is very small or

negligible. The term  $\frac{n}{N}$  is called sampling fraction. In practice, fpc can be ignored whenever

$\frac{n}{N} < 5\%$  and for many purposes even if it is as high as 10%. Ignoring fpc will result in the overestimation of variance of  $\bar{y}$ .

## Efficiency of $\bar{y}$ under SRSWOR over SRSWR

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

$$\begin{aligned} V(\bar{y}_{WR}) &= \frac{N-1}{Nn} S^2 \\ &= \frac{N-n}{Nn} S^2 + \frac{n-1}{Nn} S^2 \\ &= V(\bar{y}_{WOR}) + a \text{ positive quantity} \end{aligned}$$

Thus

$$V(\bar{y}_{WR}) > V(\bar{y}_{WOR})$$

and so, SRSWOR is more efficient than SRSWR.

## Estimation of variance from a sample

Since the expressions of variances of sample mean involve  $S^2$  which is based on population values, so these expressions can not be used in real life applications. In order to estimate the variance of  $\bar{y}$  on the basis of a sample, an estimator of  $S^2$  (or equivalently  $\sigma^2$ ) is needed. Consider  $s^2$  as an estimator of  $S^2$  (or  $\sigma^2$ ) and we investigate its biasedness for  $S^2$  in the cases of SRSWOR and SRSWR,

Consider

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 - n(\bar{y} - \bar{Y})^2 \right] \end{aligned}$$

$$\begin{aligned} E(s^2) &= \frac{1}{n-1} \left[ \sum_{i=1}^n E(y_i - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2 \right] \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n \text{Var}(y_i) - n\text{Var}(\bar{y}) \right] = \frac{1}{n-1} [n\sigma^2 - n\text{Var}(\bar{y})] \end{aligned}$$



**In case of SRSWOR**

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

and so

$$\begin{aligned} E(s^2) &= \frac{n}{n-1} \left[ \sigma^2 - \frac{N-n}{Nn} S^2 \right] \\ &= \frac{n}{n-1} \left[ \frac{N-1}{N} S^2 - \frac{N-n}{Nn} S^2 \right] \\ &= S^2 \end{aligned}$$

**In case of SRSWR**

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

and so

$$\begin{aligned} E(s^2) &= \frac{n}{n-1} \left[ \sigma^2 - \frac{N-n}{Nn} S^2 \right] \\ &= \frac{n}{n-1} \left[ \frac{N-1}{N} S^2 - \frac{N-n}{Nn} S^2 \right] \\ &= \frac{N-1}{N} S^2 \\ &= \sigma^2 \end{aligned}$$

Hence

$$E(s^2) = \begin{cases} S^2 & \text{is SRSWOR} \\ \sigma^2 & \text{is SRSWR} \end{cases}$$

An unbiased estimate of  $Var(\bar{y})$  is

$$\hat{V}(\bar{y}_{WOR}) = \frac{N-n}{Nn} s^2 \quad \text{in case of SRSWOR and}$$

$$\begin{aligned} \hat{V}(\bar{y}_{WR}) &= \frac{N-1}{Nn} \cdot \frac{N}{N-1} s^2 \\ &= \frac{s^2}{n} \quad \text{in case of SRSWR.} \end{aligned}$$

## Standard errors

The standard error of  $\bar{y}$  is defined as  $\sqrt{\text{Var}(\bar{y})}$ .

In order to estimate the standard error, one simple option is to consider the square root of estimate of variance of sample mean.

- under SRSWOR, a possible estimator is  $\hat{\sigma}(\bar{y}) = \sqrt{\frac{N-n}{Nn}}s$ .
- under SRSWR, a possible estimator is  $\hat{\sigma}(\bar{y}) = \sqrt{\frac{N-1}{Nn}}s$ .

It is to be noted that this estimator does not possess the same properties as of  $\widehat{\text{Var}}(\bar{y})$ .

Reason being if  $\hat{\theta}$  is an estimator of  $\theta$ , then  $\sqrt{\hat{\theta}}$  is not necessarily an estimator of  $\sqrt{\theta}$ .

In fact, the  $\hat{\sigma}(\bar{y})$  is a negatively biased estimator under SRSWOR.

The approximate expressions for large  $N$  case are as follows:

(Reference: Sampling Theory of Surveys with Applications, P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme, C. Asok, Iowa State University Press and Indian Society of Agricultural Statistics, 1984, India)

Consider  $s$  as an estimator of  $S$ .

Let

$$s^2 = S^2 + \varepsilon \text{ with } E(\varepsilon) = 0, E(\varepsilon^2) = S^2.$$

Write

$$\begin{aligned} s &= (S^2 + \varepsilon)^{1/2} \\ &= S \left( 1 + \frac{\varepsilon}{S^2} \right)^{1/2} \\ &= S \left( 1 + \frac{\varepsilon}{2S^2} - \frac{\varepsilon^2}{8S^4} + \dots \right) \end{aligned}$$

assuming  $\varepsilon$  will be small as compared to  $S^2$  and as  $n$  becomes large, the probability of such an event approaches one. Neglecting the powers of  $\varepsilon$  higher than two and taking expectation, we have

$$E(s) = \left[ 1 - \frac{\text{Var}(s^2)}{8S^4} \right] S$$

where

$$\text{Var}(s^2) = \frac{2S^4}{(n-1)} \left[ 1 + \left( \frac{n-1}{2n} \right) (\beta_2 - 3) \right] \text{ for large } N.$$

$$\mu_j = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^j$$

$$\beta_2 = \frac{\mu_4}{S^4} : \text{coefficient of kurtosis.}$$

Thus

$$\begin{aligned} E(s) &= S \left[ 1 - \frac{1}{4(n-1)} - \frac{\beta_2 - 3}{8n} \right] \\ \text{Var}(s) &= S^2 - S^2 \left[ 1 - \frac{1}{8} \frac{\text{Var}(s^2)}{S^4} \right]^2 \\ &= \frac{\text{Var}(s^2)}{4S^2} \\ &= \frac{S^2}{2(n-1)} \left[ 1 + \left( \frac{n-1}{2n} \right) (\beta_2 - 3) \right]. \end{aligned}$$

Note that for a normal distribution,  $\beta_2 = 3$  and we obtain

$$\text{Var}(s) = \frac{S^2}{2(n-1)}.$$

Both  $\text{Var}(s)$  and  $\text{Var}(s^2)$  are inflated due to nonnormality to the same extent, by the inflation factor

$$\left[ 1 + \left( \frac{n-1}{2n} \right) (\beta_2 - 3) \right]$$

and this does not depend on coefficient of skewness.

This is an important result to be kept in mind while determining the sample size in which it is assumed that  $S^2$  is known. If inflation factor is ignored and population is non-normal, then the reliability on  $s^2$  may be misleading.

### Alternative approach:

The results for the unbiasedness property and the variance of sample mean can also be proved in an alternative way as follows:

#### (i) SRSWOR

With the  $i^{th}$  unit of the population, we associate a random variable  $a_i$  defined as follows:

$$a_i = \begin{cases} 1, & \text{if the } i^{th} \text{ unit occurs in the sample} \\ 0, & \text{if the } i^{th} \text{ unit does not occurs in the sample } (i=1,2,\dots,N) \end{cases}$$

Then,

$$E(a_i) = 1 \times \text{Probability that the } i^{th} \text{ unit is included in the sample}$$

$$= \frac{n}{N}, \quad i=1,2,\dots,N.$$

$$E(a_i^2) = 1 \times \text{Probability that the } i^{th} \text{ unit is included in the sample}$$

$$= \frac{n}{N}, \quad i=1,2,\dots,N$$

$$E(a_i a_j) = 1 \times \text{Probability that the } i^{th} \text{ and } j^{th} \text{ units are included in the sample}$$

$$= \frac{n(n-1)}{N(N-1)}, \quad i \neq j = 1,2,\dots,N.$$

From these results, we can obtain

$$\text{Var}(a_i) = E(a_i^2) - (E(a_i))^2 = \frac{n(N-n)}{N^2}, \quad i=1,2,\dots,N$$

$$\text{Cov}(a_i, a_j) = E(a_i a_j) - E(a_i)E(a_j) = \frac{n(N-n)}{N^2(N-1)}, \quad i \neq j = 1,2,\dots,N.$$

We can rewrite the sample mean as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N a_i y_i$$

Then

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N E(a_i) y_i = \bar{Y}$$

and

$$\text{Var}(\bar{y}) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^N a_i y_i \right) = \frac{1}{n^2} \left[ \sum_{i=1}^N \text{Var}(a_i) y_i^2 + \sum_{i \neq j}^N \text{Cov}(a_i, a_j) y_i y_j \right].$$

Substituting the values of  $Var(a_i)$  and  $Cov(a_i, a_j)$  in the expression of  $Var(\bar{y})$  and simplifying, we get

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

To show that  $E(s^2) = S^2$ , consider

$$s^2 = \frac{1}{(n-1)} \left[ \sum_{i=1}^n y_i^2 - n\bar{y}^2 \right] = \frac{1}{(n-1)} \left[ \sum_{i=1}^N a_i y_i^2 - n\bar{y}^2 \right].$$

Hence, taking, expectation, we get

$$E(s^2) = \frac{1}{(n-1)} \left[ \sum_{i=1}^N E(a_i) y_i^2 - n \{ Var(\bar{y}) + \bar{Y}^2 \} \right]$$

Substituting the values of  $E(a_i)$  and  $Var(\bar{y})$  in this expression and simplifying, we get  $E(s^2) = S^2$ .

## (ii) SRSWR

Let a random variable  $a_i$  associated with the  $i^{th}$  unit of the population denotes the number of times the  $i^{th}$  unit occurs in the sample  $i=1,2,\dots,N$ . So  $a_i$  assumes values  $0, 1, 2,\dots,n$ . The joint distribution of  $a_1, a_2, \dots, a_N$  is the multinomial distribution given by

$$P(a_1, a_2, \dots, a_N) = \frac{n!}{\prod_{i=1}^N a_i!} \cdot \frac{1}{N^n}$$

where  $\sum_{i=1}^N a_i = n$ . For this multinomial distribution, we have

$$E(a_i) = \frac{n}{N},$$

$$Var(a_i) = \frac{n(N-1)}{N^2}, \quad i=1,2,\dots,N.$$

$$Cov(a_i, a_j) = -\frac{n}{N^2}, \quad i \neq j=1,2,\dots,N.$$

We rewrite the sample mean as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N a_i y_i.$$

Hence, taking expectation of  $\bar{y}$  and substituting the value of  $E(a_i) = n/N$  we obtain that

$$E(\bar{y}) = \bar{Y}.$$

Further,

$$\text{Var}(\bar{y}) = \frac{1}{n^2} \left[ \sum_{i=1}^N \text{Var}(a_i) y_i^2 + \sum_{i=1}^N \text{Cov}(a_i, a_j) y_i y_j \right]$$

Substituting, the values of  $\text{Var}(a_i) = n(N-1)/N^2$  and  $\text{Cov}(a_i, a_j) = -n/N^2$  and simplifying, we get

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

To prove that  $E(s^2) = \frac{N-1}{N} S^2 = \sigma^2$  in SRSWR, consider

$$(n-1)s^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^N a_i y_i^2 - n\bar{y}^2,$$

$$\begin{aligned} (n-1)E(s^2) &= \sum_{i=1}^N E(a_i) y_i^2 - n \{ \text{Var}(\bar{y}) + \bar{Y}^2 \} \\ &= \frac{n}{N} \sum_{i=1}^N y_i^2 - n \cdot \frac{(N-1)}{nN} S^2 - n\bar{Y}^2 \\ &= \frac{(n-1)(N-1)}{N} S^2 \end{aligned}$$

$$E(s^2) = \frac{N-1}{N} S^2 = \sigma^2$$

### Estimator of population total:

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

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

which can be estimated by

$$\begin{aligned} \hat{Y}_T &= N\hat{\bar{Y}} \\ &= N\bar{y}. \end{aligned}$$

Obviously

$$\begin{aligned}
 E(\hat{Y}_T) &= NE(\bar{y}) \\
 &= N\bar{Y} \\
 \text{Var}(\hat{Y}_T) &= N^2(\bar{y}) \\
 &= \begin{cases} N^2 \left( \frac{N-n}{Nn} \right) S^2 = \frac{N(N-n)}{n} S^2 & \text{for SRSWOR} \\ N^2 \left( \frac{N-1}{Nn} \right) S^2 = \frac{N(N-1)}{n} S^2 & \text{for SRSWOR} \end{cases}
 \end{aligned}$$

and the estimates of variance of  $\hat{Y}_T$  are

$$\widehat{\text{Var}}(\hat{Y}_T) = \begin{cases} \frac{N(N-n)}{n} s^2 & \text{for SRSWOR} \\ \frac{N}{n} s^2 & \text{for SRSWOR} \end{cases}$$

### Confidence limits for the population mean

Now we construct the  $100(1-\alpha)\%$  confidence interval for the population mean. Assume that the population is normally distributed  $N(\mu, \sigma^2)$  with mean  $\mu$  and variance  $\sigma^2$ . then  $\frac{\bar{y} - \bar{Y}}{\sqrt{\text{Var}(\bar{y})}}$

follows  $N(0,1)$  when  $\sigma^2$  is known. If  $\sigma^2$  is unknown and is estimated from the sample then

$\frac{\bar{y} - \bar{Y}}{\sqrt{\text{Var}(\bar{y})}}$  follows a  $t$ -distribution with  $(n-1)$  degrees of freedom. When  $\sigma^2$  is known, then the

$100(1-\alpha)\%$  confidence interval is given by

$$\begin{aligned}
 P \left[ -Z_{\frac{\alpha}{2}} \leq \frac{\bar{y} - \bar{Y}}{\sqrt{\text{Var}(\bar{y})}} \leq Z_{\frac{\alpha}{2}} \right] &= 1 - \alpha \\
 \text{or } P \left[ \bar{y} - Z_{\frac{\alpha}{2}} \sqrt{\text{Var}(\bar{y})} \leq \bar{y} \leq \bar{y} + Z_{\frac{\alpha}{2}} \sqrt{\text{Var}(\bar{y})} \right] &= 1 - \alpha
 \end{aligned}$$

and the confidence limits are

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

when  $Z_{\frac{\alpha}{2}}$  denotes the upper  $\frac{\alpha}{2}$  % points on  $N(0,1)$  distribution. Similarly, when  $\sigma^2$  is unknown,

then the  $100(1-\alpha)$  % confidence interval is

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

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

and the confidence limits are

$$\left[\bar{y} - t_{\frac{\alpha}{2}} \sqrt{\text{Var}\hat{(\bar{y})}} \leq \bar{y} \leq \bar{y} + t_{\frac{\alpha}{2}} \sqrt{\text{Var}\hat{(\bar{y})}}\right]$$

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

## Determination of sample size

The size of the sample is needed before the survey starts and goes into operation. One point to be kept in mind is that 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. The sample size can be determined on the basis of prescribed values of standard error of sample mean, error of estimation, width of the confidence interval, coefficient of variation of sample mean, relative error of sample mean or total cost among several others.

An important constraint or need to determine the sample size is that the information regarding the population standard deviation  $S$  should be known for these criteria. The reason and need for this will be clear when we derive the sample size in the next section. A question arises about how to have information about  $S$  beforehand? The possible solutions to this issue are to conduct a pilot survey and collect a preliminary sample of small size, estimate  $S$  and use it as known value of  $S$  it. Alternatively, such information can also be collected from past data, past experience, 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

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

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

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

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

$$\text{or } \frac{1}{n} - \frac{1}{N} \leq \frac{1}{n_e}$$

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

$$\text{where } n_e = \frac{S^2}{V}$$

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

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

If  $N$  is large, then the required  $n$  is

$$n \geq n_e \text{ and } n_{\text{smallest}} = n_e$$

## 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 requirement can be satisfied by associating a probability  $(1 - \alpha)$  with it and can be expressed as

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

Since  $\bar{y}$  follows  $N(\bar{Y}, \frac{N-n}{Nn} S^2)$  assuming the normal distribution for the population, we can write

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

which implies that

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

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

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

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

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

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

### 3. Prespecified width of confidence interval

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

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

assuming  $\sigma^2$  is known and population is normally distributed. This can be expressed as

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

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

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

$$\text{or } n \geq \frac{\frac{4Z_{\frac{\alpha}{2}}^2 S^2}{W^2}}{1 + \frac{W^2}{NW^2}}$$

The minimum sample size required is

$$n_{\text{smallest}} = \frac{\frac{4Z_{\frac{\alpha}{2}}^2 S^2}{W^2}}{1 + \frac{W^2}{NW^2}}$$

If  $N$  is large then

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

and the minimum sample size needed is

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

#### 4. Prespecified 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 the coefficient of variation of  $\bar{y}$  should not exceed a given or prespecified value of coefficient of variation, say  $C_0$ , then the required sample size  $n$  is to be determined such that

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

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

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

$$\text{or } \frac{1}{n} - \frac{1}{N} \leq \frac{C_0^2}{C^2}$$

$$\text{or } n \geq \frac{\frac{C^2}{C_0^2}}{1 + \frac{C^2}{NC_0^2}}$$

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

The smallest sample size needed in this case is

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

If  $N$  is large, then

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

$$\text{and } n_{\text{smallest}} = \frac{C^2}{C_0^2}$$

## 5. Prespecified relative error

When  $\bar{y}$  is used for estimating the population mean  $\bar{Y}$ , then the relative estimation error is defined as  $\frac{\bar{y} - \bar{Y}}{\bar{Y}}$ . If it is required that such relative estimation error should not exceed a prespecified value

$R$  with probability  $(1 - \alpha)$ , then such requirement can be satisfied by expressing it like such requirement can be satisfied by expressing it like

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

Assuming the population to be normally distributed,  $\bar{y}$  follows  $N\left(\bar{Y}, \frac{N-n}{Nn} S^2\right)$ .

So it can be written that

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

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

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

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

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. Prespecified cost

Let an amount of money  $C$  is being designated for sample survey to called  $n$  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

$$C = C_0 + nC_1$$

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

is the required sample size.

## 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  $i = 1, 2, \dots, N$

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

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

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

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

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

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

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

### (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:

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

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

### 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

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

and the estimate of variance is

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

### Confidence interval estimation of $P$

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

can write

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

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

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

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

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

### 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 remain 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

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

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

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  $\frac{m-1}{n-1}$ .

$$\begin{aligned} E\left(\frac{m-1}{n-1}\right) &= \sum_{n=m}^{m+NQ} \left(\frac{m-1}{n-1}\right) P(n) \\ &= \sum_{n=m}^{m+NQ} \left(\frac{m-1}{n-1}\right) \frac{\binom{NP}{m-1} \binom{NQ}{n-m}}{\binom{N}{n-1}} \cdot \frac{Np-m+1}{N-n+1} \\ &= \sum_{n=m}^{m+NQ-1} \left(\frac{NP-m+1}{N-n+1}\right) \frac{\binom{NP-1}{m-2} \binom{NQ}{n-m}}{\binom{N-1}{n-2}} \end{aligned}$$

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

$$\begin{aligned} \text{Var}(\hat{P}) &= E(\hat{P}^2) - [E(\hat{P})]^2 \\ &= E(\hat{P}^2) - P^2. \end{aligned}$$

Thus

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

In order to obtain an estimate of  $P^2$ , consider the expectation of  $\frac{(m-1)(m-2)}{(n-1)(n-2)}$ , i.e.,

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

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

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

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

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

For large  $N$ , the hypergeometric distribution tends to negative Binomial distribution with

probability density function  $\binom{n-1}{m-1} P^m Q^{n-m}$ . So

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

and

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

## 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  $C_1, C_2, \dots, C_k$ . Corresponding to these classes, let  $P_1 = \frac{C_1}{N}, P_2 = \frac{C_2}{N}, \dots, P_k = \frac{C_k}{N}$ , be the proportions of units in the classes  $C_1, C_2, \dots, C_k$  respectively.

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

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

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

It can be easily shown that

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

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

and

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

For estimating the number of units in the  $i^{\text{th}}$  class,

$$\hat{C}_i = N p_i$$

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

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  $c_1, c_2, \dots, c_k$  can be approximated by multinomial distribution given by

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

For this distribution

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

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

and

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