CHAPTER 1

Preliminaries and Basics of Probability Sampling

1.1 INTRODUCTION

Various government organizations, researchers, sociologist, and businesses often conduct surveys to get answers to certain specific questions, which cannot be obtained merely through laboratory experiments or simply using economic, mathematical, or statistical formulation. For example, the knowledge of the proportion of unemployed people, those below poverty line, and the extent of child labor in a certain locality is very important for the formulation of a proper economic planning. To get the answers to such questions, we conduct surveys on sections of people of the locality very often. Surveys should be conducted in such a way that the results of the surveys can be interpreted objectively in terms of probability. Drawing inference about aggregate (population) on the basis of a sample, a part of the populations, is a natural instinct of human beings. Surveys should be conducted in such a way that the inference relating to the population should have some valid statistical background. To achieve valid statistical inferences, one needs to select samples using some suitable sampling procedure. The collected data should be analyzed appropriately. In this book, we have discussed various methods of sample selection procedures, data collection, and methods of data analysis and their applications under various circumstances. The statistical theories behind such procedures have also been studied in great detail.

In this chapter we introduce some of the basic definitions and terminologies in survey sampling such as population, unit, sample, sampling designs, and sampling schemes. Various methods of sample selection as well as Hanurav's algorithm which gives the correspondence between a sampling design and a sampling scheme have also been discussed.

1.2 DEFINITIONS AND TERMINOLOGIES

1.2.1 Population and Unit

A population is an aggregate or collection of elements or objects in a certain region at a particular point in time and is often a subject of study. Each

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00001-7
Copyright © 2017 Raghunath Arnab.
Published by Elsevier Ltd.
All rights reserved.

element of the population is called a unit. Suppose we want to study the prevalence of HIV in the province of KwaZulu-Natal in 2016, the collection of all individuals, i.e., male or female and child or adult, residing in KwaZulu-Natal will be termed as population and each individual will be called a unit. Suppose we consider air pollution in a certain region. In this case, the air under consideration constitutes the population, but we cannot divide it into identifiable parts or elements. This type of population is called a continuous population.

1.2.2 Finite and Infinite Populations

A finite population is a collection of a finite number of identifiable units. The total number of elements will be denoted by N and refers to the size of the population. The students in a class, tigers in a game park, and households in a certain locality are examples of finite population as the units are identifiable and finite in number. Bacteria in a test tube, however, are identifiable, but they are very large in number. In this case $N \rightarrow \infty$, and hence it is considered an infinite population. The size of the population may be known or unknown before a survey. Sometimes, surveys are conducted to determine the unknown population size N, such as the total number of illegal immigrants or certain kinds of animals in a game park.

1.2.3 Sampling Frame

It is a list of all the units of a population with proper identification. The list is the basic material for conducting a survey. So, the sampling frame must be complete, up-to-date and free from duplication or omission of units. We denote a list of finite population or sampling frame as

$$U = (u_1, \ldots, u_i, \ldots, u_N)$$

where $u_i(i = 1, \ldots, N)$ is the ith unit of the population U. For simplicity we will denote the population U as

$$U = \{1, \ldots, i, \ldots, N\} \tag{1.2.1}$$

1.2.4 Parameter and Parameter Space

For a given population U, we may be interested in studying certain characteristics of it. Such characteristics are known as study variables. When considering a population of students in a certain class, we may be interested to know the age, height, racial group, economic condition, marks on different subjects, and so forth. Each of the variables under study is called a study variable, and it will be denoted by y. Let y_i be the value of a study variable y for the ith unit of the population U, which is generally not known before the survey. The N-dimension vector $\mathbf{y} = (y_1,\ldots, y_i,\ldots, y_N)$ is known as a parameter of the population U with respect to the characteristic y. The set of all-possible values of the vector $\mathbf{y}$ is the N-dimensional Euclidean space $R^N = (-\infty < y_1 < \infty,\ldots,-\infty < y_i < \infty,\ldots,-\infty < y_N < \infty)$ and it is known as a parameter space. In most of the cases we are not interested in knowing the parameter $\mathbf{y}$ but in a certain parametric function of $\mathbf{y}$ such as, $Y = \sum_{i=1}^{N} y_i =$ population total, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i =$ population mean, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 =$ population variance, $C_y = S_y/\overline{Y}$ $=$ population coefficient of variation, and so forth.

1.2.5 Complete Enumeration and Sample Survey

To know the value of a parameter or parametric function for a certain study variable y, we can follow two routes. The first route is to survey all the elements of the population and get all the values of y_i's, $i = 1,\ldots, N$. The second route is to select only a part of the population, which is termed as a sample. Then survey all the selected units in the sample and obtain the y-values from the selected units. From the y-values obtained in the sample, we predict (estimate) the population parameter under consideration. The first route is known as a complete enumeration or census, whereas the second route is called a sample survey.

1.2.6 Sampling and Nonsampling Errors

Obviously, using the complete enumeration method, we get the correct value of the parameter, provided all the y-values of the population obtained are correct. This would mean that there is no nonresponse, i.e., a response from each unit is obtained, and there is no measurement error in measuring y-values. However, in practice, at least for a large-scale survey, nonresponse is unavoidable, and y-values are also subject to error because the respondents report untrue values, especially when y-values relate to confidential characteristics such as income and age. The error in a survey, which is originated from nonresponse or incorrect measurement of y-values, is termed as the nonsampling error. The nonsampling errors increase with the sample size.

From a sample survey, we cannot get the true value of the parameter because we surveyed only a sample, which is just a part of the population. The error committed by making inference by surveying a part of the population is known as the sampling error. In complete enumeration, sampling error is absent, but it is subjected more to nonsampling error than sample surveys. When the population is large, complete enumeration is not possible as it is very expensive, time-consuming, and requires many trained investigators. The advantages of sample surveys over complete enumeration were advocated by Mahalanobis (1946), Cochran (1977), and Murthy (1977), to name a few.

1.2.7 Sample

A sample is an ordered sequence of elements from a population U, and it will be denoted by $s = (i_1, \ldots, i_j, \ldots, i_{n_s})$, where $i_j \in U$. The units in s need not be distinct and they may be repeated. The number of units in s, including repetition, is called the size of the sample s and will be denoted by n_s. The number of the distinct units in s is known as effective sample size and will be denoted by $\nu(s)$.

Example 1.2.1

Let $U = (1, 2, 3, 4)$ be a population of size 4, then $s = (1, 1, 2)$ is a sample of size $n_s = 3$ and effective sample size $\nu(s) = 2$.

1.2.8 Probability and Purposive Sampling

In probability sampling, a sample is selected according to a certain rule or method (known as sampling design), where each sample has a definite preassigned probability of selection. In purposive sampling or subjective sampling, the selection of sample is subjective; it totally depends on the choice of the sampler. Thus probability sampling reduces to purposive sampling when the probability of selection of a particular sample is assigned to 1.

1.3 SAMPLING DESIGN AND INCLUSION PROBABILITIES

1.3.1 Sampling Design

Let $\mathfrak{I}$ be the collection of all possible samples s. A sampling design p is a function defined on $\mathfrak{I}$, which satisfies the following conditions: (i) $p(s) \geq 0 \ \forall \ s \in \mathfrak{I}$ and (ii) $\sum_{s \in \mathfrak{I}} p(s) = 1$.

Example 1.3.1

Consider a finite population $U = (1, 2, 3, 4)$. Let $s_1 = (1, 1, 2)$, $s_2 = (1, 2, 2)$, $s_3 = (3, 2)$, and $s_4 = (4)$ be the possible samples and their respective probabilities are $p(s_1) = 0.25$, $p(s_2) = 0.30$, $p(s_3) = 0.20$, and $p(s_4) = 0.25$.

Here $\mathfrak{I} = (s_1, s_2, s_3, s_4)$ and p is a sampling design selecting the sample s_j with probability $p(s_j)$ for $j = 1, 2, 3, 4$.

1.3.2 Inclusion Probabilities

The inclusion probability of the unit i is the probability of inclusion of the unit i in any sample with respect to the sampling design p and will be denoted by π_i. Thus,

$$\pi_i = \Pr ob(i \in s) = \sum_{s \supset i} p(s) = \sum_{s \in \mathfrak{I}} I_{si} p(s)$$

where $I_{si} = 1$ if $i \in s$ and $I_{si} = 0$ if $i \notin s$ and $s \supset i$ denotes the sum over the samples containing the ith unit. Similarly, inclusion probability for the ith and jth unit $(i \neq j)$ is denoted by

$$\pi_{ij} = \sum_{s \supset i,j} p(s) = \sum_{s \in \mathfrak{I}} I_{si} I_{sj} p(s) = \pi_{ji}$$

The inclusion probabilities π_i and π_{ij} are called first- and second-order inclusion probabilities, respectively. The higher order inclusion probabilities are defined similarly. For the sake of convenience, we write $\pi_{ii} = \pi_i$.

1.3.3 Consistency Conditions of Inclusion Probabilities

The consistency conditions of the inclusion probabilities obtained by Godambe (1955) and Hanurav (1966) are given in the following theorem:

Theorem 1.3.1

(i) Godambe (1955): $\sum_{i=1}^{N} \pi_i = \nu$

(ii) Hanurav (1966): $\sum_{i \neq}^{N} \sum_{j=1}^{N} \pi_{ij} = V_p(\nu(s)) + \nu(\nu - 1)$

where $\nu = E_p(\nu(s)) =$ expected effective sample size for the design $p = \sum_{s \in \mathfrak{I}} \nu(s) p(s)$, and $V_p(\cdot)$ is the variance with respect to the design p.

Proof

(i)
$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} \left\{ \sum_{s \in \mathfrak{I}} I_{si} p(s) \right\}$$
$$= \sum_{s \in \mathfrak{I}} p(s) \sum_{i=1}^{N} I_{si}$$

Now noting $\sum_{i=1}^{N} I_{si} = \nu(s) =$ number of distinct units in s, we find

$$\sum_{i=1}^{N} \pi_i = \sum_{s \in \mathfrak{I}} p(s) \nu(s)$$
$$= \nu$$

$$
\begin{aligned}
\text{(ii)} \quad \sum_{i\neq}^{N}\sum_{j=1}^{N}\pi_{ij} &= \sum_{i=1}^{N}\left\{\sum_{j(\neq i)}^{N}\sum_{s\in\mathfrak{S}} I_{si}I_{sj}p(s)\right\} \\
&= \sum_{s\in\mathfrak{S}} p(s)\sum_{i=1}^{N} I_{si}\sum_{j(\neq i)}^{N} I_{sj} \\
&= \sum_{s\in\mathfrak{S}} p(s)\sum_{i=1}^{N} I_{si}(\nu(s)-I_{si}) \\
&= \sum_{s\in\mathfrak{S}} p(s)\left\{(\nu(s))^2-\nu(s)\right\} \\
&= E_p(\nu(s))^2 - E_p(\nu(s)) \\
&= V_p(\nu(s)) + \nu(\nu-1)
\end{aligned}
$$

1.3.4 Fixed Effective Size Design

The number of distinct units in a sample s is known as the effective sample size and is denoted by ν(s). A sampling design for which all the samples with positive probability have exactly n distinct units, i.e., $P(\nu(\mathrm{s})=n)=1$ is known as a fixed effective size (n) sampling design (FESD(n)). The class of fixed effective size sampling designs will be denoted by $\mathscr{P}_n$.

1.3.5 Fixed Sample Size Design

A sampling design p is said to be a fixed sample size (FSS) design if $p\{n_s=n\}=1$, i.e., sample size n_s is fixed as n for each of the samples s with $p(s)>0, s\in\mathfrak{S}$.

Corollary 1.3.1 (Yates and Grundy, 1953)

For a fixed effective size (ν), sampling design p, $V_p(\nu(s))=0$ and in this case, Theorem 1.3.1 yields

$$
\text{(i)} \quad \sum_{i=1}^{N}\pi_i=\nu \quad \text{and} \quad \text{(ii)} \quad \sum_{i\neq}^{N}\sum_{j}^{N}\pi_{ij}=\nu(\nu-1) \tag{1.3.1}
$$

Corollary 1.3.2

For a fixed effective size (ν) design

$$
\sum_{j(\neq i)}^{N}\pi_{ij}=(\nu-1)\pi_i \tag{1.3.2}
$$

Proof

$$\sum_{j(\neq i)}^{N} \pi_{ij} = \sum_{j(\neq i)}^{N} \sum_{s \in \mathfrak{S}} I_{si} I_{sj} p(s) = \sum_{s \in \mathfrak{S}} I_{si} \sum_{j \neq i}^{N} I_{sj} p(s)$$
$$= \sum_{s \in \mathfrak{S}} I_{si}(\nu - I_{si}) p(s) = (\nu - 1)\pi_i$$

Example 1.3.2

Consider Example 1.3.1. Here the first-order inclusion probabilities for the units 1, 2, 3 and 4 are $\pi_1 = p(s_1) + p(s_2) = 0.55$, $\pi_2 = p(s_1) + p(s_2) + p(s_3) = 0.75$, $\pi_3 = p(s_3) = 0.20$, and $\pi_4 = p(s_4) = 0.25$, respectively. The second-order probabilities are $\pi_{12} = p(s_1) + p(s_2) = 0.55$, $\pi_{13} = \pi_{14} = 0$, $\pi_{23} = p(s_3) = 0.20$, and $\pi_{24} = \pi_{34} = 0$. The expectation and variance of the effective sample size are obtained as follows:

s	s_1	s_2	s_3	s_4	Total
$p(s)$	0.25	0.30	0.20	0.25	1.00
$\nu(s)$	2	2	2	1	
$\nu(s)p(s)$	0.50	0.60	0.40	0.25	1.75
$\nu(s)^2 p(s)$	1.0	1.20	0.80	0.25	3.25

(i) $E_p(\nu(s)) = \nu = 1.75$, (ii) $V_p(\nu(s)) = 3.25 - (1.75)^2 = 0.1875$. Here we can easily verify the consistency conditions $\sum_{i=1}^{4} \pi_i = 1.75 = \nu$ and $\sum_{i\neq}^{4} \sum_{j=1}^{4} \pi_{ij} = 1.50 = \nu(\nu - 1) + V_p(\nu(s))$.

1.4 METHODS OF SELECTION OF SAMPLE

We can use the following two methods of selection of sample.

1.4.1 Cumulative Total Method

Here we label all possible samples of $\mathfrak{S}$ as $s_1, \ldots, s_i, \ldots, s_M$, where $M =$ total number of samples in $\mathfrak{S}$. Then we calculate the cumulative total $T_i = p(s_1) + \cdots + p(s_i)$ for $i = 1, \ldots, M$ and select a random sample R (say) from a uniform population with range (0, 1). This can be done by choosing a five-digit random number and placing a decimal preceding it. The sample s_k is selected if $T_{k-1} < R \leq T_k$, for $k = 1, \ldots, M$ with $T_0 = 0$.

Example 1.4.1

Let $U = (1, 2, 3, 4)$; $s_1 = (1, 1, 2)$, $s_2 = (1, 2, 2)$, $s_3 = (3, 2)$, $s_4 = (4)$; $p(s_1) = 0.25$, $p(s_2) = 0.30$, $p(s_3) = 0.20$, and $p(s_4) = 0.25$.

s	s_1	s_2	s_3	s_4
$p(s)$	0.25	0.30	0.20	0.25
T_k	0.25	0.55	0.75	1

Let a random sample $R = 0.34802$ be selected from a uniform population with range (0, 1). The sample s_2 is selected as $T_1 = 0.25 < R = 0.34802 \le T_2 = 0.55$.

The cumulative total method mentioned above, however, cannot be used in practice because here we have to list all the possible samples having positive probabilities. For example, suppose we need to select a sample of size 15 from a population size $R = 30$ following a sampling design, where all possible samples of size $n = 15$ have positive probabilities, we need to list $M = \binom{30}{15}$ possible samples, which is obviously a huge number.

1.4.2 Sampling Scheme

In a sampling scheme, we select units one by one from the population by using a preassigned set of probabilities of selection of a unit in a particular draw. For a fixed sample of size n (FSS(n)) design, we select the ith unit at kth draw with probability $p_i(k)$ for $k = 1,\ldots, n$; $i = 1,\ldots, N$. $p_i(k)$'s are subject to

$$0 \le p_i(k) \le 1 \text{ and } \sum_{i=1}^{N} p_i(k) = 1; k = 1, \ldots, n, i = 1, \ldots, N \qquad (1.4.1)$$

There are various sampling schemes available in literature. We have given some of FSS designs, which are commonly used in practice below.

1.4.3 With and Without Replacement Sampling

In a with replacement (WR) sampling scheme, a unit may occur more than once in a sample with positive probability, whereas in a without replacement (WOR) sampling scheme, all the units of the sample are distinct, i.e., no unit is repeated in a sample with positive probability.

1.4.4 Simple Random Sampling With Replacement

In a simple random sampling WR (SRSWR) sampling scheme, $p_i(k) = 1/N$ for $k = 1,\ldots, n$. So, for an SRSWR, the probability of selection of a unit at any draw is the same and is equal to $1/N$. Hence the probability of selecting i_1 at the first draw, i_2 at the second draw, and i_n at the nth draw is

$$p(i_1, i_2, \ldots, i_n) = 1/N^n \;\; \text{for} \;\; 1 \leq i_1, i_2, \ldots, i_n \leq N \tag{1.4.2}$$

1.4.5 Simple Random Sampling Without Replacement

In a simple random sampling WOR (SRSWOR),

$$p_i(k) = \begin{cases} \dfrac{1}{N-(k-1)} & \text{if } i\text{th unit is not selected in first } k-1 \text{ draws,} \\ & \quad k = 1, \ldots, n \\ 0 & \text{if the } i\text{th unit is selected in first } k-1 \text{ draws} \end{cases} \tag{1.4.3}$$

So, under SRSWOR, the probabilities of selecting units i_1 at the first draw, $i_2 (i_2 \neq i_1)$ at the second draw, and $i_n (i_n \neq i_{n-1} \neq \cdots \neq i_1)$ at the nth draw are $1/N$, $1/(N-1)$, and $1/(N-n+1)$, respectively. So the probability of selection of such a sample (i_1, i_2,…, i_n) is

$$\begin{aligned} p(i_1, i_2, \ldots, i_n) = 1/\{N(N-1)\cdots(N-n+1)\} \;\; \text{for} \\ 1 \leq i_1 \neq i_2 \neq \cdots \neq i_n \leq N \end{aligned} \tag{1.4.4}$$

1.4.6 Probability Proportional to Size With Replacement Sampling

For a probability proportional to size WR (PPSWR) sampling scheme, the probability of selecting the ith unit at any draw is $p_i\left(p_i \geq 0, \sum_{i=1}^{N} p_i = 1\right)$, which is called the normed size measure for the ith unit. So for a PPSWR sampling scheme, $p_i(k) = p_i$ for $k = 1,\ldots, n$; $i = 1,\ldots, N$. Hence the probability of selecting i_1 at the first draw, i_2 at the second draw, and i_n at the nth draw under a PPSWR sampling scheme is

$$p(i_1, i_2, \ldots, i_n) = p_{i_1} p_{i_2} \ldots p_{i_n} \;\; \text{for} \;\; 1 \leq i_1, i_2, \ldots, i_n \leq N \tag{1.4.5}$$

Clearly the PPSWR sampling scheme reduces to SRSWR sampling scheme if $p_i = 1/N$ for $i = 1,\ldots, N$.

1.4.7 Probability Proportional to Size Without Replacement Sampling

In probability proportional to size WOR (PPSWOR) sampling scheme, probability of selection of i_1 at the first draw is $p_{i_1}(1) = p_{i_1}$. Probability of selecting i_2 at the second draw is $p_{i_2}(2) = \dfrac{p_{i_2}}{1 - p_{i_1}}$ if the unit $i_1 (i_2 \neq i_1)$ is selected at the first draw and $p_{i_2}(2) = 0$ when the unit i_2 is selected at the first draw, i.e., $i_2 = i_1$. In general, the probability of selection of i_k at the kth draw is $p_{i_k}(k) = \dfrac{p_{i_k}}{1 - p_{i_1} - p_{i_2} - \cdots - p_{i_{k-1}}}$, if the units $i_1, i_2,\ldots, i_{k-1}$ are selected in any of the first $k - 1$ draws and $p_{i_k}(k) = 0$ if the unit i_k is selected in any of the first $k - 1$ draws for $k = 2,\ldots, n$; $i = 1,\ldots, N$. So, for a PPSWOR sampling scheme, the probability of selecting i_1 at the first draw, i_2 at the second draw, and i_n at the nth draw is

$$p(i_1, \ldots, i_n) = p_{i_1}\frac{p_{i_2}}{1 - p_{i_1}}\cdots\frac{p_{i_k}}{1 - p_{i_1} - \cdots - p_{i_{k-1}}}\cdots\frac{p_{i_n}}{1 - p_{i_1} - \cdots - p_{i_{n-1}}} \text{ for } 1 \leq i_1 \neq i_2 \neq \cdots \neq i_n \leq N \tag{1.4.6}$$

It should be noted that PPSWOR reduces to SRSWOR sampling scheme if $p_i = 1/N$ for $i = 1,\ldots, N$.

1.4.8 Lahiri–Midzuno–Sen Sampling Scheme

In Lahiri (1951)–Midzuno (1952)–Sen (1953) (LMS) sampling scheme, at the first draw ith unit is selected with a normed size measure p_i, after which the remaining $n - 1$ units are selected by the SRSWOR method from units not selected at the first draw, i.e., where $p_{i_1}(1) = p_{i_1}$ and $p_{i_j}(k) = \dfrac{1}{(N-1)\cdots(N-k+1)}$ for $k = 2,\ldots, n$ if the unit i_j is not selected in earlier $k - 1$ draws, otherwise $p_{i_j}(k) = 0$. Thus the probability of selecting i_1 at the first draw, i_2 at the second draw, and i_n at the nth draw under the LMS sampling scheme is

$$p(i_1, i_2, \ldots, i_n) = p_{i_1}\cdot\frac{1}{N-1}\cdots\frac{1}{N-n+1} \quad \text{for } 1 \leq i_1 \neq i_2 \neq \cdots \neq i_n \leq N \tag{1.4.7}$$

The LMS sampling scheme reduces to SRSWOR sampling scheme if $p_i = 1/N$ for every $i = 1, \ldots, N$.

1.5 HANURAV'S ALGORITHM

Hanurav (1966) established a correspondence between a sampling design and a sampling scheme. He proved that any sampling scheme results in a sampling design. Similarly, for a given sampling design, one can construct at least one sampling scheme, which can implement the sampling design. In fact, Hanurav proposed the most general sampling scheme, known as Hanurav's algorithm, using which one can derive various types of sampling schemes or sampling designs. Henceforth, we will not differentiate between the terms "sampling design" and "sampling scheme".

Let n_0 denote the maximum sample size that might be required from a sampling scheme. Then, Hanurav's (1966) algorithm is defined as follows:

$$\mathscr{A} = \mathscr{A}\{q_1(i); q_2(s); q_3(s,i)\} \tag{1.5.1}$$

where

(i) $0 \le q_1(i) \le 1, \quad \sum_{i=1}^{N} q_1(i) = 1$ for $i = 1, \ldots, N$

(ii) $0 \le q_2(s) \le 1$ for any sample $s \in \mathfrak{J}$, where $\mathfrak{J}$ be the set of all possible samples.

(iii) $q_3(s, i)$ is defined when $q_2(s) > 0$ and subject to $0 \le q_3(s,i) \le 1$, $\sum_{i=1}^{N} q_3(s,i) = 1$ for $i = 1, \ldots, N$

Samples are selected using the following steps:

Step 1: At the first draw a unit i_1 is selected with probability $q_1(i_1)$; $i_1 = 1, \ldots, N$

Step 2: In this step, we decide whether the sampling procedure will be terminated or continued. Let $s_{(1)} = i_1$ be the unit selected in the first draw. A Bernoulli trial is performed with success probability $q_2(s_{(1)})$. If the trial results in a failure, the sampling procedure is terminated and the selected sample is $s_{(1)} = i_1$. On the other hand, if the trial results in a success, we go to step 3.

Step 3: In this step, a second unit i_2 is selected with probability $q_3(s_{(1)}, i_2)$ and we denote $s_{(2)} = (i_1, i_2)$. After selection of the sample $s_{(2)}$, we go back to step 2 and perform a Bernoulli trial with success probability $q_2(s_{(2)})$. If the trial results in a failure, then the sample procedure is terminated and the selected sample is $s_{(2)}$. Otherwise, another unit i_3 is selected with probability $q_3(s_{(2)}, i_3)$, and we denote $s_{(3)} = (i_1, i_2, i_3)$ as the selected sample. This procedure is

continued until the sampling procedure is terminated. The sampling procedure is terminated with probability 1 after a selection of a sample of size n_0, as we assign $q_2\left(s_{(n_0)}\right) = 0$.
The probability of selection of a sample $s_{(n)} = (i_1, \ldots, i_n)$ using the algorithm $\mathscr{A}$ is

$$p\left(s_{(n)}\right) = q_1(i_1)\cdot q_2(i_1)\cdot q_3(i_1, i_2)\cdot q_2(i_1, i_2)\cdot\cdots\cdot q_3(i_1, i_2\ldots, i_n)\cdot \{1 - q_2(i_1, i_2\ldots, i_n)\}$$

Corollary 1.5.1

Hanurav's (1966) algorithm reduces to an FSS (n) sampling scheme, if

$$q_2(i_1) = q_2(i_1, i_2) = \cdots = q_2(i_1, i_2\ldots, i_{n-1}) = 1 \text{ and } q_2(i_1, i_2\ldots, i_n) = 0 \text{ for } i_1, i_2, \ldots, i_n = 1, 2, \ldots, N.$$

The following examples show that (i) SRSWR, (ii) SRSWOR, (iii) PPSWR, (iv) PPSWOR, and (v) LMS sampling schemes are particular cases of Hanurav's algorithm.

Example 1.5.1

SRSWR of size n:

Here we choose (i) $q_1(i_1) = 1/N$, (ii) $q_2(i_1) = \cdots = q_2(i_1, i_2, \ldots, i_{n-1}) = 1$ and $q_2(i_1, i_2, \ldots, i_n) = 0$ for $i_1, i_2, \ldots, i_n = 1, 2, \ldots, N$, and (iii) $q_3(s, i) = 1/N$ for $i = 1, \ldots, N$.

Example 1.5.2

SRSWOR of size n:

Here we choose (i) $q_1(i_1) = 1/N$, (ii) $q_2(i_1) = \cdots = q_2(i_1, i_2, \ldots, i_{n-1}) = 1$ and $q_2(i_1, i_2, \ldots, i_n) = 0$ for $i_1, i_2, \ldots, i_n = 1, 2, \ldots, N$, and (iii) $q_3(s, i) = 1/(N - k)$ for $i = 1, \ldots, N$ if $s = (i_1, \ldots, i_k)$ does not contain the unit i otherwise $q_3(s, i) = 0$.

Example 1.5.3

PPSWR of size n:

Here we choose (i) $q_1(i_1) = p_{i_1}$, (ii) $q_2(i_1) = \cdots = q_2(i_1, i_2, \ldots, i_{n-1}) = 1$ and $q_2(i_1, i_2, \ldots, i_n) = 0$ for $i_1, i_2, \ldots, i_n = 1, 2, \ldots, N$, and (iii) $q_3(s, i) = p_i$ for $i = 1, \ldots, N$.

Example 1.5.4

PPSWOR of size n:

Here we choose (i) $q_1(i_1) = p_{i_1}$, (ii) $q_2(i_1) = \cdots = q_2(i_1, i_2, \ldots, i_{n-1}) = 1$ and $q_2(i_1, i_2, \ldots, i_n) = 0$ for $i_1, i_2, \ldots, i_n = 1, 2, \ldots, N$, and (iii) $q_3(s, i) = p_i/(1 - p_{i_1} - p_{i_2} - \cdots - p_{i_{k-1}})$ for $i = 1, \ldots, N$, if $s = (i_1, \ldots, i_{k-1})$ does not contain the unit i and $i_1 \neq \cdots \neq i_{k-1}$; $q_3(s, i) = 0$ if s contains i.

Example 1.5.5

LMS of size n:

Here we choose (i) $q_1(i_1) = p_{i_1}$, (ii) $q_2(i_1) = \cdots = q_2(i_1, i_2, \ldots, i_{n-1}) = 1$ and $q_2(i_1, i_2, \ldots, i_n) = 0$ for $i_1, i_2, \ldots, i_n = 1, 2, \ldots, N$, and (iii) $q_3(s, i) = 1/(N - k)$ for $i = 1, \ldots, N$, if $s = (i_1, \ldots, i_{k-1})$ does not contain the unit i and $i_1 \neq \cdots \neq i_{k-1}$; $q_3(s, i) = 0$, if s contains i.

A correspondence between a sampling design and a sampling scheme is given in the following theorem:

Theorem 1.5.1

(i) Sampling according to the algorithm $\mathscr{A} = \mathscr{A}\{q_1(i); q_2(s); q_3(s, i)\}$ results in a sampling design.

(ii) For a given sampling design p, there exists an algorithm $\mathscr{A}$, which results in the design p.

Proof

(i) Here we have to show $\sum_{s \in \mathfrak{I}} p(s) = 1$.

Let $\mathfrak{I}_k$ = collection of all samples whose size is k. Then,

$$\mathfrak{I} = \bigcup_{k=1}^{n_0} \mathfrak{I}_k \text{ and } \sum_{s \in \mathfrak{I}} p(s) = \sum_{k=1}^{n_0} \sum_{s \in \mathfrak{I}_k} p(s). \tag{1.5.2}$$

Now,

$$\sum_{s \in \mathfrak{I}_1} p(s) = \sum_{i_1=1}^{N} p(i_1) = \sum_{i_1=1}^{N} q_1(i_1)\{1 - q_2(i_1)\} = 1 - \sum_{i_1=1}^{N} q_1(i_1) q_2(i_1) \tag{1.5.3}$$

$$\begin{aligned} \sum_{s \in \mathfrak{I}_2} p(s) &= \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} p(i_1, i_2) = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} q_1(i_1) q_2(i_1) q_3(i_1, i_2)\{1 - q_2(i_1, i_2)\} \\ &= \sum_{i_1=1}^{N} q_1(i_1) q_2(i_1) - \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} q_1(i_1) q_2(i_1) q_3(i_1, i_2) q_2(i_1, i_2) \end{aligned} \tag{1.5.4}$$

$$\begin{aligned}\sum_{s\in\mathfrak{S}_{n_0-1}} p(s) &= \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0-1}=1}^{N} p(i_1,i_2,\ldots,i_{n_0-1})\\
&= \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0-1}=1}^{N} q_1(i_1)q_2(i_1)q_3(i_1,i_2)\cdots q_2(i_1,i_2,\ldots,i_{n_0-2})\\
&\quad q_3(i_1,i_2,\ldots,i_{n_0-1})\{1-q_2(i_1,i_2,\ldots,i_{n_0-1})\}\\
&= \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0-2}=1}^{N} q_1(i_1)q_2(i_1)q_3(i_1,i_2)\cdots q_2(i_1,i_2,\ldots,i_{n_0-2})\\
&\quad -\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0-1}=1}^{N} q_1(i_1)q_2(i_1)q_3(i_1,i_2)\cdots q_2(i_1,i_2,\ldots,i_{n_0-2})\\
&\quad q_3(i_1,i_2,\ldots,i_{n_0-1})q_2(i_1,i_2,\ldots,i_{n_0-1})
\end{aligned} \tag{1.5.5}$$

$$\begin{aligned}\sum_{s\in\mathfrak{S}_{n_0}} p(s) &= \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0}=1}^{N} p(i_1,i_2,\ldots,i_{n_0})\\
&= \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0}=1}^{N} q_1(i_1)q_2(i_1)q_3(i_1,i_2)\cdots q_2(i_1,i_2,\ldots,i_{n_0-1})\\
&\quad q_3(i_1,i_2,\ldots,i_{n_0})\{1-q_2(i_1,i_2,\ldots,i_{n_0})\}\\
&= \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_{n_0}=1}^{N} q_1(i_1)q_2(i_1)q_3(i_1,i_2)\cdots q_2(i_1,i_2,\ldots,i_{n_0-1})
\end{aligned} \tag{1.5.6}$$

(noting $q_2(i_1,i_2,..,i_{n_0})=0$ since n_0 is the maximum sample size)

Finally, from (1.5.2) to (1.5.6), we get

$$\sum_{s\in\mathfrak{S}} p(s) = \sum_{k=1}^{n_0}\sum_{s\in\mathfrak{S}_k} p(s) = 1$$

(ii) Here we have given a sampling design p where $\mathfrak{S}$ = all possible samples and $p(s)$ is the probability of selection of a sample s $(\in\mathfrak{S})$. Here we are to find q_1, q_2, and q_3 so that sampling according to the algorithm $\mathscr{A}(q_1,q_2,q_3)$ implements the design p.

Let $\mathfrak{S}_{(i)}=\{s|i_1=i\}$ = collection of samples, whose first element is i; $\mathfrak{S}_{(i,j)}=\{s|i_1=i,i_2=j\}$ = collection of samples, whose first element is i and the second element is j; $\mathfrak{S}_{(j_1,\ldots,j_k)}$'s are similarly defined.

Let $\beta(i_1, i_2,\ldots, i_n)$ = probability of selection of the sample $(i_1, i_2,\ldots, i_n)$ $= p(i_1, i_2,\ldots, i_n)$, where the unit i_1 is selected at the first draw, i_2 at the second draw, and i_n at the nth draw.

$\alpha(j_1) = \sum_{s\in \mathfrak{J}_{(j_1)}} p(s)$, $\alpha(j_1,\ldots,j_k) = \sum_{s\in \mathfrak{J}_{(j_1,\ldots,j_k)}} p(s)$ are defined similarly.

Here we check $\mathfrak{J} = \bigcup_{i=1}^{N} \mathfrak{J}_{(i)}$, $\mathfrak{J}_{(i)} = \bigcup_{j=1}^{N} \mathfrak{J}_{(i,j)} \cup (i)$, $\sum_{i=1}^{N} \alpha(i) = 1, \sum_{j=1}^{N} \alpha(i,j) + \beta(i) = \alpha(i)$, etc. Now following Hanurav (1966), we define

$$q_1(i_1) = \alpha(i_1)$$

$$q_2(i_1, i_2, \ldots, i_k) = \begin{cases} 1 - \dfrac{\beta(i_1, i_2, \ldots, i_k)}{\alpha(i_1, i_2, \ldots, i_k)} \\ \quad \text{if} \quad \alpha(i_1, i_2, \ldots, i_k) > 0 \\ \quad 0 \quad \text{otherwise} \end{cases}$$

$$q_3\{(i_1, i_2, \ldots, i_k), i_{k+1}\} = \begin{cases} \dfrac{\alpha(i_1, i_2, \ldots, i_{k+1})}{\alpha(i_1, i_2, \ldots, i_k) - \beta(i_1, i_2, \ldots, i_k)} \\ \quad \text{if} \quad q_2(i_1, i_2, \ldots, i_k) > 0 \\ \quad 0 \quad \text{otherwise} \end{cases}$$

So, the probability of drawing a sample $(i_1, i_2,\ldots, i_n)$ by the algorithm $\mathscr{A}$ is

$$\begin{aligned} p(i_1, i_2, \ldots, i_n) &= q_1(i_1) \times q_2(i_1) \times q_3(i_1, i_2) \times q_2(i_1, i_2) \times \cdots \times q_3(i_1, i_2, \ldots, i_n) \\ &\quad \times \{1 - q_2(i_1, i_2, \ldots, i_n)\} \\ &= \alpha(i_1) \times \left(1 - \frac{\beta(i_1)}{\alpha(i_1)}\right) \times \cdots \times \frac{\alpha(i_1, i_2, \ldots, i_n)}{\alpha(i_1, i_2, \ldots, i_{n-1}) - \beta(i_1, i_2, \ldots, i_{n-1})} \\ &\quad \times \frac{\beta(i_1, i_2, \ldots, i_n)}{\alpha(i_1, i_2, \ldots, i_n)} = \beta(i_1, i_2, \ldots, i_n) \end{aligned}$$

Example 1.5.6

Let us consider the sampling design where the population $U = \{1, 2, 3\}$; $\mathfrak{J}$ consists of the samples $s_1 = (1, 1)$, $s_2 = (3)$, and $s_3 = (2, 3)$ with respective probabilities $p(s_1) = 0.2$ and $p(s_2) = p(s_3) = 0.4$.

Here, $n_0 = 2 =$ maximum sample size. $\mathfrak{I}_1 = \{s_1\} \mathfrak{I}_2 = \{s_3\}$, $\mathfrak{I}_3 = \{s_2\}$; $\mathfrak{I}_{11} = \{s_1\}$ and $\mathfrak{I}_{1j} = \phi =$ (null set) for j = 2, 3; $\mathfrak{I}_{23} = \{s_3\}$, $\mathfrak{I}_{2j} = \phi$ for $j = 1, 2$ and $\mathfrak{I}_{3j} = \phi$ for $j = 1, 2, 3$; $\beta(3) = p(s_2) = 0.4$; $\beta(j) = 0$ for $j = 1, 2$; $\beta(1, 1) = p(s_1) = 0.2$, $\beta(2, 3) = p(s_3) = 0.4$ and other values of $\beta(i, j)$'s are equal to zero.

$\alpha(1) = p(\mathfrak{I}_1) = 0.2$, $\alpha(2) = p(\mathfrak{I}_2) = 0.4$, $\alpha(3) = p(\mathfrak{I}_3) = 0.4$; $\alpha(1, 1) = p(s_1) = 0.2$, $\alpha(1, j) = p(\phi) = 0$ for j = 2, 3; $\alpha(2, 3) = p\{s_3\} = 0.4$ and $\alpha(2, j) = p(\phi) = 0$ for $j = 1, 2$; $\alpha(3, j) = p(\phi) = 0$ for $j = 1, 2, 3$.

$q_1(1) = \alpha(1) = 0.2$, $q_1(2) = \alpha(2) = 0.4$, $q_1(3) = \alpha(3) = 0.4$;

$q_2(3) = 1 - \beta(3)/\alpha(3) = 1 - 1 = 0$, $q_2(j) = 1 - \beta(j)/\alpha(j) = 1$ for $j = 1, 2$, $q_2(1, 1) = 1 - \beta(1, 1)/\alpha(1, 1) = 0$, $q_2(1, j) = 0$ for $j = 2, 3$ since in this case $\alpha(1, j) = 0$; $q_2(2, 3) = 1 - \beta(2, 3)/\alpha(2, 3) = 0$ and $q_2(2, j) = 1 - \beta(2, j)/\alpha(2, j) = 0$ for $j \neq 3$ as $\alpha(2, j) = 0$; $q_2(3, j) = 1 - \beta(3, j)/\alpha(3, j) = 0$ as $\alpha(3, j) = 0$ for $j = 1, 2, 3$;

$q_3(1, 1) = \alpha(1, 1)/\{\alpha(1) - \beta(1)\} = 1$, $q_3(1, j) = 0$ for $j = 2, 3$ since $\alpha(1, j) = 0$; $q_3(2, 3) = \alpha(2, 3)/\{\alpha(2) - \beta(2)\} = 1$, $q_3(2, j) = 0$ for $j \neq 3$ as $\alpha(2, j) = 0$ for $j \neq 3$ and $q_3(3, j) = 0$ as $\alpha(3, j) = 0$ for j = 1, 2, 3.

Now using $q_1(\cdot)$, $q_2(\cdot)$, and $q_3(\cdot)$ we can check,

$$\begin{aligned} p(s_1) = p\{(1, 1)\} &= q_1(1) \times q_2(1) \times q_3(1, 1) \times \{1 - q_2(1, 1)\} \\ &= 0.2 \times 1 \times 1 \times (1 - 0) = 0.2 \end{aligned}$$

$$\begin{aligned} p(s_2) = p\{(3)\} &= q_1(3) \times \{1 - q_2(3)\} \\ &= 0.4 \times (1 - 0) = 0.4 \end{aligned}$$

$$\begin{aligned} p(s_3) = p\{(2, 3)\} &= q_1(2) \times q_2(2) \times q_3(2, 3) \times \{1 - q_2(2, 3)\} \\ &= 0.4 \times 1 \times 1 \times (1 - 0) = 0.4 \end{aligned}$$

1.6 ORDERED AND UNORDERED SAMPLE

A sample is said to be an ordered sample if it retains information about which draw selects which unit. So, from an ordered sample, we know the number of times a particular unit is selected in a sample and also in which draw it was selected. If we pick up the set of distinct units from an ordered sample and arrange them in ascending order of their labels, then the resulting sample is known as an unordered sample. Thus an unordered sample does not retain information about which draw a particular unit was selected and its multiplicity. Let $s = (i_1, \ldots, i_k, \ldots, i_{n_s})$ be an ordered sample of size n_s, where the unit i_k is selected at the kth draw. Let $\tilde{s} = \left(j_1, \ldots, j_k, \ldots, j_{\nu(s)}\right)$ be the set of distinct units in s of size $\nu(s)$ with $j_1 < \cdots < j_l < \cdots < j_{\nu(s)}$, then $\tilde{s}$ is an unordered sample obtained from s.

Example 1.6.1

Suppose from a population $U = (1, 2, 3, 4, 5)$, a sample of three units is selected as follows; On the first draw the unit 5, second draw unit 2, and at the third draw the unit 5 is selected. Then the sample $s = (5, 2, 5)$ is an ordered sample as we know, from the sample, that the unit 5 is selected twice, once in the first draw and again in the third draw whereas the unit 2 is selected in the second draw. Now, selecting the distinct units of the sample s and arranging in ascending order, we get an unordered sample $\widetilde{s} = (2, 5)$.

1.7 DATA

After selection of a sample s, we collect information on one or more characters of interest from the selected units in the sample s. Consider the simplest situation where a single character y is of interest, and y_i is the value of the character obtained from the ith unit. The information related to the units selected in a sample and their y-values obtained from the survey are known as data and will be denoted by d. Thus data corresponding to an ordered sample $s = (i_1,\ldots, i_k,\ldots, i_{n_s})$ will be denoted by

$$\begin{aligned} d(s) &= \left((i_1, y_{i_1}), \ldots, (i_k, y_{i_k}), \ldots, \left(i_{n_s}, y_{i_{n_s}}\right)\right) \\ &= ((j, y_j), j \in s). \end{aligned}$$

The data $d(s)$ based on the ordered sample are known as ordered data.

Similarly, the data obtained from the unordered sample $\widetilde{s} = \left(j_1, \ldots, j_l, \ldots, j_{\nu_s}\right)$ are known as unordered data and are denoted by

$$\begin{aligned} d(\widetilde{s}) &= \left(\left(j_1, y_{j_1}\right), \ldots, \left(j_k, y_{j_k}\right), \ldots, \left(j_{\nu_s}, y_{j_{\nu(s)}}\right)\right) \\ &= \left((j, y_j), j \in \widetilde{s}\right). \end{aligned}$$

If the label part of the data, i.e., information of the selected units in the sample, is deleted from the data, then the resulting data are called unlabeled data. Thus unlabeled data from an ordered and unordered sample may be denoted by $\left(y_{i_1}, \ldots, y_{i_k}, \ldots, y_{i_{n_s}}\right)$ and $\left(y_{j_1}, \ldots, y_{j_k}, \ldots, y_{j_{\nu(s)}}\right)$, respectively.

1.7.1 Sample Space

The sample space corresponding to a sampling design p is defined as the collection of all possible data that could be obtained from the design. The sample space corresponding to an ordered and unordered design will be denoted by $\mathscr{X}_o = (d(s), s \in \mathfrak{J})$ and $\mathscr{X} = (d(\widetilde{s}), \widetilde{s} \in \mathfrak{J})$, respectively.

1.8 SAMPLING FROM HYPOTHETICAL POPULATIONS

Let X be a random variable with a distribution function $F(x) = P(X \leq x)$. To draw a sample from this population, we use the property that $F(x)$ follows uniform distribution over (0, 1). Let R be a random sample from a uniform distribution. Then $x = F^{-1}(R)$ is a random sample from a population, whose distribution function is $F(x)$.

1.8.1 Sampling From a Uniform Population

Here we select a five-digit random number (selection of more digits gives better accuracy) from a random number table and then place a decimal point preceding the digits. The resulting number is a sample from the uniform distribution over (0, 1). For example, if the selected five-digit random number is 56342 the selected sample from a uniform population (0, 1) is $R = 0.56342$.

1.8.2 Sampling From a Normal Population

Suppose we want to select a random sample from a normal population with mean $\mu = 50$ and variance $\sigma^2 = 25$. We first select a five-digit random number 89743 and put a decimal place preceding it. The resulting number $R = 0.89743$ is a random sample from a uniform distribution (0, 1). A random sample x from a normal population $N(\mu, \sigma)$ with mean $\mu(=50)$ and variance $\sigma^2(=25)$ is obtained from the equation,

$$P(X \leq x) = 0.89743 \text{ i.e., } P\left(\frac{X-\mu}{\sigma} = z \leq \frac{x-50}{5}\right) = 0.89743.$$

Noting that $\dfrac{X-\mu}{\sigma} = z$ is a standard normal variable, we find from the normal deviate table, $\dfrac{x-50}{5} \cong 1.27$. Hence $x = 56.35$ is a random sample from $N(50, 5)$.

1.8.3 Sampling From a Binomial Population

Suppose we want to select a sample from a binomial population with $n = 5$ and $p = 0.342$. Let X be a Bernoulli variable with a success probability $p = 0.342$. Then, $P\{X = 1\} = p$ and $P\{X = 0\} = 1 - p$. We first select five independent random samples $R_1 = 0.302$, $R_2 = 0.987$, $R_3 = 0.098$, $R_4 = 0.352$, and $R_5 = 0.004$ from a uniform distribution over (0, 1) using Section 1.8.1. From the random samples R_i, select a random sample from Bernoulli population X_i, which is equal to 1 (success) if $R_i \leq p(=0.342)$

and $X_i = 0$ (failure) if $R_i > p$. Then $Y = X_1 + X_2 + X_3 + X_4 + X_5 = 1 + 0 + 1 + 0 + 1 = 3$ is a random sample from the Binomial population with $n = 5$ and $p = 0.342$.

1.9 EXERCISES

1.9.1 Define the following terms giving suitable examples: (i) population, (ii) sampling frame, (iii) sample, (iv) sampling scheme, (v) sampling design, and (vi) effective sample size.

1.9.2 **(a)** Define inclusion probabilities of the first two orders. Compute inclusion probabilities of the first two orders of the following sampling designs: (i) SRSWR, (ii) SRSWOR, and (iii) PPSWR.

(b) Find (i) expectation and (ii) variance of the number of distinct units in a sample of size 5, selected from a population of size 10, by the SRSWR method.

1.9.3 Let the expected effective sample size of a sampling design be $\nu = E(\nu(s)) = [\nu] + \theta$, where $[\nu]$ is the integer part of ν. Then show that

(i) $\theta(1-\theta) \le Var(\nu(s)) \le (N-\nu)(\nu-1)$ and (ii) $\nu(\nu-1) + \theta(1-\theta) \le \sum_{i\neq}\sum_{j} \pi_{ij} \le N(\nu-1)$ (Hanurav, 1966).

1.9.4 **(a)** Let π_i^* and π_{ij}^* be the exclusion (noninclusion) probabilities for the ith, and ith and jth $(i \neq j)$ units, then show that $\pi_i^*\pi_j^* - \pi_{ij}^* = \pi_i\pi_j - \pi_{ij}$ (Lanke, 1975a,b).

(b) Show that the first two order exclusion probabilities of units in SRSWOR sampling of size n selected from a population of size N are $(N-n)/N$ and $(N-n)(N-n-1)/\{N(N-1)\}$, respectively.

1.9.5 Let π_{ijk} be the inclusion probability of the unit i, j and $k (i \neq j \neq k)$ for a fixed effective size ν design. Show that (i) $\sum_{k(\neq i,j)} \pi_{ijk} = (\nu-2)\pi_{ij}$ and (ii) $\sum_{i\neq}\sum_{j\neq}\sum_{k} \pi_{ijk} = \nu(\nu-1)(\nu-2)$

1.9.6 Consider the sampling design where $U = (1, 2, 3, 4)$ and

Sample (s)	(1, 2, 1)	(1, 4, 3)	(3)	(4)	(3, 2, 1)	(2, 2)
Probability $p(s)$	0.2	0.1	0.4	0.1	0.1	0.1

(a) Calculate (i) inclusion probabilities of first two orders, (ii) $E(\nu(s))$, and (iii) $Var(\nu(s))$

(b) Select a sample using the cumulative total method

1.9.7 Use Hanurav's algorithm to select a sample using the following sampling designs:

(a) $U = (1, 2, 3, 4, 5, 6)$

Sample (s)	(1, 4)	(2, 5)	(3, 6)
Probability $p(s)$	1/3	1/3	1/3

(b) $U = (1, 2, 3, 4)$

Sample (s)	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
Probability $p(s)$	1/6	1/6	1/6	1/6	1/6	1/6

(c) $U = (1, 2, 3, 4, 5)$

Sample (s)	(1, 2, 2)	(1, 3, 4)	(4)	(1, 5, 3)	(1, 1, 4)	(3, 4)
Probability $p(s)$	0.20	0.10	0.25	0.25	0.10	0.10

1.9.8 Using a random number table, select a sample of size 5 from the following populations:

(i) Uniform distribution over (0, 1)
(ii) Uniform distribution over (10, 100)
(iii) Bernoulli population with parameter $p = 0.1234$.
(iv) Binomial distribution with parameters $n = 8$ and $p = 0.673$.
(v) Hypergeometric distribution with $N_1 = 10$, $N_2 = 15$, and $n = 8$.
(vi) Poisson distribution with parameter $\lambda = 4$.
(vii) Normal population with mean $\mu = 50$ and standard deviation $\sigma = 5$.
(viii) Chi-square distribution with degrees of freedom 10.
(ix) Bivariate normal distribution with parameters $\mu_1 = 50$, $\mu_2 = 60, \sigma_1^2 = 16, \sigma_2^2 = 25$, and $\rho = 0.8$.
(x) Cauchy population $f(x|\theta) = 1/[\pi\{1 + (x - \theta)^2\}]$; $\theta = 5$, $-\infty < x < \infty$

1.9.9 The following table gives a list of households in 10 localities.

Localities	1	2	3	4	5	6	7	8	9	10
Number of households	5	15	20	10	15	25	15	10	10	15

Select 15 households at random by (i) SRSWR and (ii) SRSWOR methods.

1.9.10 Select five points at random in a (i) circle of radius 5 cm, and (ii) square of sides 5 cm.

1.9.11 The following table gives the number of students in different sections and grades. Select a sample of size 5 by (i) SRSWR and (ii) SRSWOR methods.

Grades	Section 1	Section 2	Section 3	Section 4
A	10		5	5
B	12	30	10	15
C	40	25	20	
D	5	25	30	20
E		20	5	10