University of Khartoum
Faculty of Mathematical Sciences and Informatics

METHODS FOR SELECTING UNEQUAL PROBABILITY SAMPLING

Prepared by:

1. **Manar Jamaleldin Elrayah Elfaki 016-236**
2. **Tibyan Abdalaa Rajab Hassan 016-110**

Supervised by:

Dr. Ibrahim Elabid

بسم الله الرحمن الرحيم

"وَمَا تَوْفِيقِي إِلَّا بِاللَّهِ عَلَيْهِ تَوَكَّلْتُـوَإِلَيْهِ أُنِيبُ" (سورة هود: 88)

Dedication:

To our friend **Omer Hassan Hamad** may Allah grant him the highest ranks in Jannah.
You will forever remain in our hearts, and we will never forget you...

Abstract:

The ultimate goal in sampling is to obtain a sample from a population in order to estimate an unknown population parameter, usually a total or mean of some variable of interest. Unequal probability sampling is when the units in the population do not all have an equal chance of being included in a sample. The probabilities of inclusion are usually made proportional to some auxiliary variable available for all population units. Where unequal probability sampling is practicable, it will give much better estimates than equal probability sampling. The present investigation is concerned with unequal probability sampling of a finite population of units.

In this research, unequal probability sampling methods are reviewed with a particular focus on Probability Proportional to Size (PPS) sampling methods. It contrasts and compares key methods, including Lahiri's Method, the Cumulative Totals Method, and the Splitting Method, their theoretical underpinnings, practical consequences, and efficiency. To demonstrate the methods in action, a simulated dataset was created in R and analyzed.

The basis of comparison was effectiveness and efficiency using sample variance as one of the measures of performance. The results gave insight into the strengths and weaknesses of each method, providing valuable information to help researchers choose the most suitable method for scenarios with unequal probabilities.

By closing gaps in the knowledge of unequal probability sampling, this research provides practical guidance on the application of these methods in applied research environments.

Table of Contents:

List of Figures:

List of Table:

Chapter One
Introduction

1.1 Introduction:

Sampling techniques in statistical research and data analysis make it possible to collect representative data from larger populations. Understanding different approaches to sampling becomes much more relevant as a means whereby researchers are pursuing increased accuracy and efficiency methods of collecting data. This paper, therefore, reviews methods for selecting unequal probability sampling.

In recent years, unequal probability sampling has caught considerable attention due to the fact that it has the potential for increasing efficiency in estimation and reducing bias under certain situations. When the selection probabilities for units that are most informative or representative of the population characteristics of interest are high, in many cases, more precise estimates can be obtained with smaller sample sizes than would be required by the methods of equal probability sampling.

This research attempts to present a review of different unequal probability sampling techniques. We will present the theoretical backgrounds, practical uses, advantages, and limitations of these methods. Based on both classical and modern perspectives, this review hopes to present an understanding to researchers and practitioners of the techniques available and to provide guidance in the selection of an appropriate method for a particular research problem.

1.2 Research Problem:

In the field of education, this method has not been extensively detailed or thoroughly explored. This research provides a straightforward and clear review of unequal probability sampling.

Existing methods of unequal probability sampling, such as Lahiri's Method, Cumulative Totals Method, and Splitting Method, are popular but not comprehensively evaluated in the context of their practical efficiency, variability, and effectiveness in practice. Researchers typically find it challenging to choose the most appropriate method due to a shortage of comparative research and practical guidelines. This research answers the need to evaluate these methods systematically, using theoretical and practical approaches, to establish their merits, limitations, and applicability.

1.3 Aims & Objective of The Research:

1.3.1 Aims of The Research:

Unequal probability sampling is crucial because it gives rise to greater efficiency and precision in statistical estimation because it assigns higher selection probabilities to more informative units, boosting representativeness and reducing bias. It is of most use in complicated surveys, resource allocation, and rare event or subpopulation investigations since it ensures sufficient representation of critical subgroups. By exploiting auxiliary information, unequal probability sampling achieves more efficient allocation of resources, better data quality, and more accurate results, which are essential for decision-making and policy-making. Uses range from environmental studies, where larger polluting plants are sampled more frequently, to market research on high-value consumers and health surveys on high-risk populations. A comparison of methods like Lahiri's Method, Cumulative Totals Method, and Splitting Method helps researchers identify the most appropriate methods for their specific needs, hence leading to better research outcomes and decision-making.

The primary aim of this research is to provide a comprehensive understanding of unequal probability sampling and its methods, focusing on its principles, applications, and significance in research methodologies and to evaluate and compare different sampling methods for unequal probability selection, including Lahiri's Method, Cumulative Totals Method, and Splitting Method, in order to identify their efficiency, effectiveness, and applicability in generating representative samples.

1.3.2 Objective of The Research:

The objective of this research is to provide a comprehensive review of the
Unequal probability sampling,

1. To learn in detail the theoretical underpinnings, models, key concepts, assumptions, and examples of usage of unequal probability methods of sampling, i.e., Probability Proportional to Size (PPS) methods.

2. To generate a simulated dataset in R and carry out Lahiri's Method, Cumulative Totals Method, and Splitting Method, and compare their efficiency on the basis of statistical parameters such as variance and mean.

3.In order to compare the effectiveness of these three techniques in reducing sample data variability, and to identify their strengths and weaknesses under different research conditions.

4.To highlight the central role of the methods of sampling in achieving accurate and reliable statistical analysis, and to provide practical guidelines for the selection of the most appropriate method of sampling in case of unequal probabilities.

Chapter Two
Literature Review

Introduction:

Unequal probability sampling is an integral technique that could greatly enhance the precision and relevance of data obtained in such complicated research scenarios. Its application has spanned many different areas: for the study of rare events, measuring subgroup-specific effects, or improving the representation of critical population groups. Outcomes of this kind of sampling can provide appropriate guidance on making decisions, formulating resource allocations, and defining intervention designs. In this chapter, we will outline the foundational concepts and principles of unequal probability sampling to provide a comprehensive understanding of its practical applications and importance.

2.1 Basic concepts:

2.1.1 Survey

Today the word “survey” is used most often to describe a method of gathering information from a sample of individuals. (Scheuren, n.d., 9)

A survey is a research method used for collecting data from a predefined group of respondents to gain information and insights about various topics of interest. Surveys are commonly used in social sciences, marketing research, public opinion polling, and various other fields to collect quantitative and qualitative data. They can be administered through different modes, including face-to-face interviews, telephone interviews, mail questionnaires, and online forms. (Fink, 2003, 1)

Data collection in statistical research is a very crucial step, and basically, there are two approaches towards the collection of data: **complete survey or census** and **sample surveys**. Each has its merits and demerits that make it appropriate for various research situations.

Complete Survey (Census)

A complete survey, or census is a comprehensive and official data collection operation that aims to count and gather information about every individual in the population. (B., King, T. & Valente, 2013, 407-425)

The advantages of a census include comprehensive data as the data from a complete survey encompasses the entire population, thereby ruling out any possibility of bias or sampling error. This makes sure that the results obtained are highly accurate and reflect the true nature of the population characteristics under study(Pedigo & Buntin, 1994,

714)). It also allows for detailed analysis enabling the researchers to carry out detailed analyses of subgroups or even cross-tabulations with complete abandon, without concerns over sample size.(Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau, 2011)

Furthermore, census data is vital for policy and decision-making as it is frequently used by governments and organizations in making informed decisions over the apportionment of resources, planning, and policy formulation. (United Nations Statistics Division, 2008)

However, there are disadvantages associated with a census, including high cost, note that surveying can be very expensive, especially if the population is large enough, since it requires vast resources for data collection and processing. (United Nations Statistics Division, 2008,p.45).

Additionally, it is time consuming as the process of collecting data from each member of the population does take time, thus delaying the availability of results. (Blair et al., 2013, p.88)

Logistical challenges also arise, organizing a complete survey as this needs a great deal of forethought and management in order to have good coverage and the data collected must be very accurate. (Bethlehem, 2009, p.101)

Sample Survey

The data is collected from a portion of the population. Again, the information obtained from the sample is used to make inferences about the whole population. Proper sampling techniques are to be followed in order to ensure that the sample drawn is representative of the population(Cochran, 1977).

The advantages of sample surveys include cost effective , as they are generally less expensive compared to complete surveys since relatively fewer resources are invested in data collection and analysis of the sample(Cochran, 1977,p.1-2).

They are also **time efficient** because sampling reduces the number of respondents. Thus, data collection and results can be obtained sooner, which is especially valuable in sensitive studies(Cochran, 1977, p.1-2).

Additionally, sample surveys offer practicality, as there are greater practicality in the sampling of huge or dispersed populations since the survey of such a population may be literally impossible to conduct(Cochran, 1977, p.1-2).

However, sample surveys have disadvantages, including sampling error as surveying only a subset population that is surveyed, there is the possibility of sampling error. In case the sample is not representative, this may result in inaccurate or biased results(Cochran, 1977, p.14-15).There is also limited depth of information,Most of the surveys may give breadth at the expense of depth. Sample surveys cannot capture the details a complete survey does. Therefore, sample surveys fail at times to show hidden things, Moreover, sample surveys exhibit inflexibility as making changes or adding questions after the survey has been fielded becomes very challenging and difficult to make changes(Fowler, 2014).

2.1.2 Sampling

Populations and Samples:

Research is done to draw informed conclusions about the parameters of populations. For instance, what is the likelihood of our target audience buying our product if certain features are included? Or how satisfied are our customers with the products and services we offer? Making sense of the research process requires an understanding of differences between populations and samples, all of the entities (orders, stores, customers, sales, units, etc.) about which we want to make inferences or decisions are referred to as a population. The study is referred to be a census when researchers measure every aspect of the population. A census is frequently impractical for a variety of reasons, chief among them being time and expense. A smaller portion of the population is typically chosen for research. We refer to this smaller subset as a sample. Only the individual sample members are measured when a sample is used.

Why not conduct a census; a study of every individual in the population? There are three factors that limit the ability to conduct a census in research situations: (Boyd, n.d., 2)

1. Identifying all persons in a population as a single group may be too impractical or impossible. (Boyd, n.d., 2)
2. Even if the whole population could be accounted for, it might only be prohibitively expensive in collecting data from every individual. (Boyd, n.d., 2)
3. Certain research methods render the object being researched non-functional and therefore impossible to analyze in full. (Boyd, n.d., 2)

Selection of units with replacement (WR):

The probability of selection of a unit will not change, and the probability of selecting a specified unit is the same at any stage. There is no redistribution of the probabilities after a draw. (Professor Shalabh, n.d.)

Selection of units without replacement (WOR):

The probability of selection of a unit will change at any stage, and the probabilities are redistributed after
each draw. (Professor Shalabh, n.d.)

2.1.3 Non-Probability Sample

Sampling methods are classified into two categories: **probability sampling** and **non-probability sampling**. We will begin by defining non-probability sampling.
Oftentimes, it is not practical to sample randomly or use a formal sampling frame. For example, there might not be a list compiled of the people in some specific population of interest. There is no list of single fathers with children between the ages of 5 and 8 who do their shopping at a mall on weekdays. When the study of such populations is imperative, one must still be able to identify a sample from which to gather needed information. (Boyd, n.d.,p.9-10)
The method of sampling in these situations is called non-Probability sampling because not everyone of the pertinent population has the chance to be included in a sample using these procedures.
Non-probability sampling is the sample method in which participants are selected based on nonrandomized criteria, including availability, geographical proximity, and expertise or any other factor considered relevant for the representation of a given research problem. (Nikolopoulou, 2022)
This approach is further followed by the following types:

1.Convenience sampling:

Convenience sampling: a non-probability sample in which units are selected for inclusion in the sample based on their ease of access to the researcher. Units may be selected because they are geographically proximate, available at a particular point in time, or willing to participate. Sometimes called accidental sampling, this is a non-random method of sampling. (Nikolopoulou, 2022)

2.Purposive Sampling

purposive sampling A collective term describing a suite of sampling techniques in which participants are deliberately selected due to their possession of particular characteristics or qualities. Sometimes called judgmental sampling, selection of units-that is, cases or individuals or organisations-to be investigated, depends on the discretion of the researcher this method is frequently used in qualitative and mixed methods research designs, particularly when addressing unique cases or specific issues that require targeted exploration. (Nikolopoulou, 2022)

3.Snowball Sampling

Snowball sampling is a technique used when there is no list to refer to, or for those that are very hard to access, such as for most hidden populations. This usually applies in research targeting social problems such as addiction, homelessness, or sex work this method begins with identifying a first participant who agrees to participate in the research and then asks him or her to refer other individuals who could also meet the criteria of the study ,alternatively, snowball sampling can also be used in studies focusing on people with specific experiences or those who use particular products. In this case, the researcher uses personal networks or connections to reach a target population. (Nikolopoulou, 2022)

4.Quota Sampling

Quota sampling involves the selection of a fixed number or percentage of units, a quota, divided into subgroups or strata, according to certain characteristics, such as individuals, cases, or organizations, and selected in a nonrandom manner.the strata are such that the groups must always be mutually exclusive. Data from earlier studies or related data can be used at this stage to make possible estimates of the quota. This estimate can only be used to guide persons on how many units there are to select in each Subgroup. Data collection involves the recruitment for participants until the quota as determined in each subgroup have been filled.

2.1.4 Probability Sample

The probability samples are taken as the gold standard in the variety of sampling methodologies and provide reasonable assurance that the study findings generalize to the target population. In probability sampling, every individual in the given population has an equal chance to be selected for this study. (Acharya et al., n.d., 1)

This approach is further followed by the following types:

1. Simple random sampling
2. Systematic random sampling
3. Stratified random sampling
4. Cluster sampling

Equal Probability Sampling:

Equal probability random sampling is one in which every unit in the population has equal and known probability of being selected. (Lohr, 2010) In equal probability sampling methods, every unit in the population has an equal chance of being selected for the sample. This ensures that every member of the population has an equal chance of being included in the sample and is usually defined as 1/N where N is the population size. (Särndal et al., 2003, p.66)

In the present section, each of these methods will be described briefly and illustrated with examples.

1. Simple Random Sampling(SRS):

Simple Random Sampling (SRS) is the simplest and most common method of selecting a sample, in which the sample is selected unit by unit , with equal probability of selection for each unit at each draw. In other words, simple random sampling is a method of selecting a sample s of II units from a population n of size N by giving equal probability of selection to all units. It is a sampling scheme in which all possible combinations of II units may be formed from the population of N units with the same chance of selection. (Singh, 2003, p.71)

Simple random sampling can be used in two scenarios:(a) When a unit has been selected, observed, and then put back in the population before drawing the next lot, and the process is repeated n times, the result will be a simple random sample of n units. This method is known as simple random sampling with replacement, abbreviated as SRSWR. (Singh, 2003, p.71),

(b) If a unit is selected, observed, and not returned to the population before the next draw, and this process is continued until n distinct units are selected, with repetitions ignored, it is called simple random sampling without replacement, abbreviated as SRSWOR. Let us now examine the properties of estimators for the population mean, variance, and proportion in both cases. (Singh, 2003, p.71)

Advantages of simple random sampling include Impartial random selection are crucial and representativeness of the population,Randomization helps to offset the confounding effects of known and unknown factors and each sample has an equal probability of selection, However, there are disadvantages to this method. The procedure is cumbersome and rarely used. Also a complete list of the population is needed, which is not always readily available. Additionally, it is difficult to use when the population is widely dispersed and heterogeneous and there is a risk of sampling error (Golzar, 2022).

Simple Random Sampling method:

1. **Chit Method or Lottery Method:**

Assign each element of the population a number or name. Those numbers or names are then put into a receptacle (e.g. a bowl or box). It is a lottery draw wherein numbers are drawn at random to create the sample (Singh, 2003).Then there are two possibilities :

<u>With Replacement Sampling (SRSWR):</u>

After selecting a chit, return it to the box before drawing the next one,any unit can be selected many different times.

Suppose a population consists of N = 3 blocks(A, B, C) and sample size 2, **The possible ordered samples** are : AA, AB, AC , BA, BB, BC, CA, CB, CC.

Thus a total 9 samples of size 2 can be drawn from the population of size 3, which in fact is given by $3^2 = 9$. (Singh, 2003)

Without Replacement Sampling (SRSWOR):

After selecting a chit, we don't return the chit to the box ,any unit can only be selected once. (Singh, 2003)

Suppose a population consists of N = 3 blocks A, B and C, and sample size 2, The possible samples are : AB,AC, BC . Thus a total of 3 samples of size 2 can be drawn from the population of size 3, which in fact is given by ${}^{3}C_{2} = 3$. (Singh, 2003)

2. **Random Number Tables:**

   A random number table is a set of numbers used for drawing random samples. The numbers are usually compiled by a process involving a chance element, and in their simplest form, consist of a series of digits 0 to 9 occurring at random with equal probability. (Singh, 2003)

   Suppose we are given a population of N = 225 units and we want to select a sample of say n = 36 units from it. To pick up a random sample of 36 units out of a population of 225 units , use any three columns from the random number table. (Singh, 2003)

3. **Random Numbers Generated by Computers:**

   Most modern methods of sampling involve the use of computational tools such as Excel, Python, or R that have the option of generating random numbers. They provide fast, exact results that scale well to large population sizes. (*Random — Generate Pseudo-Random Numbers — Python 3.13.1 Documentation*, n.d.)

   For example: The random.sample() function in Python can be used to return random samples from a population.

4. **Systematic Random Sampling:**

In systematic random sampling, a random starting unit is selected from the whole population. After this, every nth unit is selected down the list.

For example: For a population of 100 individuals, if the 3rd unit is randomly selected as the starting point and the interval is 5, the selected units would be: 3, 8, 13, 18, and so on. (Cochran, 1977)

5. **Using Apps or Software for Random Sampling:**

Applications and software tools used in the process of random sampling include Google Sheets, MS Excel, and statistical programs like SPSS. These execute the work of sampling in a very easy and accurate manner.

For example: In MS Excel, random values can be assigned to the elements of the population using the RAND() function. Then, the list is sorted and the sample as required is selected. (*Microsoft Corporation. (N.d.). RAND Function - Excel Help.*, n.d.)

2. **Stratified Random Sampling:**

The procedure of partitioning the population into groups, called **strata**, and then drawing a sample independently from each stratum, is known as **stratified sampling**.

If the sample drawn from each stratum is a random one, the procedure is then termed as **stratified random sampling**. (Singh & Singh Mangat, 2010, p.102)

For example, when investigating customer satisfaction for a dog-walking business, you may want to differentiate between people of different ages. In this case, you would divide the population into groups based on their age and then draw a random sample from each age group to collect data.

Some broad principles should be kept in mind while using stratified sampling: the strata should be non-overlapping and should together comprise the whole population. The units forming any stratum should be similar with respect to the study variable to reduce variability within each stratum. When it is difficult to stratify the population with respect to the study variable or a highly correlated auxiliary variable, administrative convenience may be considered as the basis for stratification (Singh & Singh Mangat, 2010, p.103).

Stratified random sampling is particularly suitable for studies with heterogeneous populations containing unique subgroups where accuracy, comparisons among subgroups, or balanced representation are important. It is especially powerful when subgroup analysis is required.

Advantages of stratified sampling include minimizing human bias in the selection process by ensuring that probabilistic methods are used in arriving at the sample. The eventual outcome will be a sample that more truly reflects the population being studied, so long as there is minimal data that is missing. It ensures that the sampling will capture all major representatives of the population by dividing the population into distinct subgroups, or strata, and sampling from each.

Additionally, Since the units included in the sample are probabilistically selected, stratified sampling allows the researcher to make statistical generalizations—that is, from the sample to a wider population—validly. Therefore, it enhances external validity in the study's results. (Sharma, 2017)

However, stratified sampling also has disadvantages. It assumes that the population can be divided into distinct, non-overlapping subgroups comprehensively. When this cannot be effectively or accurately done, the methodology becomes impractical.A common misapplication is scaling subgroup sample sizes based on data availability rather than subgroup size or variability. This can easily result in biased or unrepresentative samples.The method often assumes that each subgroup is equally important. If there are significant variations between subgroups, this assumption may not hold, leading to inaccurate conclusions.When subgroup variances differ widely, stratification may require scaling by variance. However, it is challenging to simultaneously adjust sample sizes proportionally to both subgroup size and variance, making it difficult to balance representation and efficiency .Efficient resource allocation among groups with different variances and means is often complex. Simple proportional allocation may be insufficient when subgroup characteristics vary significantly. (Sharma, 2017)

Here steps for performing a stratified random sampling:

1. **Specify population and subgroups**

   Determine and access the complete target population,can the population be divided into distinct, non-overlapping subgroups clearly defined by characteristics that are relevant to the study Determine the population of interest and Ensure that population can be divided into subgroups (strata). (Cochran, 1977)

2. **Divide the population into subgroups**

   Divide the population into non-overlapping subgroups (strata) based on relevant characteristics like age, income, or region. (Kish, 1965)

3. **Choose the sample size for subgroups**

   This step is to make sure the sample size for each subgroup is proportionate to the entire population. While subgroups that are less represented in the population are less represented in the sample, subgroups more represented in the population are also more represented in the sample. Decide on a total sample size that's large enough to draw statistical conclusions for each stratum.

   Decide total sample size and allocate it:

- Proportional to the stratum size.

  This is one of the two stratified sampling types. The number of elements assigned to the different strata is proportional to the representation of the strata in the target population. In proportionate terms. That is to say, in the target population, the scope of the sample taken from each stratum is proportional to the relative size of that stratum. (Etikan, 2021)

- Disproportionally for high-variance or critical subgroups

  This is another stratified sampling technique. The number of elements sampled in disproportional stratified random sampling from each stratum is not equal to their population representation. In the population, the element has no fair chance of being included in the sample. There is no smearing of the fraction with each stratum. On the other hand, there are distinct sampling fractions in the strata. To estimate population variables, the population arrangement must be applied as weights to compensate for the disproportionality in the sample. On the other hand, for some research activities, disproportionate stratified sampling may be more appropriate than proportionate stratified sampling. (Etikan, 2021)

4. **Take random samples of the subgroups**
   Use random sampling methods within each stratum to select participants,Merge the samples from all strata and ensure representativeness using statistical checks (Lohr, 2021, #) This presents a stratified random sampling of the original population.

Methods of sample allocation to different strata :

1. **Equal Allocation:**

   In case of equal allocation, number of sampling units selected from each stratum is equal.

   Thus for h = 1, 2, ... ,L,

Sample size for h-th stratum in case of equal allocation :

$$n_h = \frac{n}{L}$$

Total sample size for fixed total cost :

$$n = \frac{L\,(C - c_0)}{\sum\limits_{h=1}^{L} c_h}$$

It may be pointed out here, that this method of sample allocation is used when strata sizes do not differ much from each other, and the information about the variation within the strata is lacking. (Singh & Singh Mangat, 2010, 108-109)

2. **Proportional Allocation**

The sample size from each stratum is proportional to its size in the total population, This means $n_h \propto N_h$

Sample size for h-th stratum in case of proportional allocation :

$$n_h = \frac{n}{N} N_h$$

Total sample size for fixed total cost :

$$n = \frac{C - c_0}{\sum\limits_{h=1}^{L} W_h c_h}$$

The allocation is likely' to be nearly optimum for a fixed sample size, when the strata variances are almost same (Singh & Singh Mangat, 2010, 108-109)

Where:

n_h : Sample size for stratum h

N_h : Population size of stratum h

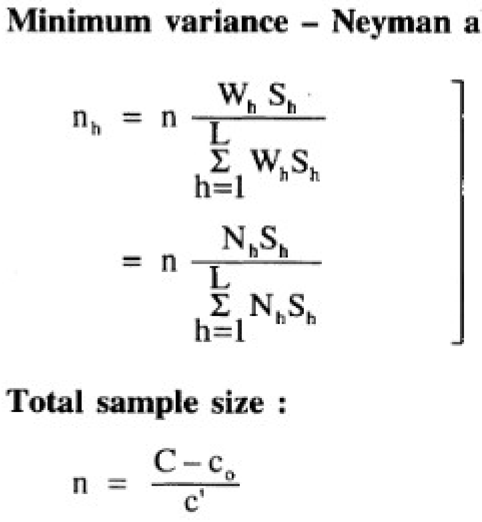
N : Total population size

n : Total sample size

3. **Optimum/Neyman Allocation**

Optimum or Neyman allocation is a statistical method used in stratified sampling to allocate sample sizes across strata in such a way as to **minimize the variance of an estimator** for a fixed total sample size or cost. Alternatively, it can be used to **minimize the cost** of the survey while maintaining a specified maximum variance for the estimator. (Singh & Singh Mangat, 2010, 114)

the strata sample sizes for Neyman allocation are given as:



Minimum variance – Neyman allocation :

$$\left.\begin{aligned} n_h &= n\,\frac{W_h S_h}{\sum\limits_{h=1}^{L} W_h S_h} \\ &= n\,\frac{N_h S_h}{\sum\limits_{h=1}^{L} N_h S_h} \end{aligned}\right]$$

Total sample size :

$$n = \frac{C - c_o}{c'}$$

Where:

N_h : The population size of each stratum.

S_h : The variability (standard deviation) within each stratum.

C_h: The cost of surveying one unit in each stratum

W_h :The weight of the population in stratum

3. Systematic Random Sampling

The method in which only the first unit is selected at random, the
rest being automatically selected according to a predetermined pattern. (Singh & Singh Mangat, 2010, 145)
Suppose we want to select a systematic sample of size n from a population
consisting of N units. The method of LS sampling is employed when N is a multiple of n, that is, N=nk where k is an integer.

As a hypothetical example of systematic sampling, assume that, in a population of 10,000 people, a statistician selects every 100th person for sampling. The sampling intervals can also be systematic, such as choosing a new sample to draw from every 12 hours.

Another example Suppose a researcher wants to select a sample of 50 **individuals** from a population of **200** people. To determine the sampling interval (k):

k=200/50 =4

if we select the 3ed individual ,and then select every **4th individual** , using this way we make sure we spread across the entire population. (Cochran, 1977, 100-105)

You can use the following steps to create a systematic sample :

Define your population—this is the group from which you are sampling.

Settle on a sample size—determine how many subjects you want or need to sample from the population to get a reflective idea of it.

Assign every member of the population a number—if the group consists of, for example, 10,000 people, start by lining them up and assigning numbers to each.

Decide the sampling interval—this can be calculated by dividing the population size by the desired sample size.

Choose a starting point—this can be achieved by selecting a random number.

Identify members of your sample—if your starting point is 15 and your sampling interval is 100, the first member of the sample would be 115, and so on. (Systematic Sampling, n.d.)

Systematic sampling can be used whenever you want the benefits of randomly sampling the population you're studying. It is particularly useful in situations where you don't have details of the entire population before starting your study, as systematic sampling is rule-based and allows you to apply the chosen interval directly to the data.

Systematic sampling is most appropriate in instances where there is some type of order or regularity within the population being studied. For example, when surveying customers entering a particularly filled store, a systematic sampling method enables you to pick every ***nth*** customer, ensuring a representative set of customers across different times of day or the week. This approach mitigates bias that might arise from only examining customers who arrive during specific time frames (Cochran, 1977).

Advantages of Systematic Random Sampling include that it's easier to conduct than a simple random sample. Also it spreads the sample more evenly over the population**.**

However, systematic sampling also has disadvantages. It assumes that the systematic sampling can fail when there is a hidden periodic trait in the population. If the sampling interval aligns with the periodic cycle of this trait, the sample becomes biased,Also the alignment of the sampling interval with the population's periodic traits disrupts the randomness of the sampling process,and the sample may overrepresented or underrepresented certain characteristics, making it unrepresentative of the entire population.(Sharma, 2017, 750)

Types of Systematic Sampling:

Here are the types of systematic sampling:

1. **Systematic Random Sampling** This is the classic form of systematic sampling where the subject is selected at a predetermined interval. For example, if a researcher wants to select a sample of 100 students from a population of 1000, they could use systematic random sampling by selecting every 10th student from a list sorted in random order. This approach ensures that each member of the population has an equal chance of being selected, while still maintaining a systematic sampling pattern. (Singh & Singh Mangat, 2010)

2. **Linear Systematic Sampling**

   Suppose we want to select a systematic sample of size n from a population consisting of N units. The method of LS sampling is employed when N is a multiple of n, that is, N=nk where k is an integer. So, instead of randomly choosing the sampling interval, this is when you create a skip pattern moving in a linear path. No longer picking every nth member from the population, the selection itself follows a fixed order e.g. select every 5th, then every 7th, then every 9th, etc. Linear systematic sampling is useful when the population has some definite order, to wade through the population systematically. (Singh & Singh Mangat, 2010, 145-146)

3. **Circular Systematic Sampling**

   This is where a sample loops to the beginning when it finishes playing. This means that the sampling interval, after selecting the last member of the population, goes back to the beginning, and carries on with the selection. In cases such as those involving populations that may have cyclical patterns or populations that do not have a clear beginning or end, circular systematic sampling is commonly utilized. (Singh & Singh Mangat, 2010, 148)

4. Cluster Random Sampling

The cluster sampling consists of forming suitable clusters of contiguous population units, and surveying all the units in a sample of clusters selected according to an appropriate sampling scheme. (Singh & Singh Mangat, 2010,248)

A cluster sample is a sampling method where the researcher divides the entire population into separate groups, or clusters. Then, a random sample of these clusters is selected. All observations within the chosen clusters are included in the sample. This method is typically used when the population is large, widely dispersed, and inaccessible. The clusters should ideally mirror the characteristics of the population as a whole.

Suppose that the company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

another example: You're investigating how satisfied people in your town are with dog parks.You divide the town into multiple neighborhoods and draw a random sample from each neighborhood to collect data using a survey. (Simkus, 2023)

Steps to Conduct Cluster Sampling: Here are the steps to perform cluster sampling:First Decide the target audience and also the sample size and Create a sampling frame by using either an existing framework or creating a new one for the target audience. Evaluate frameworks based on coverage and clustering and make adjustments accordingly. These groups will be varied, considering the population, which can be exclusive and comprehensive. Members of a sample are selected individually.Then determine the number of groups by including the same average members in each group. So make sure each of these groups is distinct from one another.and choose clusters by applying a random selection. It is bifurcated into two-stage and multi-stage subtypes based on the number of steps followed by researchers to form clusters. (Simkus, 2023)

Cluster sampling is used when the target population is too large or spread out, and studying each subject would be costly, time-consuming, and improbable. (Simkus, 2023)

This sampling technique is used in an area or geographical cluster sampling for market research. A broad geographic location can be expensive to survey compared to surveys sent to clusters divided based on region. The sample numbers must be increased to achieve accurate results, but the cost savings make this process of rising clusters attainable. (Singh & Singh Mangat, 2010)

Types of Cluster Sampling:

Here are the types of Cluster sampling:

1. **Single-stage cluster sampling**

   This means that all the elements in the selected clusters are viewed as a sampling unit, hence the name a single-stage cluster. First, researchers will split the total sample as needed into the number of clusters they want them to be based on how big each cluster should be. They then take a very small random sample of the clusters and within the selected clusters they measure every unit. (Simkus, 2023)

2. **Double-stage cluster sampling**

   In two-stage cluster sampling, the sample consists of a random subsample of individual units from all the units selected in each of the clusters. This method will be less accurate than one-stage sampling, and should only be employed when it is impractical, costly, or time-consuming to sample and test the entire cluster. (Simkus, 2023)

3. **Multi-stage cluster sampling**

   This method of cluster sampling includes the same procedure as double-stage sampling, with a few additional steps. In multi-stage sampling researchers will continue to randomly draw elements from the clusters until we can obtain a manageable sample size. (Simkus, 2023)

2.2 Unequal Probability Sampling:

Most survey designs are typically based on the selection of N units at random with equal probabilities and without replacement from a population of N units(denoted by $U = \{1, 2, ..., N\}$. Indeed, in many circumstances, it is more advantageous to select the units with unequal probabilities. The following specific approach is very applicable when for all units in the population, $i = 1, 2, ..., N$ there is available a measure of size (x_i) thought to be related to the characteristic of interest (y_i),□ for which one desires an estimate of its total for the whole population (denoted by Y). (Hartley & Rao, n.d., #)

One common way the size measures x_i can be used is through the selection of units with probabilities proportional to their size, x_i which is referred to as **Probability Proportional to Size (PPS)**. This is very often done in sample surveys through the selection of primary sampling units in multi-stage sampling designs. (Hartley & Rao, n.d., #)

The theory of unequal probability sampling agrees with that of multinomial sampling when sampling is carried out with replacement. It is however known from the theory of equal-probability sampling that sampling with replacement results in less precise estimates compared to estimators based on samples drawn without replacement. Precision gained by not replacing units during sampling is in fact measured by the **finite population correction** factor $\frac{n}{N}$ (Hartley & Rao, n.d., #)

This created the expectation that equal gains in precision using unequal probability sampling without replacement were possible. However, unequal probability sampling without replacement had a much harder theoretical development fraught with serious mathematical and computational complications and is incomplete to this date. (Hartley & Rao, n.d., #)

Usually the objective is to estimate the total of some variable y which has value y_ii for unit i. Thus we want to estimate $Y = \sum_{i=1}^{N} y_i$ All the y_i values are unknown before a sample has been selected. In order to use unequal probability sampling, we need some auxiliary information. It is often the case that we know the value of another variable $x_i > 0$ for each unit $i \epsilon u$, and we suspect that y is approximately proportional to x_i. The following example illustrates one possible situation. (Grafström, 2010)

If we want to estimate the total amount of pollution from a set of factories, then we may know or strongly suspect that larger factories generate more pollution than smaller ones. If we have some auxiliary information x about the size of the factories, then that information can be used. Such information may be the number of employees, the size of the buildings, or the number of units produced last year, and so on. In this situation, we want to sample large factories with higher probabilities than small factories since large factories will contribute more to the total amount of pollution. By doing so, we can get a much better estimate than if the factories are selected with equal probabilities. (Grafström, 2010)

The information available before the sample is selected includes the labels, $i = 1,2, ..., N$, of the units, and the value of z_i for each unit i. The aim is to choose each unit with a probability $\pi_i = c\, z_i$, where c is a positive constant.

In many applications, the preference is to take samples of fixed size n as this design often leads to more efficient estimators and can control the costs of the sample collection. For a fixed sample size, say n, it has to be guaranteed that $\sum_{i=1}^{N} \pi_i = n$.

Suppose the sampling probabilities $\frac{\pi_i}{p_i}$ are pre-specified satisfying $\sum_{i=1}^{N} \pi_i = n$, then draw the sample according to these sampling probabilities. (Grafström, 2010)

In survey sampling, Hansen-Hurwitz estimators and Horvitz-Thompson estimators are widely used for the estimation of population totals or means when the design is unequal probability sampling. A short overview of such estimators is given below.

- **The Hansen-Hurwitz Estimator :**

  The Hansen-Hurwitz Estimator (HHE), has been proposed in case of sampling with replacement. Let consider an ordered sample of m independent draws (k_1,$k_2 \cdots k_m$). At each draw, the probability of selecting an individual k is p_k. (Osier, 2025)

$$\hat{T}_{HH} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k}$$

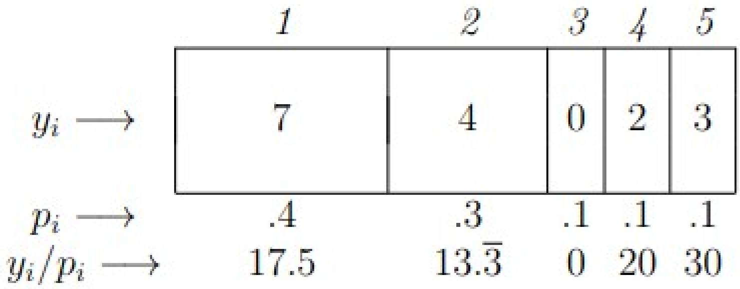
The variance of $\hat{T}_{HH}$ is given by:

$$var(\hat{T}_{HH}) = \frac{V_1}{m}$$

Where:

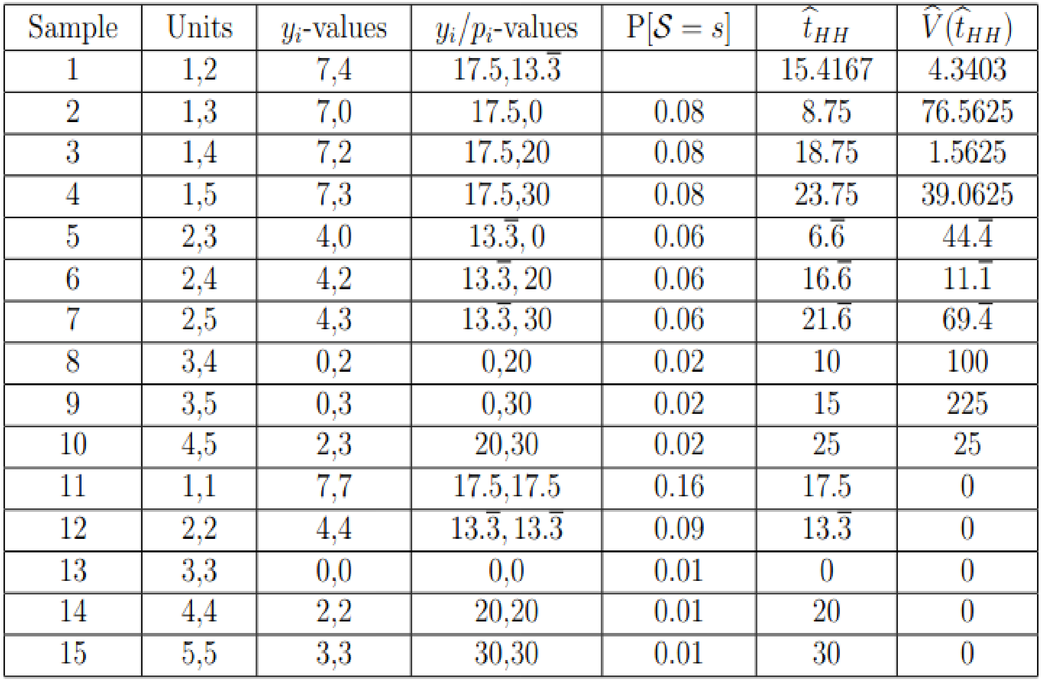
- $\hat{T}_{HH}$: The Hansen-Hurwitz estimate of the population total
- y_k: The value of the characteristic of interest for unit k
- p_k : the probability of selecting an individual k
- m : The the sample size
- $V_1 = \sum_{i \epsilon U}^{\square} p_i \left(\frac{y_i}{p_i} - Y \right) \square^2$

Example: Hansen-Hurwitz estimation with selection probabilities p_i that are proportional to unit size and with p_i (approximately) directly proportional to y_i The total abundance t = 16. There are N = 5 sampling units. The figure shows the units,labeled 1 to 5, and the five y_i values. (Borkowski, n.d.)



	1	2	3	4	5
$y_i \longrightarrow$	7	4	0	2	3
$p_i \longrightarrow$	.4	.3	.1	.1	.1
$y_i/p_i \longrightarrow$	17.5	$13.\overline{3}$	0	20	30

Table 2.1: Hansen-Hurwitz estimator Table



Sample	Units	y_i-values	y_i/p_i-values	P[$\mathcal{S} = s$]	$\widehat{t}_{HH}$	$\widehat{V}(\widehat{t}_{HH})$
1	1,2	7,4	$17.5, 13.\overline{3}$		15.4167	4.3403
2	1,3	7,0	17.5,0	0.08	8.75	76.5625
3	1,4	7,2	17.5,20	0.08	18.75	1.5625
4	1,5	7,3	17.5,30	0.08	23.75	39.0625
5	2,3	4,0	$13.\overline{3}, 0$	0.06	$6.\overline{6}$	$44.\overline{4}$
6	2,4	4,2	$13.\overline{3}, 20$	0.06	$16.\overline{6}$	$11.\overline{1}$
7	2,5	4,3	$13.\overline{3}, 30$	0.06	$21.\overline{6}$	$69.\overline{4}$
8	3,4	0,2	0,20	0.02	10	100
9	3,5	0,3	0,30	0.02	15	225
10	4,5	2,3	20,30	0.02	25	25
11	1,1	7,7	17.5,17.5	0.16	17.5	0
12	2,2	4,4	$13.\overline{3}, 13.\overline{3}$	0.09	$13.\overline{3}$	0
13	3,3	0,0	0,0	0.01	0	0
14	4,4	2,2	20,20	0.01	20	0
15	5,5	3,3	30,30	0.01	30	0

- **The Horvitz-Thompson Estimator :**

The Horvitz-Thompson Estimator (HTE), proposed by Horvitz and Thompson (1952), is a statistical technique designed to estimate population totals or means in cases where sampling occurs with unequal probabilities and without replacement. By accounting for the unequal inclusion probabilities of the sampled units, this estimator ensures that the resulting estimate remains unbiased. (Horvitz & Thompson, 1952) It showed that the only linear unbiased estimator for any without replacement sampling algorithm, where the inclusion probabilities are well defined as π_i for i = 1, . . . , N , was

$$\hat{T}_{HT} = \sum_{i \epsilon s}^{\square} \frac{y_i}{\pi_i}$$

Horvitz and Thompson also showed the variance of this estimator to be and note that $i \neq j$

$$var(\hat{T}_{HT}) = \frac{1}{2} \sum_{i \epsilon U}^{\square} \sum_{j \epsilon U}^{\square} (\pi_i \pi_j - \pi_{ij}) \frac{y_i y_j}{\pi_i \pi_j}$$

Where:

- $\hat{T}_{HT}$: The Horvitz-Thompson estimate of the population total
- y_i: The value of the characteristic of interest for unit i
- s : The set of sampled units
- π_i : The probability of selecting ith units in the sample
- π_j : The probability of selecting jth units in the sample
- π_{ij} : The probability of selecting pair of ith and jth units in the sample

Let's introduce the following dummy variable δi=1 if i∈S and δi=0 otherwise. δi is a random variable whose expectation is given by:

E(δi) = πiHence, we have: $\sum_{i \epsilon U}^{\square} \square \pi_i$=$\sum_{i \epsilon U}^{\square} \square$E(δi)=E($\sum_{i \epsilon U}^{\square} \square$δi)=E($n_s$)

Result : Assuming π_k>0 ∀k∈U, $\hat{T}_{HT}$ is an unbiased estimator of Y

The proof is easy after the HT is rewritten as a sum over all population elements using the δi dummies:

$$\hat{T}_{HT} = \sum_{k \epsilon s}^{\square} \frac{y_k}{\pi_k} = \sum_{k \epsilon U}^{\square} \frac{y_k}{\pi_k} \delta k$$

Then the expectation is given by:

$$E(\hat{T}_{HT}) = \sum_{k \epsilon u}^{\square} \frac{y_k}{\pi_k} E(\delta k) = \sum_{k \epsilon u}^{\square} \frac{y_k}{\pi_k} \pi_k = Y$$

The preceding result illustrates that unequal probability sampling can produce estimators with higher precision compared to simple random sampling or other equal-probability methods. Although this may initially appear counterintuitive to those less familiar with survey sampling, it emphasizes an important principle: utilizing auxiliary information can greatly improve sampling precision.

Example: Horvitz-Thompson estimation with inclusion probabilities π_i that are proportional to size and πi is (approximately) directly proportional to y_i, The total abundance t = 16. There are N = 5 sampling units. The figure shows the units,labeled 1 to 5, and the five y_i values. (Borkowski, n.d.)

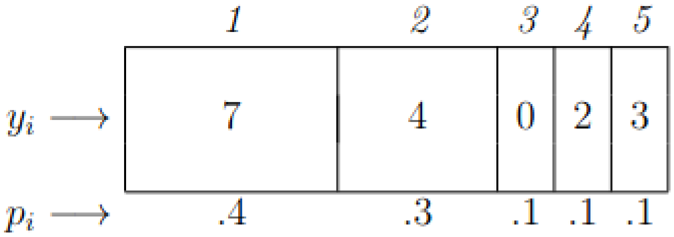
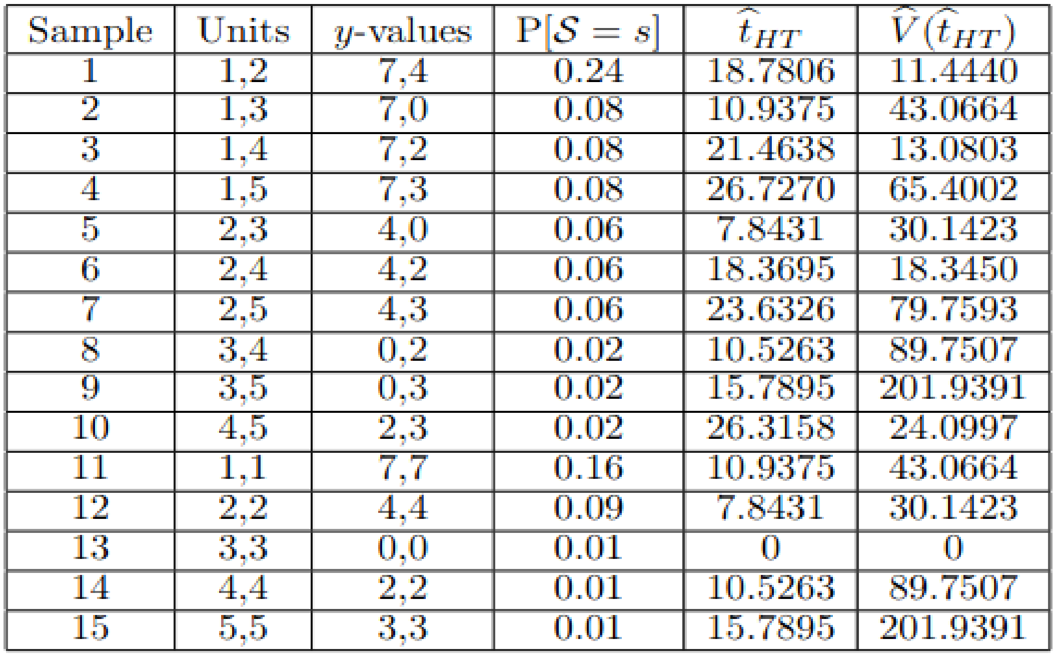


Table 2.2: Hansen-Hurwitz estimator Table



Sample	Units	y-values	$P[\mathcal{S} = s]$	$\widehat{t}_{HT}$	$\widehat{V}(\widehat{t}_{HT})$
1	1,2	7,4	0.24	18.7806	11.4440
2	1,3	7,0	0.08	10.9375	43.0664
3	1,4	7,2	0.08	21.4638	13.0803
4	1,5	7,3	0.08	26.7270	65.4002
5	2,3	4,0	0.06	7.8431	30.1423
6	2,4	4,2	0.06	18.3695	18.3450
7	2,5	4,3	0.06	23.6326	79.7593
8	3,4	0,2	0.02	10.5263	89.7507
9	3,5	0,3	0.02	15.7895	201.9391
10	4,5	2,3	0.02	26.3158	24.0997
11	1,1	7,7	0.16	10.9375	43.0664
12	2,2	4,4	0.09	7.8431	30.1423
13	3,3	0,0	0.01	0	0
14	4,4	2,2	0.01	10.5263	89.7507
15	5,5	3,3	0.01	15.7895	201.9391

The inclusion probabilities are

$$\pi_1 = .24 + 3(.08) + .16 = .64,$$

π_2 = .24 + 3(.06) + .09 = .51, and

π_3 = π_4 = π_5 = .08 + .06 + 2(.02) + .01 = .19.

1. Cumulative Total Method:

Let there be N units in the population and the auxiliary variable X has values $X_1, X_2, X_3, \ldots, X_N$,respectively, on these population units, serially arranged as the first, second, third, …, Nth unit, which are known by some means.

Let the total of X_i Values be given by $X = \sum_{i=1}^{N} X_i$

The cumulative total method is essentially based on assigning each of the unit a set of consecutive natural numbers such that the number of numbers in the set is equal to the size of the unit(Singh & Kumar, n.d., #). Let us arrange the unit labels (i), the values of the study variable (Yi), the corresponding sizes of units (Xi), Cumulative Totals (Ti) and selection probabilities (pi) serially as shown in the following table:

Table 2.3: Cumulative Total Method Table

Serial Number of Units (Labels i)	**Values of the Study Variable** (Yi)	**Size Measure of Units** (Xi)	**Cumulative Totals** (Ti)	**Probability of Selection** $pi = (Xi/X)$
1	Y_1	X_1	$T_1 = X_1$	X_1/X
2	Y_2	X_2	$T_2 = X_1 + X_2$	X_2/X
3	Y_3	X_3	$T_3 = X_1 + X_2 + X_3$	X_3/X
4	Y_4	X_4	$T_4 = X_1 + X_2 + X_3 + X_4$	X_4/X
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
N	Y_N	X_N	$T_N = X_1 + X_2 + \ldots + X_N$	X_N/X

Total	**—**	$X = \sum_{i=1}^{N} X_i$	**—**	**1.00**

Example:

a village has 8 orchards containing 50, 30, 25, 40, 26, 44, 20 and 35 trees respectively. A sample of 3 orchards has to be selected with replacement and with probability proportional to the number of trees in the orchards. We prepare the following cumulative total table(Jayaraman, n.d.):

Table 2.4: Cumulative Total Method Example

Serial number of the orchard	Size (x_i)	Cumulative size	Numbers associated
1	50	50	1 - 50
2	30	80	51 - 80
3	25	105	81 -105
4	40	145	106 -145
5	26	171	146 - 171
6	44	215	172 - 215
7	20	235	216 - 235
8	35	270	236 - 270

Now, we select three random numbers between 1 and 270. The random numbers selected are 200, 116 and 47. The units associated with these three numbers are 6th, 4th, and 1st respectively. And hence, the sample so selected contains units with serial numbers, 1, 4 and 6. (Jayaraman, n.d.)

The cumulative total method drawback that this procedure involves writing down the successive cumulative totals. This is time consuming and tedious if the number of units in the population is large,This problem is overcome in Lahiri's method. (Singh & Kumar, n.d., #)

2. Lahiri's Method:

We have noticed that the cumulative total method involves writing down the successive cumulative totals which is time consuming and tedious, especially with large populations. Lahiri in 1951 suggested an alternative procedure which avoids the necessity of writing down the cumulative totals. Lahiri’s method consists in selecting a pair of random numbers, say (i,j) such that i from 1 to N and 1 $\leq J \geq M$; where M is the maximum of the sizes of the N units of the population.

If $j \leq X_i$, the **i th unit is selected** otherwise, the pair of random numbers is rejected and another pair is chosen. For selecting a sample of n units, the procedure is to be repeated till n units are selected. This procedure leads to the required probabilities of selection. (Lahiri, 1951)

For instance, to select a sample of 3 orchards from the population in the previous example in this section, by Lahiri’s method by PPS with replacement, as N = 8,

M = 50 and n = 3, we have to select three pairs of random numbers such that the first random number is less than or equal to 8 and the second random number is less than or equal to 50. Referring to the random number table, three pairs selected are (2, 23) (7,8) and (3, 30). As in the third pair j >X_i , a fresh pair has to be selected. The next pair of random numbers from the same table is (2, 18) and hence, the sample so selected consists of the units with serial numbers 2, 7 and 2. Since the sampling unit 2 gets repeated in the sample, the effective sample size is two in this case. In order to get an effective sample size of three, one may repeat the sampling procedure to get another distinct unit. (Lahiri, 1951)

Lahiri (1951) introduced a new method, which does not need cumulative totals, for selecting a PPSWR sample, but in this method, we need to know the maximum value of X_i, which we denote by X_0. Sometimes it is not possible to know the maximum value of X_i, e.g., $X_{\square}$ = Number of errors in a book. In such cases, we choose X_0 to be more than the maximum among all values of X_i. In other words, we choose X_i such that $X_0 \geq Max(X_1, X_2, \ldots, X_N)$.

using Lahiri's method are as follows:

1. Select a random number R_i such that $1 \leq R_i \leq N$.
2. Select another random number R_j such that $1 \leq R_j \leq X_0$.
3. Compare the magnitude of the random number R_j with that of X_R. If the magnitude of $R_j \leq X_{Ri}$, then the unit with serial number R_i is selected to be

included in the sample; otherwise, it is rejected. The process is continued until we select a random sample of the desired size. (Lahiri, 1951)

The advantages of Lahiri's method that it does not require writing down all cumulative totals for each unit and Sizes of all the units need not be known beforehand. We need only some number greater than the maximum size and the sizes of those units which are selected by the choice of the first set of random numbers 1 to N for drawing samples under this scheme.

However, there are disadvantage associated with Lahiri's method. It results in the wastage of time and efforts if units get rejected.A draw is ineffective if one of the ineffective random numbers is selected.

Theorem :

Probability of selection of the iii-th unit for Lahiri's method is P_i , for i=1,…,N (Arnab, 2017)

Proof:

Let a_i Probability that the first pair of random numbers selects unit i.

=Probability of selection of random number i $(1 \leq i \leq N)$

and a random number R less than or equal to X_i = X_i/NM:

And let b_i Probability that the first pair of random numbers fails to select any unit:

$$= \sum_{i=1}^{N} \frac{1}{N} \frac{M - x_i}{M} = 1 - \frac{X}{NM}$$

Hence the probability of selection of i-th unit using this method is:

=Prob(First pair of random numbers selects i-th unit)+

Prob(First pair of random numbers fails to select any unit)

×Prob(Second pair of random numbers selects i-th unit)+⋯

¿$a_i + b\,a_i + b^2 a_i + \cdots = a_i(1-b)-1 = x_i \,/ X = p_i$

Example:

Select a sample of four households using Lahiri's method for the data given in Table 2.3. Here the population size is N=9 and we may take $M=6\geq Max\{x_i\}$ $1\leq i\leq 9$. (Arnab, 2017, 120)

Table 2.5: Lahiri's Method Example

Random number between 1 and 9 (j)	4	5	3	2	5
Random number between 1 and 6 (R)	6	2	4	2.5	2.5
x_i	4	2.5	3	3.5	2.5
Unit selected	-	5	3	2	5

So, the selected ordered sample by Lahiri's method is s_0=(5,3,2,5)

3. Splitting Method:

A broad class of sampling methods without replacement and with unequal probabilities is defined. In this approach, the inclusion probability vector is divided into several new inclusion probability vectors. One of them is selected at random, reducing the original problem of sampling without replacement with unequal probabilities to a simpler unequal probability problem. This process of splitting is done iteratively with the inclusion probability vectors, progressively simplifying the sampling problem at each step.

The flexibility of this approach enables the creation of new unequal probability sampling procedures with ease. Furthermore, the splitting method generalizes other well-known procedures, such as the Midzuno method, the elimination procedure, and the Chao procedure.

A sufficient condition is provided to ensure that the splitting method satisfies the Sen–Yates–Grundy condition. It is further demonstrated that the elimination procedure adheres to the sufficient condition established by Gabler.(Jean-Claude & Tillé, n.d., #)

Splitting into Two Vectors

The fundamental idea behind this technique is straightforward: each π_k is divided into two parts $\pi_k^{(1)}$ and $\pi_k^{(2)}$ that satisfy the following relations:

$$\pi_k=\lambda\pi_k^{(1)}+(1-\lambda)\pi_k^{(2)};0\leq\pi_k^{(1)}\leq 1,0\leq\pi_k^{(2)}\leq 1,\sum_{k\epsilon U}^{\square}\pi_k^{(1)}=\sum_{k\epsilon U}^{\square}\pi_k^{(1)}=n,$$

where λ is a parameter that can be freely chosen, provided $0<\lambda<1$. The method consists of selecting nn units with unequal probabilities, reducing the problem to another sampling problem with unequal probabilities. If the splitting process results in some $\pi_k^{(1)}$and the$\pi_k^{(2)}$ being equal to 0 or 1, the sampling problem becomes simpler in subsequent steps since the splitting will then apply to a smaller population.(Jean-Claude & Tillé, n.d., #)

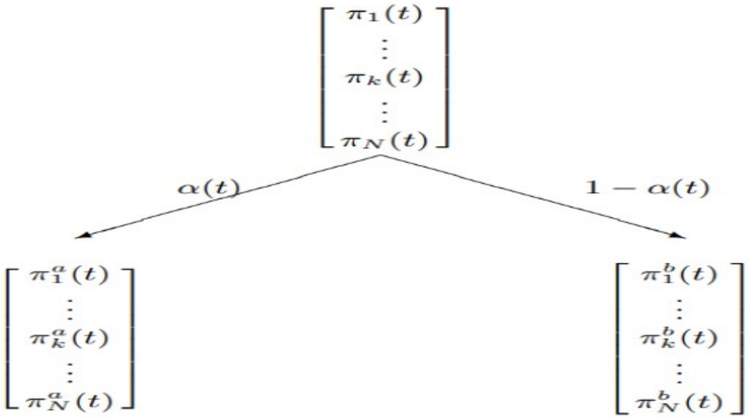


Figure 2.1 : Splitting into Two Vectors

Minimum Support Design

Wynn (1977, Theorem 1) proved it is always possible, for any vector of preassigned first-order inclusion probabilities, to construct a sampling design that involves only N samples s such that $p(s)$. Any such design is referred to as the **minimal support design**. However, Wynn's proof is not constructive.

The minimal support design corresponds to the most direct application of the two-part splitting technique and is closely related to the simplest "game"

proposed by Hedayat et al. (1989, Theorem 2.1) in the context of a procedure for emptying boxes.

Let $\pi_{(1)},.....,\pi_{(k)},.....,\pi_{(N)}$ denote the ordered inclusion probabilities. Then define:

$$\lambda=min\left(1-\lambda_{(N-n)},\lambda_{(N-n+1)}\right)$$

$$\pi_{(k)}^{(1)}=\begin{cases}0 & \text{if } k\leqslant N-n,\\ 1 & \text{if } k>N-n,\end{cases}\quad \pi_{(k)}^{(2)}=\begin{cases}\pi_{(k)}/(1-\lambda) & \text{if } k\leqslant N-n,\\ (\pi_{(k)}-\lambda)/(1-\lambda) & \text{if } k>N-n.\end{cases}$$

(Jean-Claude & Tillé, n.d., #)

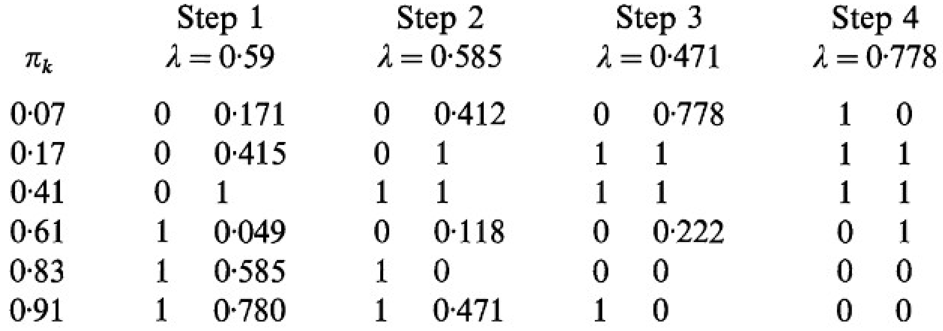
Example(1). Consider $N=6,n=3$, and $\pi=(0.07,0.17,0.41,0.61,0.83,0.91)$. In this case, the breakdown is completed in 4 steps. The vector of inclusion probabilities is divided into two parts, as shown in columns 2 and 3 of Table below . With probability $\lambda=0.59$, the sample $\{4,5,6\}$ is selected and, with probability $1-\lambda=0.411$, Another sampling is applied with unequal probabilities given by $(0.171,0.415,1,0.049,0.585,0.780)$. At step 2, the splitting is applied again to this vector, and, in 4 steps, the sample is selected. The sampling design is thus given by: (Tillé, 2006, #)

$$p(\{4,5,6\})=0.59\,,p(\{3,5,6\})=(1-0.59)\times 0.585=0.24\,,$$

$$p(\{2,3,6\})=(1-0.59-0.24)\times 0.471=0.08\,,$$

$$p(\{1,2,3\})=(1-0.59-0.24-0.08)\times 0.778=0.07\,,$$

$$p(\{2,3,4\})=(1-0.59-0.24-0.08-0.07)=0.02.$$



	Step 1		Step 2		Step 3		Step 4	
π_k	$\lambda=0{\cdot}59$		$\lambda=0{\cdot}585$		$\lambda=0{\cdot}471$		$\lambda=0{\cdot}778$	
0·07	0	0·171	0	0·412	0	0·778	1	0
0·17	0	0·415	0	1	1	1	1	1
0·41	0	1	1	1	1	1	1	1
0·61	1	0·049	0	0·118	0	0·222	0	1
0·83	1	0·585	1	0	0	0	0	0
0·91	1	0·780	1	0·471	1	0	0	0

<u>Splitting into Simple Random Sampling</u>

This method divides the inclusion probabilities into two parts. Define:

$$\lambda=min\{\frac{\pi_{(1)}*N}{n}\,,\frac{N}{N-n}(1-\pi_{(N)})\},$$

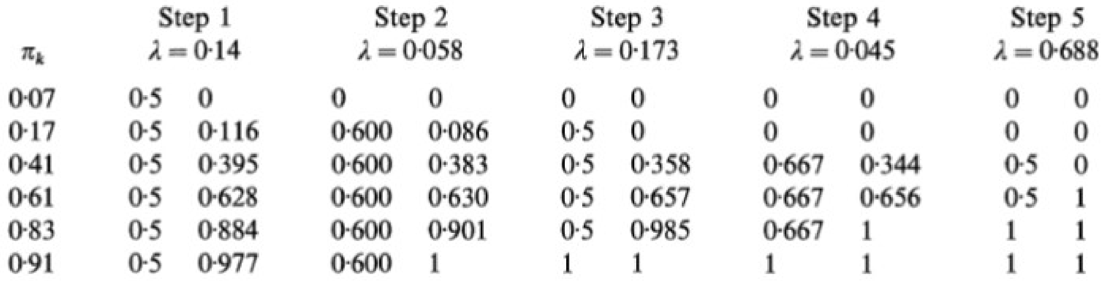
and compute, for $k\epsilon U$,

$$\pi_{(k)}^{(1)}=\frac{n}{N},\pi_{(k)}^{(2)}=\frac{N}{N-n}(1-\pi_{(N)}).$$

If $\lambda=\pi_{(1)}.\frac{N}{n}$ then $\pi_{(1)}^{(2)}=0$; and if $\lambda=\left(1-\pi_{(N)}\right).\frac{N}{N-b}$, then $\pi_{(N)}^{(2)}=1$.

At the next step, the problem reduces to selecting a sample of size $n-1$ or n from a population of size $N-1$ The problem can be fully resolved within at most $N-1$ steps. (Jean-Claude & Tillé, n.d., #)

Example(2) Table below presents the method's outcome using the same $\pi_{(k)}$ as in Example1.(Tillé, 2006, #)



	Step 1		Step 2		Step 3		Step 4		Step 5	
π_k	$\lambda=0{\cdot}14$		$\lambda=0{\cdot}058$		$\lambda=0{\cdot}173$		$\lambda=0{\cdot}045$		$\lambda=0{\cdot}688$	
0·07	0·5	0	0	0	0	0	0	0	0	0
0·17	0·5	0·116	0·600	0·086	0·5	0	0	0	0	0
0·41	0·5	0·395	0·600	0·383	0·5	0·358	0·667	0·344	0·5	0
0·61	0·5	0·628	0·600	0·630	0·5	0·657	0·667	0·656	0·5	1
0·83	0·5	0·884	0·600	0·901	0·5	0·985	0·667	1	1	1
0·91	0·5	0·977	0·600	1	1	1	1	1	1	1

The Pivotal Method

The pivotal method involves splitting the vector of inclusion probabilities into two parts. In this approach, only two inclusion probabilities are altered in each step. The method selects two units, denoted as i and j , and modifies their inclusion probabilities accordingly. (Jean-Claude & Tillé, n.d., #)

If $\pi i+\pi j>1$, then $\lambda=\frac{(1-\pi j)}{(2-\pi i-\pi j)}$:

$$\pi_k^{(1)}=\begin{cases}\pi_k & \text{if } k\in U\backslash\{i,j\},\\ 1 & \text{if } k=i,\\ \pi_i+\pi_j-1 & \text{if } k=j,\end{cases}\qquad \pi_k^{(2)}=\begin{cases}\pi_k & \text{if } k\in U\backslash\{i,j\},\\ \pi_i+\pi_j-1 & \text{if } k=i,\\ 1 & \text{if } k=j.\end{cases}$$

If $\pi i+\pi j<1$, then $\lambda=\frac{\pi i}{(\pi i+\pi j)}$:

$$\pi_k^{(1)}=\begin{cases}\pi_k & \text{if } k\in U\backslash\{i,j\},\\ \pi_i+\pi_j & \text{if } k=i,\\ 0 & \text{if } k=j,\end{cases}\qquad \pi_k^{(2)}=\begin{cases}\pi_k & \text{if } k\in U\backslash\{i,j\},\\ 0 & \text{if } k=i,\\ \pi_i+\pi_j & \text{if } k=j.\end{cases}$$

In the first case, the value of one is assigned to only a single inclusion probability. In the second case, the value of zero is assigned to only one inclusion probability. This process effectively reduces the population size to$N-1$. Within a maximum of N steps, the procedure provides a solution. This method is notable for its simplicity and can be executed in a fully sequential manner, such as by performing a single pass through a data file. However, a limitation of this approach is that many of the joint inclusion probabilities are equal to zero.

To address this issue, the procedure can be randomized, similar to systematic sampling, by selecting the units i and j randomly.

Chapter Three
Methodology & Implementation

3.1 Study Design:

A study design serves as a blueprint for conducting research or analysis, outlining how data is collected, the methods employed, and the approach to data analysis. For the given sampling methods—**Lahiri's Method**, **Cumulative Totals Method**, and **Splitting Method**—the study design focuses on **Probability Proportional to Size Sampling (PPS)**.

This research involved a literature review of existing studies and articles on methods for selecting unequal probability samples. The objective was to develop a comprehensive understanding of these methods and identify scenarios where each is most effectively applied. Additionally, a simulated dataset was generated in R, serving as a practical example to demonstrate and compare the performance of these sampling methods. The analysis included applying each method to the dataset and evaluating their results.

3.2 Data Source:

Simulated datasets are artificial data that are primarily created to resemble the real world and its phenomena for analytical testing purposes. Therefore, unlike secondary data, which is usually pre-existing and collected for other purposes, simulated datasets are specifically tailored according to a study's or experiment's requirements. A simulated dataset was generated in R to serve as a controlled example for applying and comparing the sampling methods. The dataset included EmployeeID, Age, Department, Salary, YearsExperience.

3.3 Data Analysis Methods:

The sampling methods—Lahiri's Method, Cumulative Totals Method, and Splitting Method—were applied to analyze the simulated dataset. The implementation was conducted using the R programming system, utilizing specific add-in software packages called **dplyr**, which is useful for performing data manipulation operations, and the **SDaA Package** for sampling techniques.

Key sampling parameters, such as selection probabilities and cumulative totals, were used to evaluate the efficiency and effectiveness of each method. The variance of the sampled data was calculated to compare the methods, with lower variance indicating higher efficiency. Additionally, random number generation and probability proportional to size (PPS) techniques were implemented to ensure fairness in the selection process.

The analysis shed light on the representativeness of each sampling method and their relative strengths in dealing with datasets of unequal probabilities.

3.4 Application using Software:

In this section we will take a look at how we can apply Methods in R.

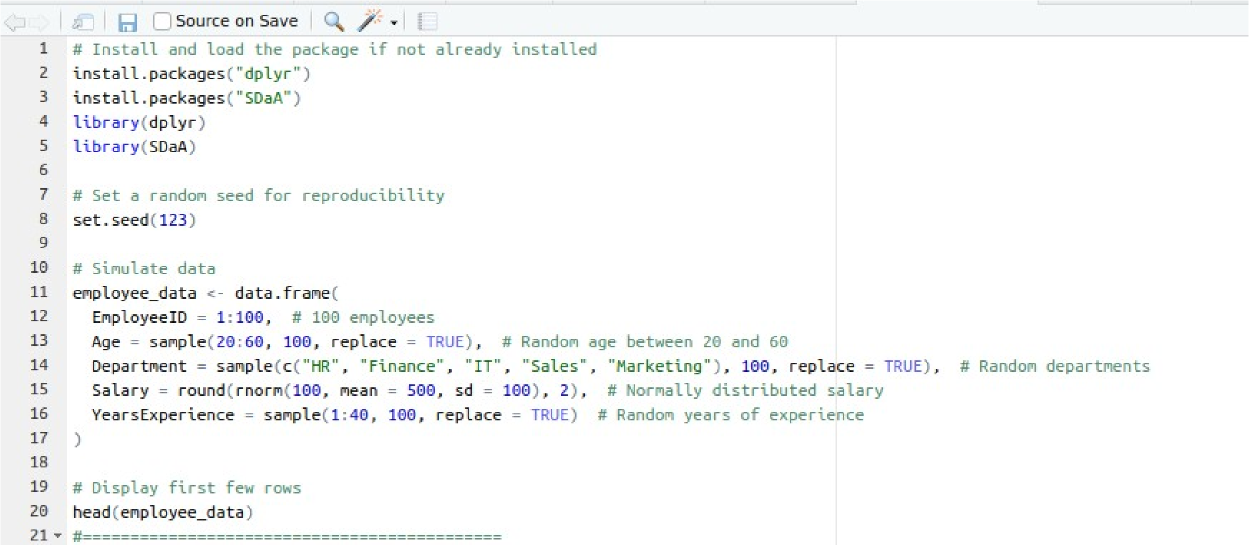
First of all we simulate a dataset about employees in the company with their details salary information and years of experience in R .

The following R code demonstrates making a simulated dataset for employees. First we use the dplyr library, a popular data manipulation tool. A random seed is set using the command set.seed(123) to ensure reproducible results, making the randomly created data consistent between runs.

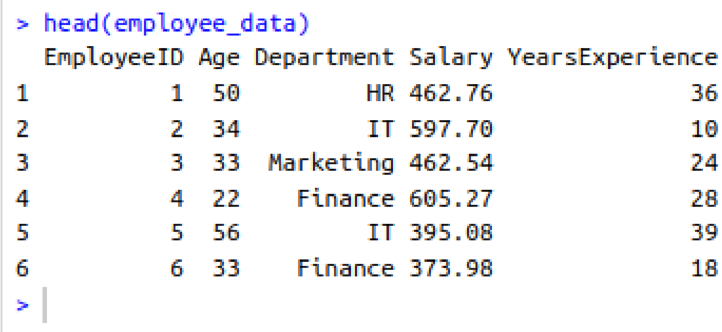
The dataset is created using the data.frame() function with the following variables in employee_data:

1. **EmployeeID**: A unique identifier for each employee, represented by a sequence of integers from 1 to 100.
2. **Age**: Randomly generated ages ranging from 20 to 60, using the sample() function.
3. **Department**: Randomly assigned department names (such as HR, Finance, IT, Sales, and Marketing) for each employee, also generated with sample().
4. **Salary**: Normally distributed salaries with a mean of 50,000 and a standard deviation of 10,000, generated using rnorm(). The values are rounded to two decimal places for precision.
5. **YearsExperience**: Randomly assigned years of experience between 1 and 40.

Finally, the code concludes by displaying the first few rows of the dataset using the head() function. This dataset serves to simulate employee information for analysis or testing purposes.

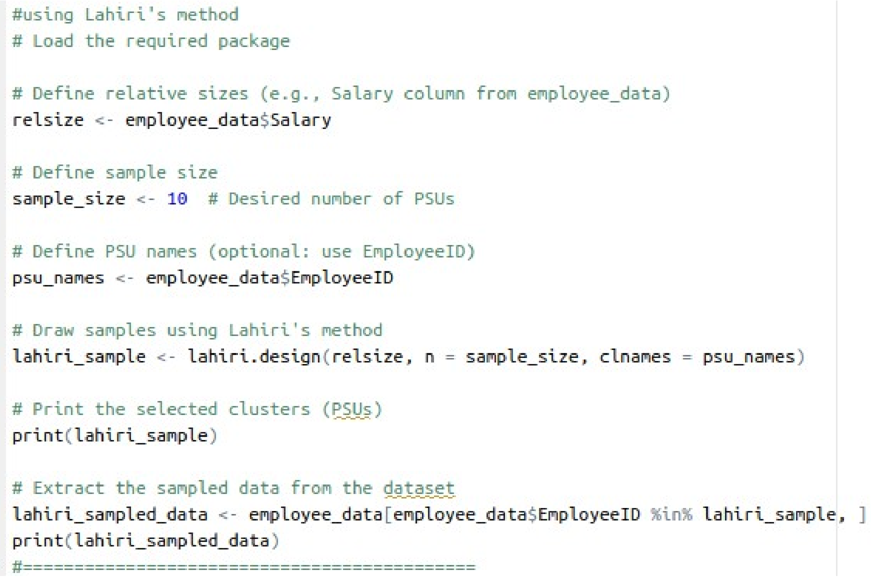


```r
# Install and load the package if not already installed
install.packages("dplyr")
install.packages("SDaA")
library(dplyr)
library(SDaA)

# Set a random seed for reproducibility
set.seed(123)

# Simulate data
employee_data <- data.frame(
  EmployeeID = 1:100,  # 100 employees
  Age = sample(20:60, 100, replace = TRUE),  # Random age between 20 and 60
  Department = sample(c("HR", "Finance", "IT", "Sales", "Marketing"), 100, replace = TRUE),  # Random departments
  Salary = round(rnorm(100, mean = 500, sd = 100), 2),  # Normally distributed salary
  YearsExperience = sample(1:40, 100, replace = TRUE)  # Random years of experience
)

# Display first few rows
head(employee_data)
#============================================
```

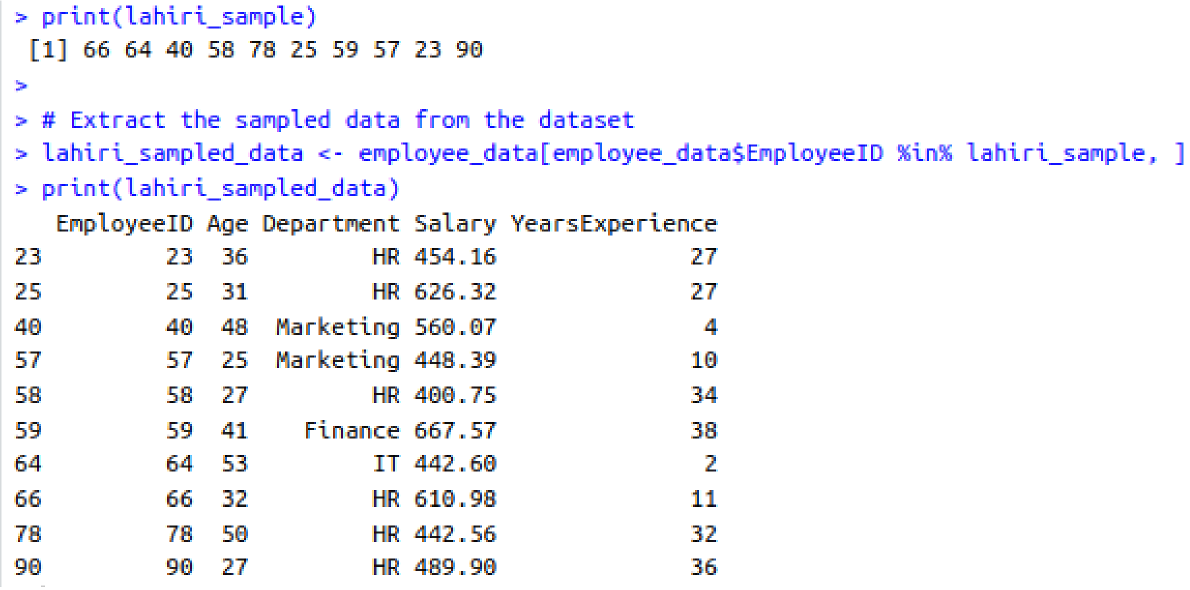


```
> head(employee_data)
  EmployeeID Age Department Salary YearsExperience
1          1  50         HR 462.76              36
2          2  34         IT 597.70              10
3          3  33  Marketing 462.54              24
4          4  22    Finance 605.27              28
5          5  56         IT 395.08              39
6          6  33    Finance 373.98              18
> 
```

Next, we use **Lahiri's Method** to sample employees based on their salaries with the help of the SDaA package. The process starts by defining the relative sizes (relsize), which refer to the Salary column in the employee_data dataset, representing the importance of each employee in the sampling process. A sample size of 10 employees is chosen, ensuring 10 individuals are selected. The EmployeeID column is used to uniquely identify each employee. The lahiri.design() function is then applied to create a sample of employees, with selection probabilities aligned with their respective salaries. The selected EmployeeIDs are stored in lahiri_sample, and the corresponding rows from employee_data are extracted into lahiri_sampled_data.

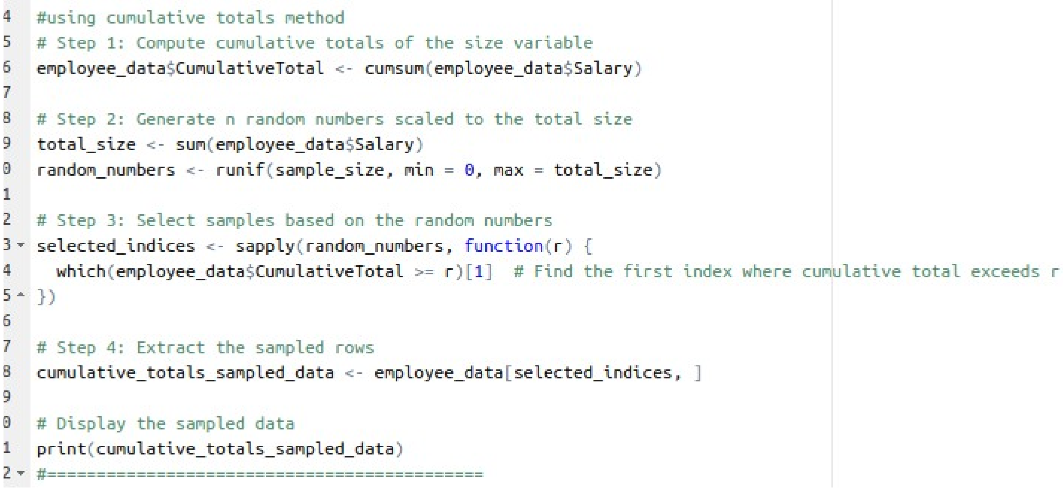


```
#using Lahiri's method
# Load the required package

# Define relative sizes (e.g., Salary column from employee_data)
relsize <- employee_data$Salary

# Define sample size
sample_size <- 10  # Desired number of PSUs

# Define PSU names (optional: use EmployeeID)
psu_names <- employee_data$EmployeeID

# Draw samples using Lahiri's method
lahiri_sample <- lahiri.design(relsize, n = sample_size, clnames = psu_names)

# Print the selected clusters (PSUs)
print(lahiri_sample)

# Extract the sampled data from the dataset
lahiri_sampled_data <- employee_data[employee_data$EmployeeID %in% lahiri_sample, ]
print(lahiri_sampled_data)
#============================================
```

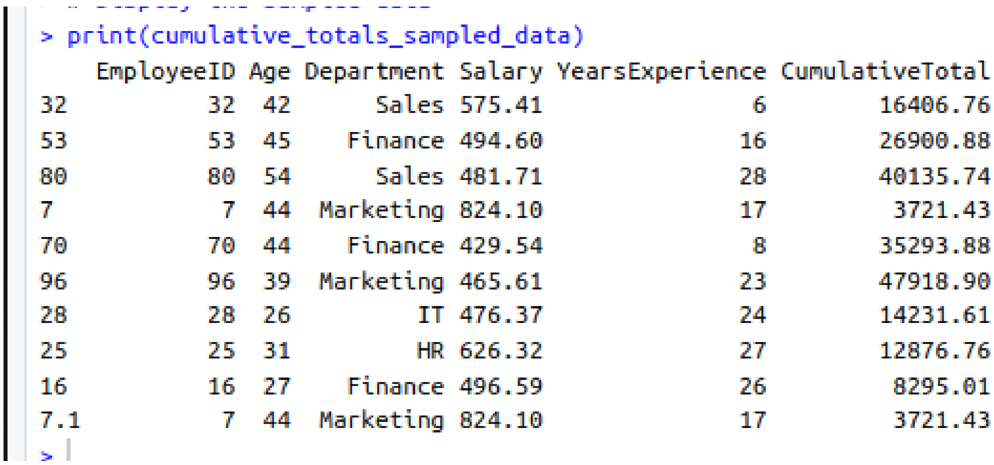


```
> print(lahiri_sample)
 [1] 66 64 40 58 78 25 59 57 23 90
>
> # Extract the sampled data from the dataset
> lahiri_sampled_data <- employee_data[employee_data$EmployeeID %in% lahiri_sample, ]
> print(lahiri_sampled_data)
   EmployeeID Age Department Salary YearsExperience
23         23  36         HR 454.16              27
25         25  31         HR 626.32              27
40         40  48  Marketing 560.07               4
57         57  25  Marketing 448.39              10
58         58  27         HR 400.75              34
59         59  41    Finance 667.57              38
64         64  53         IT 442.60               2
66         66  32         HR 610.98              11
78         78  50         HR 442.56              32
90         90  27         HR 489.90              36
```

We also use the **Cumulative Totals Method** for sampling. This approach begins by calculating the cumulative total of the Salary column in the `employee_data` dataset. Next, random numbers are generated and scaled to fall between 0 and the total sum of all salaries. For each random number, the method selects the first employee whose cumulative salary total is greater than or equal to that value. The indices of these selected employees are then used to retrieve the corresponding rows from the dataset, creating the sampled data.

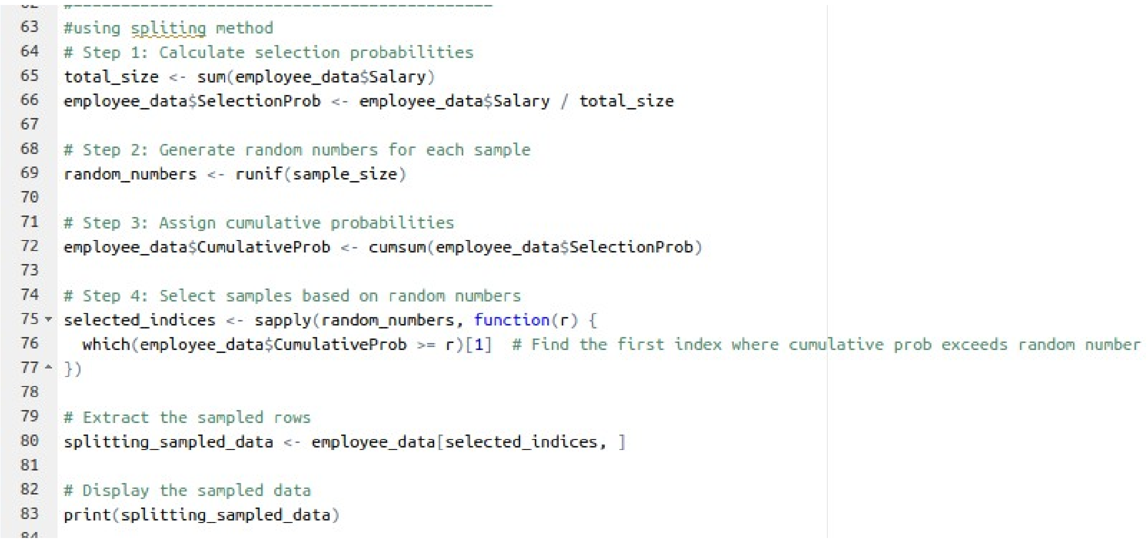


```
#using cumulative totals method
# Step 1: Compute cumulative totals of the size variable
employee_data$CumulativeTotal <- cumsum(employee_data$Salary)

# Step 2: Generate n random numbers scaled to the total size
total_size <- sum(employee_data$Salary)
random_numbers <- runif(sample_size, min = 0, max = total_size)

# Step 3: Select samples based on the random numbers
selected_indices <- sapply(random_numbers, function(r) {
  which(employee_data$CumulativeTotal >= r)[1]  # Find the first index where cumulative total exceeds r
})

# Step 4: Extract the sampled rows
cumulative_totals_sampled_data <- employee_data[selected_indices, ]

# Display the sampled data
print(cumulative_totals_sampled_data)
#============================================
```

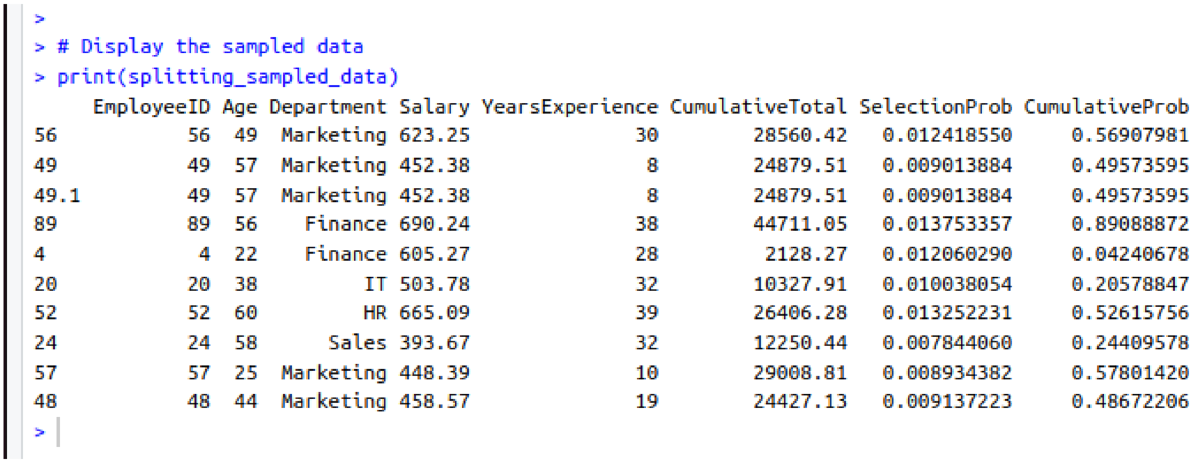


```
> print(cumulative_totals_sampled_data)
    EmployeeID Age Department Salary YearsExperience CumulativeTotal
32          32  42      Sales 575.41               6        16406.76
53          53  45    Finance 494.60              16        26900.88
80          80  54      Sales 481.71              28        40135.74
7            7  44  Marketing 824.10              17         3721.43
70          70  44    Finance 429.54               8        35293.88
96          96  39  Marketing 465.61              23        47918.90
28          28  26         IT 476.37              24        14231.61
25          25  31         HR 626.32              27        12876.76
16          16  27    Finance 496.59              26         8295.01
7.1          7  44  Marketing 824.10              17         3721.43
> 
```

The **Splitting Method** starts by calculating each employee's selection probability, which is determined by their Salary as a proportion of the total salary (SelectionProb). Random numbers are then generated to act as sample thresholds. Next, cumulative probabilities (CumulativeProb) are computed for all employees. For each random number, the method selects the first employee whose cumulative probability exceeds it. The rows corresponding to the selected employees are extracted from the dataset, stored in splitting_sampled_data, and printed. This method ensures that employees with higher salaries have a greater chance of being included in the sample.



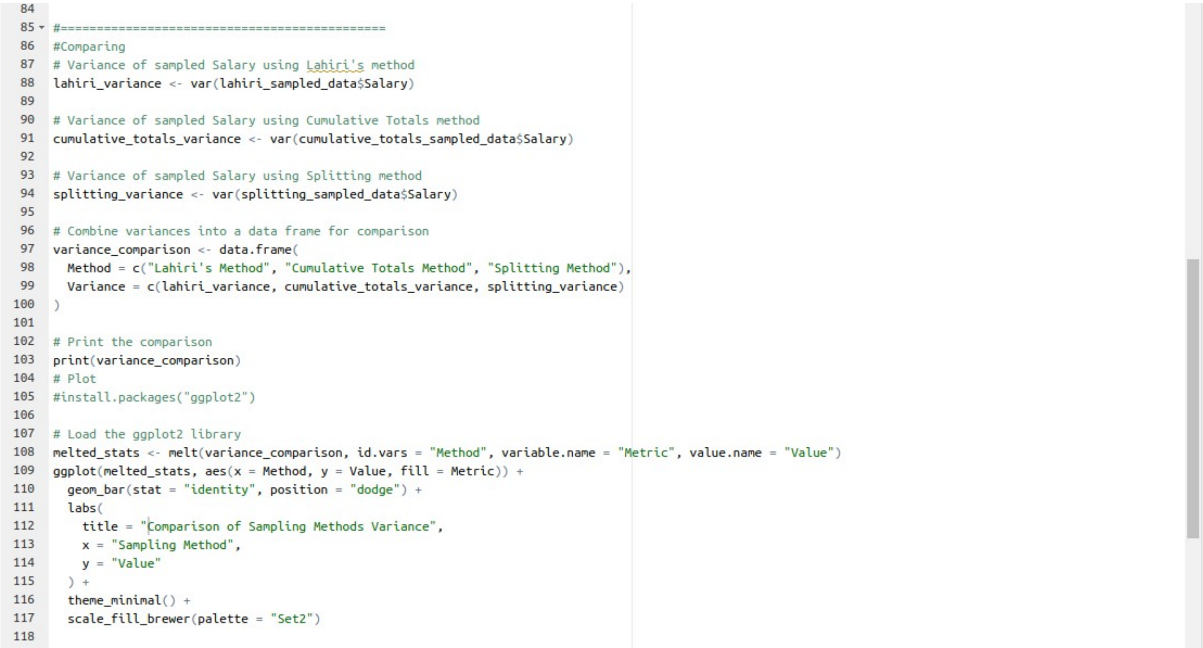
```r
#---------------------------------------------
#using spliting method
# Step 1: Calculate selection probabilities
total_size <- sum(employee_data$Salary)
employee_data$SelectionProb <- employee_data$Salary / total_size

# Step 2: Generate random numbers for each sample
random_numbers <- runif(sample_size)

# Step 3: Assign cumulative probabilities
employee_data$CumulativeProb <- cumsum(employee_data$SelectionProb)

# Step 4: Select samples based on random numbers
selected_indices <- sapply(random_numbers, function(r) {
  which(employee_data$CumulativeProb >= r)[1]  # Find the first index where cumulative prob exceeds random number
})

# Extract the sampled rows
splitting_sampled_data <- employee_data[selected_indices, ]

# Display the sampled data
print(splitting_sampled_data)
```



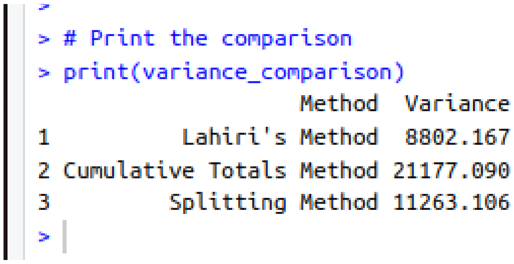
```
>
> # Display the sampled data
> print(splitting_sampled_data)
     EmployeeID Age Department Salary YearsExperience CumulativeTotal SelectionProb CumulativeProb
56           56  49  Marketing 623.25              30        28560.42   0.012418550     0.56907981
49           49  57  Marketing 452.38               8        24879.51   0.009013884     0.49573595
49.1         49  57  Marketing 452.38               8        24879.51   0.009013884     0.49573595
89           89  56    Finance 690.24              38        44711.05   0.013753357     0.89088872
4             4  22    Finance 605.27              28         2128.27   0.012060290     0.04240678
20           20  38         IT 503.78              32        10327.91   0.010038054     0.20578847
52           52  60         HR 665.09              39        26406.28   0.013252231     0.52615756
24           24  58      Sales 393.67              32        12250.44   0.007844060     0.24409578
57           57  25  Marketing 448.39              10        29008.81   0.008934382     0.57801420
48           48  44  Marketing 458.57              19        24427.13   0.009137223     0.48672206
> 
```

Chapter Four
Results & Discussion

This chapter presents the study's findings, focusing on a comparative analysis of three sampling methods: **Lahiri's Method**, the **Cumulative Totals Method**, and the **Splitting Method**, applied to a simulated dataset. The results are evaluated based on sampling efficiency, measured through the variance of the sampled data and the distribution of salaries. To illustrate the performance of each method, visualizations like boxplots are used, highlighting their strengths and weaknesses. The discussion interprets these findings, assessing how well each method produces representative samples and exploring their practical applications in scenarios requiring probability proportional to size sampling.



```

#=============================================
#Comparing
# Variance of sampled Salary using Lahiri's method
lahiri_variance <- var(lahiri_sampled_data$Salary)

# Variance of sampled Salary using Cumulative Totals method
cumulative_totals_variance <- var(cumulative_totals_sampled_data$Salary)

# Variance of sampled Salary using Splitting method
splitting_variance <- var(splitting_sampled_data$Salary)

# Combine variances into a data frame for comparison
variance_comparison <- data.frame(
  Method = c("Lahiri's Method", "Cumulative Totals Method", "Splitting Method"),
  Variance = c(lahiri_variance, cumulative_totals_variance, splitting_variance)
)

# Print the comparison
print(variance_comparison)
# Plot
#install.packages("ggplot2")

# Load the ggplot2 library
melted_stats <- melt(variance_comparison, id.vars = "Method", variable.name = "Metric", value.name = "Value")
ggplot(melted_stats, aes(x = Method, y = Value, fill = Metric)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Comparison of Sampling Methods Variance",
    x = "Sampling Method",
    y = "Value"
  ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2")

```



```
>
> # Print the comparison
> print(variance_comparison)
                    Method  Variance
1          Lahiri's Method  8802.167
2 Cumulative Totals Method 21177.090
3         Splitting Method 11263.106
> 
```

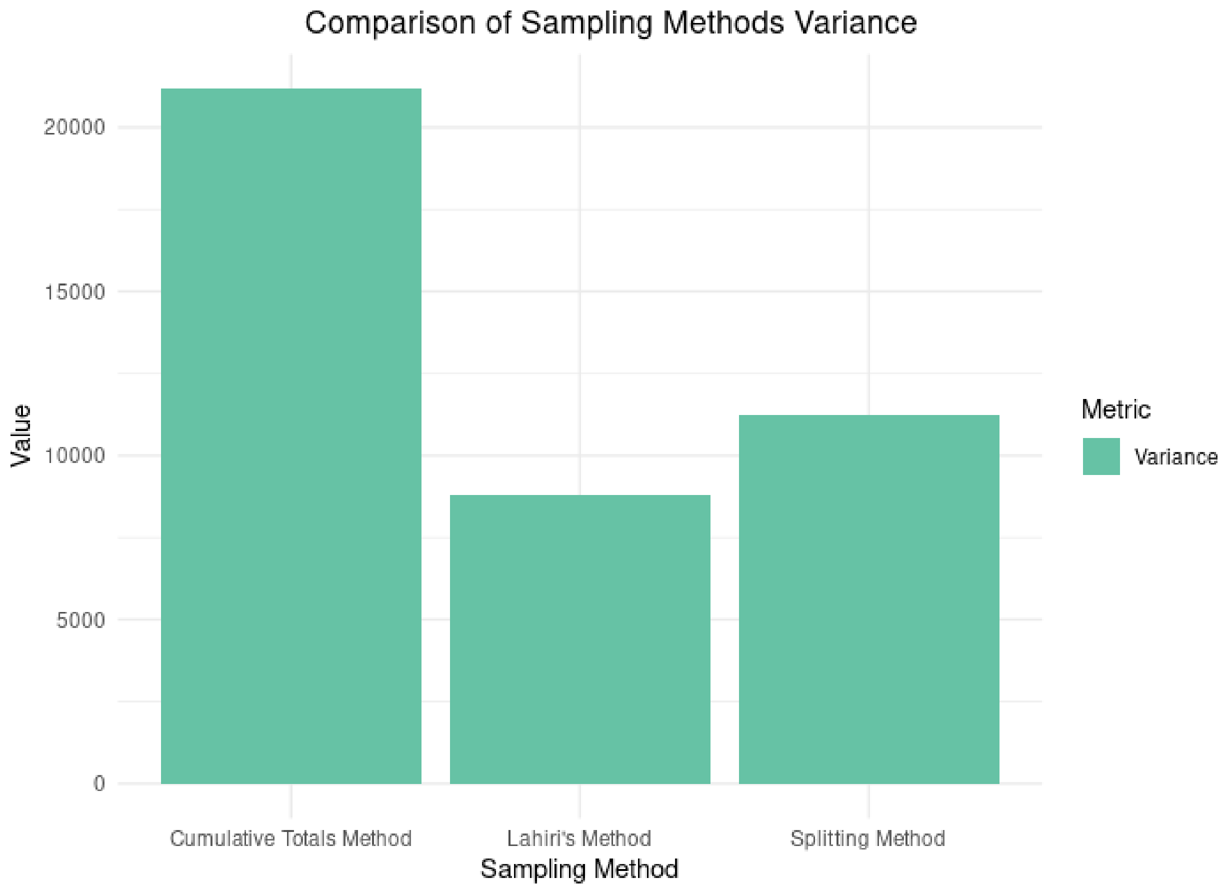


Figure 4.1 : Comparison of sampling method Variance

The output compares the variance of sampled salaries across three sampling methods, Among them, Lahiri's Method stands out with the lowest variance 8802.167, making it the most efficient and reliable for generating representative samples. The Splitting Method comes next with a moderate variance 11263.106, offering a balance between efficiency and simplicity, making it a practical alternative. On the other hand, the Cumulative Totals Method has the highest variance 21177.090, suggesting it is the least efficient approach, with the widest spread in the sampled data. Overall, Lahiri's Method proves to be the most effective, while the Cumulative Totals Method is the least favorable for scenarios requiring consistency and minimal variability.

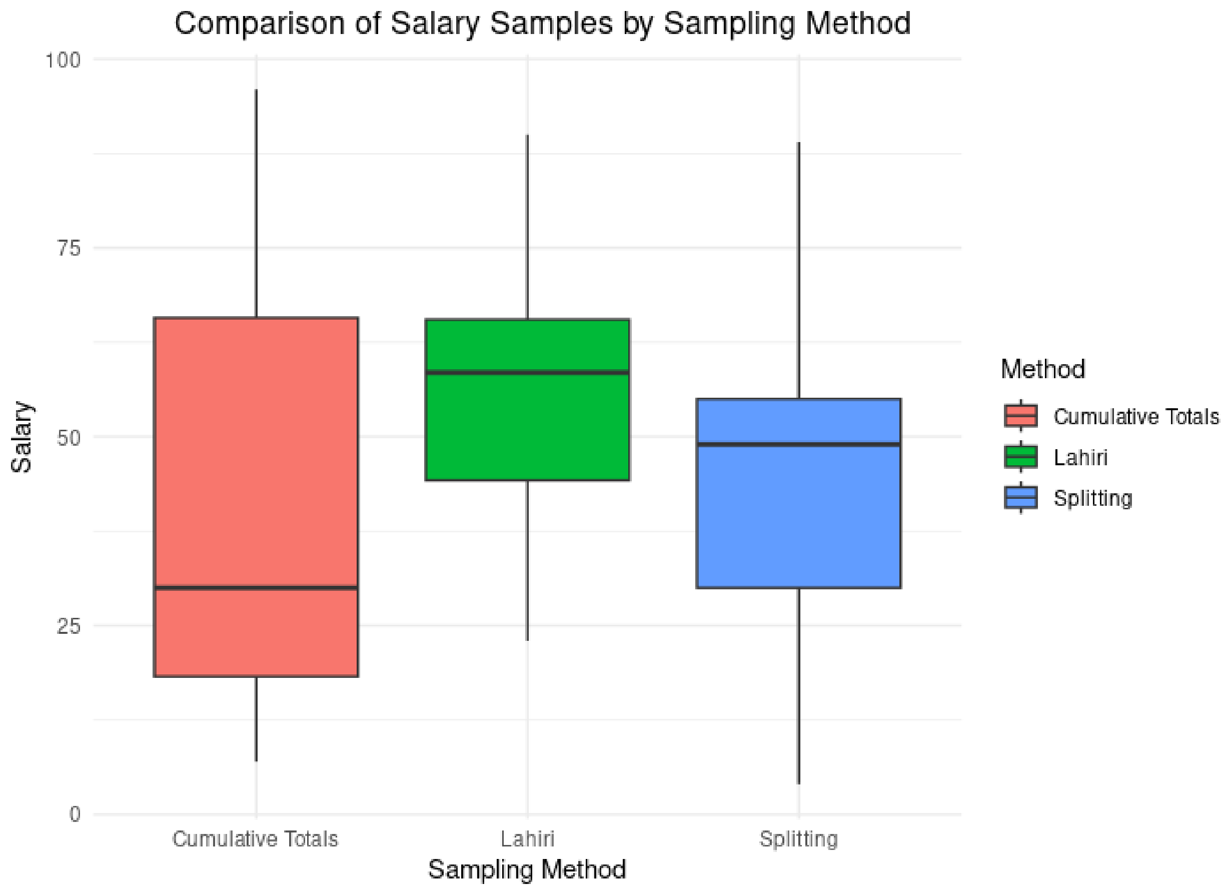


Figure 4.2 : Comparison of salary by sampling method

We use software packages called **ggplot2 ,** the boxplot compares salary distributions for our three sampling methods, providing insight into their efficiency and variability. The Cumulative Totals Method shows the greatest variability, with a wide range of salaries and a lower median, making it the least efficient option. In contrast, Lahiri's Method stands out with the lowest variability and a stable median, reflecting its consistency and efficiency. The Splitting Method falls in between, showing moderate variability—its spread is narrower than that of the Cumulative Totals Method but slightly wider than Lahiri's offering a reasonable balance between simplicity and effectiveness.

In summary, Lahiri's Method proves to be the most reliable, delivering consistent and representative samples, while the Cumulative Totals Method is the least desirable due to its high variability.

Chapter Five
Conclusion & Recommendations

In conclusion, the comparison of the Splitting Method, the Cumulative Totals Method, and Lahiri's Method reveals significant variations in their effectiveness and reliability. With the lowest variation and the most constant, representative samples, Lahiri's Method is clearly the most successful. For situations when simplicity is important, the Splitting Method provides a workable substitute with moderate variability. On the other hand, the Cumulative Totals Method has the greatest degree of unpredictability, rendering it the least effective and less appropriate for applications that demand accurate sampling.

These findings support the recommendation of Lahiri's Method for activities requiring accurate and trustworthy sampling, especially when reducing variability is crucial. When mild variability is acceptable and computing economy and simplicity are more critical, the splitting method may be a viable choice.

Based on these results, Lahiri's Method is recommended for tasks that demand precise and reliable sampling, particularly when minimizing variability is critical. The Splitting Method can serve as a good option when computational simplicity and efficiency are more important, and moderate variability is acceptable. However, caution is advised when using the Cumulative Totals Method, as its high variability could compromise the representativeness of the sample.

Scenarios Where Each Sampling Method Performs Best or Struggles

Table 5.1: Scenarios Sampling Method Performs Best or Struggles

Method	**Performs Best**	**Struggles**
Lahiri's Method	- High precision requirements - Small sample sizes - Complex populations	- Computationally intensive - Not ideal for large datasets
Cumulative Totals Method	- Quick sampling - Large datasets - Scenarios prioritizing execution time	- High variability - Samples may not be fully representative
Splitting Method	- Balanced efficiency and precision - Moderate dataset sizes - Iterative adjustments	- Moderate bias and variance - More computationally demanding than Cumulative Totals

References

- Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (n.d.). Sampling: Why and How of it. *INDIAN JOURNAL OF MEDICAL SPECIALITIES*, (2013).
- Arnab, R. (2017). *Survey Sampling Theory and Applications*. Academic Press.
- Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. Bethlehem, J. (2009). Applied survey methods: A statistical perspective. John Wiley & Sons.
- B., King, T., B., & Valente. (2013). *The modern census: evolution, examples and evaluation. International Statistical Review*.
- Blair, J., Czaja, R. F., & Blair, E. A. (2013). *Designing Surveys: A Guide to Decisions and Procedures* (3th ed.). sage publications.
- Borkowski, J. (n.d.). *Horvitz-Thompson Estimators*. https://math.montana.edu/jobo/st446/documents/ho6.pdf
- Boyd, P. (n.d.). Sampling Concepts. *MBA Faculty Conference Papers & Journal Articles*, (2002).
- Cochran, W. G. (1977). *Sampling Techniques*. Wiley.
- Etikan, I. R. R. (2021). *Comparison of quota sampling and stratComparison of quota sampling and stratified random samplingified random sampling*. Biom Biostat Int J.
- Fink, A. (2003). *The survey handbook. sage*.
- Fowler, F. J. (2014). *Survey Research Methods*. SAGE Publications.

- Golzar, J. (2022, December 16). *Simple Random Sampling Introduction*. International Journal of Education & Language Studies. Retrieved January 2, 2025, from https://www.ijels.net/article_162982_c72b367615dfd1f4d7bd9d4ff60cbef6.pdf
- Grafström, A. (2010). *On unequal probability sampling designs* [Doctoral dissertation]. Umeå University, Department of Mathematics and Mathematical Statistics.
- Hartley, H. O., & Rao, J. N. K. (n.d.). Sampling with Unequal Probabilities and without Replacement. *The Annals of Mathematical Statistics,*, ((Jun., 1962)), 350-374. https://www.jstor.org/stable/2237517
- Henderson, T. (2006, October). *Estimating the Variance of the Horvitz-Thompson Estimator* [A thesis submitted in partial fulfillment of the requirements for the degree requirements of Bachelor of Commerce].
- Horvitz, & Thompson. (1952). *A Generalization of Sampling without Replacement from a Finite Universe*. Journal of the American Statistical Association.
- Jayaraman, K. (n.d.). *A Statistical Manual For Forestry Research*. A Statistical Manual For Forestry Research. Retrieved January 12, 2025, from https://www.fao.org/4/x6831e/x6831e12.htm
- Jean-Claude, J.-C., & Tillé, Y. (n.d.). Unequal probability sampling without replacement through a splitting method. *Biometrika*, *85*(1), 89–101. https://www.jstor.org/stable/2337311
- Kish, L. (1965). *Survey Sampling*. Wiley.

- Lahiri, D. B. (1951). *Lahiri's method for PPS sampling*. https://www.scribd.com/document/455368650/Lahiris-method
- Lohr, S. L. (2021). *Sampling: Design and Analysis*. CRC Press.
- *Microsoft Corporation. (n.d.). RAND Function - Excel Help.* (n.d.). Microsoft Support. Retrieved January 1, 2025, from https://support.microsoft.com/excel
- Nikolopoulou, K. (2022, July 20). *What Is Non-Probability Sampling?* | *Types & Examples*. Scribbr. Retrieved December 31, 2024, from https://www.scribbr.com/methodology/non-probability-sampling/?utm_source=chatgpt.com
- Nikolopoulou, K. (2022, August 9). *What Is Convenience Sampling?* | *Definition & Examples*. Scribbr. Retrieved December 31, 2024, from https://www.scribbr.com/methodology/convenience-sampling/?utm_source=chatgpt.com
- Osier, G. (2025, 1 5). *Survey data in Economics and Finance.* https://bookdown.org/content/78e0de27-f084-46fb-be7c-c77a199d2abf/unequal-probability-sampling.html#the-hansen-hurwitz-estimator
- Pedigo, L. P., & Buntin, G. D. (Eds.). (1994). *Handbook of sampling methods for arthropods in agriculture*. Boca Raton, FL: CRC press.
- Professor Shalabh. (n.d.). *Varying Probability Sampling*. Chapter 7 Varying Probability Sampling. Retrieved January 12, 2025, from http://home.iitk.ac.in/~shalab/sampling/chapter7-sampling-varying-probability-sampling.pdf

- Raghunath, A. (2017). *Survey Sampling Theory and Applications*. Academic Press.
- *random — Generate pseudo-random numbers — Python 3.13.1 documentation*. (n.d.). Python Docs. Retrieved January 1, 2025, from https://docs.python.org/3/library/random.html
- Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau. (2011). *Survey Methodology* (second ed.). John Wiley & Sons.
- Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer New York.
- Scheuren, F. (n.d.). *What is a Survey*.
- Sharma, G. (2017). Pros and cons of different sampling techniques. *International journal of applied research*.
- Simkus, J. (2023, July 31). *Cluster Sampling: Definition, Method and Examples*. Simply Psychology. Retrieved January 3, 2025, from https://www.simplypsychology.org/cluster-sampling.html#Applications
- Singh, R., & Singh Mangat, N. (2010). *Elements of Survey Sampling*. Kluwer Academic Publishers.
- Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael 'selected' Amy*. Kluwer Academic Publishers.
- Singh, S. K., & Kumar, P. (n.d.). *Sampling Methods*. Indira Gandhi National Open University (IGNOU). https://egyankosh.ac.in/bitstream/123456789/97923/3/Unit-4.pdf

- *Systematic Sampling*. (n.d.). Chapter 11 Systematic Sampling. Retrieved January 3, 2025, from https://home.iitk.ac.in/~shalab/sampling/chapter11-sampling-systematic-sampling.pdf
- Tillé, Y. (2006). The Splitting Method. In *Sampling Algorithms* (pp. 99-121). Springer. https://link.springer.com/content/pdf/10.1007/0-387-34240-0_6.pdf
- United Nations Statistics Division. (2008). Principles and recommendations for population and housing censuses. *United Nations Publications (Revision 2)*.