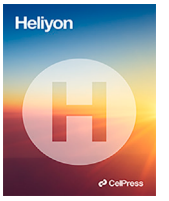
Heliyon 9 (2023) e21407

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Constructing efficient strata boundaries in stratified sampling using survey cost

Karuna G. Reddy *, M.G.M. Khan

School of Information Technology, Engineering, Mathematics and Physics, The University of the South Pacific, Suva, Fiji

A R T I C L E I N F O

Dataset link: https://ada.edu.au

MSC:
62-11
62D05
62P25
91B82

Keywords:
Optimum stratification
Stratified random sampling
Survey cost
Sample allocation
Dynamic programming

A B S T R A C T

For maximum precision in population parameter estimation under the Stratified sampling design, the optimum strata boundaries (OSB) could be constructed based on a continuous study variable rather than a set of categorical variables. If constructed optimally, the OSB results in homogenous units within each stratum leading to optimal stratum sample sizes (OSS) as well. The OSB and OSS may not remain optimum if the problem is considered in terms of a fixed total sample size, especially when a survey design involves a fixed budget. This article suggests a methodology for computing the OSB and OSS when the per unit stratum measurement costs for the survey or its probability density function are known. To plan for such a stratified survey, we demonstrate a design-based stratification empirically by using Wave 18 of the HILDA Survey general release dataset where we estimate the mean level of Gamma-distributed annual total disposable income in Australia, which could potentially be an important variable for policy decision-making. We also provide numerical illustrations for hypothetical study variables that follow exponential and right-triangular distributions respectively. The findings indicate that the suggested method is satisfactory in the sense that it is either more efficient or relatively comparable with other methods aimed at improving the accuracy of population parameter estimates. The proposed technique has been implemented in the updated `stratifyR` package.

1. Introduction

One of the popular methods of survey design used today is stratified random sampling. In this design, when dividing a population into strata, it is important that the samples in each stratum are as homogeneous as possible - this results in the best estimates of key population characteristics such as means and totals. Over the years, stratified sampling has evolved dramatically with every effort now aimed at improving the estimates to be as close as possible to the true population parameters. Many a time, to aid in decision-making, National Statistical Offices (NSO), government departments, and private organisations call for such surveys to be conducted in a short timeframe, with minimal effort and a fixed cost without compromising the quality and precision of the estimates. Without a doubt, these surveys have to be planned and administered quickly and carefully.

The Stratified sampling technique has many advantages over other sampling methods, particularly when dealing with diverse populations. Some of the key advantages of stratified sampling include improved representativeness that helps minimize parameter estimation; more precise and efficient estimation than simple random sampling; greater reliability and generalizability of findings; increased statistical power in hypothesis testing and inferential statistics; guarantees a sufficient representation of rare subgroups,

* Corresponding author.
E-mail addresses: karuna.reddy@usp.ac.fj, karunaredz@gmail.com (K.G. Reddy).

https://doi.org/10.1016/j.heliyon.2023.e21407
Received 30 March 2023; Received in revised form 3 October 2023; Accepted 20 October 2023

Available online 24 October 2023

leading to more robust analysis; is flexible and can be employed with different sampling techniques, which allows researchers to tailor the sampling method to the specific study needs; it enables researchers to conduct meaningful comparative analysis between subgroups; can lead to reduced variability in estimates and improve the precision of the statistical analysis. With these advantages, however, stratified sampling does require a more extensive initial effort to identify and define strata accurately, especially in diverse populations [1–4].

While planning, essentially, there are two major design issues in stratified sampling and these are defining the optimum strata boundaries (OSB) and choosing the optimum sample sizes (OSS), where the observations are allocated to the defined strata. The optimal allocations are usually disproportionate where some strata have higher sampling fractions than others, which means it is desirable to have a larger sample size for strata with minority populations that have larger variances. If the sample units are assigned correctly to strata, it usually results in less cost and more precision. Fundamentally, the following two concepts are targeted in our estimation efforts: (i) minimise the costs for the survey implementation for a specified precision; (ii) maximize the precision of estimates under the constraints of a fixed cost. The latter will be investigated in this paper where the stratification boundaries will be determined for the study variable.

The precision of survey estimates and the expense of survey execution are typically trade-offs made while developing a survey. One of the major constraints to the formulation of important decision-making is the absence or inadequacy of information on expenses related to various components of survey implementation. As a result, this research considers the optimisation of the variance function subject to very basic cost restrictions, providing a formal mathematical development of the survey precision and cost. There are usually limitations in the availability of information with regards to cost and variance, hence, this approach is aimed at providing rough approximations towards the stratified design based on cost [1–3]. The information of the study variable could be obtained from a most recent survey as it would best represent the study variable when the survey being planned. Under the constraint of total survey cost, the technique employs the concept of minimising the variance of the population mean. Inherent to the problem of optimum stratification is the problem of optimum sample size allocation because the determination of OSS is based on the strata that have already been constructed.

In this paper, the key idea in obtaining maximum precision in estimates is that the OSB is created by partitioning the range of the main population characteristics under study at suitable points ensuring that the total stratum variance is minimised for a given sample size. The concept of the determination of OSB has a very rich literature and initially investigated by Dalenius where his primary stratification variable was the study variable [5,6]. Many other researchers [7–11] developed various approximation techniques for calculating the OSB, all of who suggested some form of approximate solutions to the problem introduced by Dalenius [5,6].

Then, it was suggested by [12] to construct the OSB by dividing the square root of cumulative frequency at equal intervals. Although the cumulative root frequency method is quite a popular method being used today, [13] claimed that it has some arbitrariness which makes it difficult to implement. Later, [14] worked towards developing a method known as the Geometric method, which worked well for skewed populations but fell short in many other datasets [15–17].

In order to determine the OSB, many scholars have suggested algorithms, and one such approach calls for formulating the OSB as an optimisation problem and solving it with the dynamic programming (DP) technique [18]. This approach was examined by [19] and suggestions for further improvements to the approaches were made using various populations. This approach was also used by [7] to obtain the OSBs, where the population domains of the two stratification variables were divided into distinct subsets so that the variable of interest's precision was maximised. A technique was then suggested by [20] for obtaining the precise value for the OSB when the frequency distribution of the study variable is known and the number of strata is predetermined. They formulated the problem of determining OSB under Neyman allocation as a Mathematical Programming Problem (MPP) and solving it using the DP approach. However, they did not consider the problem when the cost of the survey is also a constraint, which leads to the complex process of survey costing. We explain below some of the issues that may exist in practical situations, which led to the proposed method in this study.

Surveys come in all shapes and sizes, with some certainly having high price tags and for some, we are not even sure how much it would actually cost in order to obtain appropriate sample sizes for accurate estimates of the population parameters. Since nothing is known about the survey costs (except that the survey designer is given a total survey budget or the available measurement efforts), it is imperative to develop a specific model which is suitable for the survey being undertaken. Several cost models could potentially exist in the discovery of the best sample designs for a particular survey, where every survey may be represented by a unique cost model. The conventional costing formula is that: Total cost = Fixed costs + Variable Strata Costs, where the fixed costs are expenses that must be paid no matter what sample size is selected (such as the price of developing, testing, and programming the questionnaire) and the variable costs are the per-unit expenditures associated with reaching out to the sample units, interviewing them, contacting the non-respondents, etc. Depending on how many sample instances are fielded, these variable costs change.

Surveys also use different modes of data collection, which may have different costs from stratum to stratum depending on the mode used. For example, the interviewer's travel costs in a survey could be much higher for a face-to-face interview compared to an interview via telephone. Similar to this, the quantity of callbacks or follow-up attempts may have an impact on how much less expensive mail or telephone modes are. There may also be a cost differential between mail and online surveys with the latter being proportionally more expensive in terms of fixed costs and the former being more expensive in terms of variable costs. Expressing it as a cost ratio, face-to-face surveys tend to cost twice as much per unit as telephone surveys [21–23]. A key aspect of the typically higher (compared to telephone) cost of face-to-face mode of interviewing is the amount of travelling time taken by an interviewer to reach the respondent, and in addition to that, each additional callback significantly raises the cost of the survey.

In a CAI (computer-assisted interviewing) method, the necessary equipment is provided to the interviewers in the field. The inclusion of this equipment cost could potentially increase the costs in face-to-face mode of interviews compared to telephone the mode. Conversely, the mode of the survey does not affect the fixed costs associated with developing the CAI instrument, so the stratum sample size is what may affect the overall cost differential. There is generally a smaller cost difference between telephone and mail surveys - literature has it that these ratios are between 1.2 [24,25] to around 1.7 [21,26,27]. Online surveys are substantially less expensive than mail surveys, however, the relative costs of the two approaches may vary depending on the fixed and variable expenses as well as the quantity of survey work required. Compared to mail surveys, online surveys usually have larger fixed costs which may include the general infrastructure cost and the questionnaire-development cost. For a fully electronic survey which is typically mounted online, and based on invitations and reminders (i.e., emails), the per-unit costs are almost negligible. Compared to online surveys, the fixed costs for mail surveys are typically lower, but the variable costs (printing, postage, etc.) are higher. Generally, the respective cost strategies for these two methods depend on the number of units that the fixed costs are divided over in total. Thus, it's important to establish a suitable cost model using ordinary functions of sample design features like the number of strata chosen and the number of interviewees, travelling distance to the sample from the administrative centres, mode of data collection, etc.

Survey researchers often concentrate on reducing survey errors, however, survey costs and errors inversely affect each other - raising one lowers the other. We have sampling and non-sampling errors that exist within a survey, which include coverage, non-response, sampling, and measurement errors. The majority of techniques used to lessen these errors have direct cost effects on the survey. [2]. Thus, the idea in this research is to combine the two, i.e., estimate the stratification boundaries by attempting to lessen the error of survey estimates by incorporating cost.

The majority of cost models now in use are linear functions of some survey parameter [28] (such as the number being sampled in each stratum), yet many non-linearities appear to exist in reality [2]. Also, the majority of cost models have those parameters as continuous, however, costs frequently experience discontinuities when certain adjustments come with administrative modifications in design or when different modes of data collection have varying costs depending on different locations being sampled [29]. This can often be explained by the fact that surveys frequently make use of substantial and costly discrete administrative divisions. For instance, if a survey's data gathering effort expands, more field supervisors (or administrative offices) need to be hired. The cost per interview in the sample is significantly impacted by these additions, which represent significant financial outlays. The majority of cost models are deterministic and presume applicability across every replication of the survey, however costs might vary significantly due to random events in sample selection or interviewer selection [2]. Furthermore, most cost models are perhaps only relevant to a specific type of surveys (such as those with a specific administrative structure or a specific number of interviewers), despite the fact that some modes of data collection make this evident. Therefore, by paying greater attention to cost model construction and parameter estimates, survey designs could be improved.

This research will consider the problem of determining OSB and OSS by taking into account the cost factor in the design of sample surveys, which has somewhat lacked in survey methodological research. Ideally, whilst planning for a survey, the total budget (which could also be termed as 'measurement efforts' where monetary values cannot be assigned) is usually fixed but the per-unit measurement costs (c_h) between strata vary because sampling units are located in different geographical areas or regions. As a result, the OSB computed using the fixed total sample size alone might no longer be the best option for a given survey cost. It is reasonable to say that while planning for a stratified sampling survey, it is important to consider the total budget that has been allocated to the survey, the predetermined number of strata, fixed total sample size, and also the average cost of per unit measurement within the strata. The proposed methodology addresses the gaps that exist in literature as stated in earlier paragraphs by first presenting the formulation of the problem as an MPP, implementing the solution process into practise using the DP technique, and also demonstrating the application with a numerical illustration using a real dataset.

In the ensuing sections, we provide a background to the context of the research, present the proposed methodology for the problem of stratification and sample allocation where survey costs are involved, discuss the solution procedure using dynamic programming technique, and apply the proposed method using the HILDA survey data. In the implementation, we present the formulation of the specific problem, estimate the average stratum costs, compute the OSB and OSS for the variable under study, and compare the results with other established methods. A discussion of the results is provided as we proceed through the paper and a conclusion is also presented at the end.

2. Background

The national statistics offices or government statistical agencies often conduct household surveys using an integrated system of household surveys to obtain critical information on the demographics, socio-economic status, and a variety of other key areas to help in planning, formulate policies and implement programs. These include household income and expenditure surveys, general social surveys, demographic and health surveys, surveys of education and work, and so on. In addition to these, market or other small-scale surveys are carried out on an ad-hoc basis by different organisations and additionally serve as a beneficial information source and contribute to national policy choices and development strategies.

Notably, modern-day surveys are based on stratified multistage cluster designs due to the fact that household frame listings or addresses are either unavailable or incomplete. This compels surveyors to choose a sample of geographic units first, from which, a household list is constructed. Households are then chosen from those lists. A multistage architecture is also used to reduce data collection costs. Stratified sampling, either as a single-stage or multi-stage design, has been a common phenomenon in a variety of different fields as it can be used to determine very accurate estimates (strata-based and overall) of certain population characteristics.

Literature has it that explicit stratification using categorical variables like geographic regions, age groups, ethnicities, etc. do not result in homogeneous strata and results in imprecise population estimates. The usage of continuous variables has proven to produce more accurate results.

As an example, and also used in the application of the proposed method, let us consider the main study variable, of a survey yet to be conducted, as the 'total household disposable income' (after taxes and transfers, henceforth referred to as 'income'). The study of current household income would most definitely lead to new insights into the development and evaluation of sound policies, programs, and services (such as in education and healthcare) to ensure they are delivering value to the people and communities who need them. For example, in Australia, if the state regulatory authorities want to enact policies and laws which directly depend on income (and other similar variables), it is important to first obtain a very recent and accurate estimate of such a population characteristic. Being a heterogeneous continuous variable, it might be useful to create optimal cut-off points or thresholds that might affect different communities or areas within the state in a particular way. The thresholds that define the homogeneous strata based on income could potentially be of importance, especially in the development of policies for those groups.

So the question is how do we carry out this explicit form of stratification? Stratification is most efficient when the strata mean varies considerably (i.e., small variability within each stratum) and while constructing strata, efforts are geared towards achieving as highly different strata means as possible. Stratification crucially relies on available frame data before sampling can be done, i.e., on the individual population units of the main study variable, y. For example, if we state that the survey is stratified by region and education then these characteristics are known for every unit on the frame before sampling. In Australia, researchers could potentially use the Census data or the MADIP (Multi-Agency Data Integration Project) dataset provided by the ABS as the sampling frame for planning of more general surveys. In most countries, while designing the survey, this information on the units of the population is usually not available or almost impossible to achieve in population surveys such as address-based sampling or RDD designs. A survey would not be required at all if it did.

For design-based stratification, a legitimate option would be to use data from recent surveys, which always add value to the whole planning and design process. Some researchers have studied population data (such as the MADIP dataset provided by the ABS on income) and obtained stratification-based estimates which improve the accuracy of population estimates. However, since population datasets are not easily available or get outdated with the census being conducted every decade, the method involves planning a stratified sampling design using data from similar surveys that have recently been conducted.

Thus, in the application of the proposed method, we consider the positive and negative total household disposable income variables (**tifditp** and **tifditn**) from Wave 18 of the HILDA survey general release data [30] where we created the income variable by subtracting them. Considering it as a readily-available super-population, we can plan for a stratified survey with a fixed budget allocated for the survey together with an estimate of the cost of obtaining information for every household. Due to this reason, one is always interested in being able to effectively sample by selecting just the right number of households so that maximum precision in the population estimates is achieved. For obvious reasons, organisations doing the data collection would keep these costs confidential. Hence, one might be inclined to imagine a distribution of cost for a particular survey, either roughly estimated or based on the costs of a previous survey. The socio-economic continuous variable of **income** is a particularly important one - it is the amount of money that a household earns each year after taxes and represents the money available to an individual in a household for spending on goods or services. Suppose our aim of the stratified survey is towards the development of policies pertaining to income for the different states within Australia.

3. Proposed method

For the planning of such surveys, prior data are normally utilized where the survey costs are mostly not available to us unless the surveys were conducted in-house. These costs are often estimated (via a traditional linear cost model) based on certain complex survey criteria, such as location, distance from the administrative centre, mode of data collection, and so on. Thus, for the wide applicability of the proposed method, stratification is conducted in two phases. Considering the complexities of the per-unit measurement cost of the survey, the first phase of stratification computes the initial OSB to estimate the average stratum costs. In the second phase, the final stratification of the main variable is constructed based on these average stratum costs.

Thus, we present an MPP-style broad formulation of the stratification problem subject to a cost constraint, wherein, we deal with the cost model and present the idea of solving a non-linear programming problem using the method of dynamic programming to obtain the optimum strata boundaries. We also utilize the idea of optimal allocation to calculate the stratum sample sizes.

3.1. The problem of stratification & sample allocation using survey cost

In stratified random sampling, if L strata of size $N_h; h = 1, 2, ..., L$ are formed from a population of size N, the population mean of the desired study variable can be estimated [1,31] as given in equation (1) below.

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h. \tag{1}$$

For an SRSWOR of n_h sampling points, the mean of the desired variable at the i^{th} point in h^{th} stratum (i.e., y_{hi}) can be computed using the desired variable's unweighted mean given by equation (2).

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} \tag{2}$$

In stratified sampling, equation (3) presents the unbiased estimator for the population mean ($\bar{Y}$).

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h, \tag{3}$$

with its variance given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 S_h^2, \tag{4}$$

where $W_h = \frac{N_h}{N}$ denotes the stratum weight and $S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2$ denotes the variance of the desired variable in h^{th} stratum and n_h is the stratum sample size. To calculate a sample variance estimate for the estimated population mean, the unknown stratum variances are normally replaced by the estimated sample variances.

The question is, how do we allocate a predefined total sample size (n) among L? Out of several allocation methods, the most popular of them are 'proportional' and 'optimum' allocations. The choice of the best allocation scheme would determine how precise is the estimate of the population mean. The best allocation scheme is affected by

1. how many units there are overall in each stratum (n_h),
2. the variability of the measurements within each stratum (S_h^2), and
3. the cost associated with obtaining an observation from each stratum (c_h).

In sampling literature, if there is no measurement cost involved, the proportional allocation $\left(n_h = \frac{nN_h}{N} = nW_h;\ h = 1, 2, ..., L\right)$ maintains a steady sampling fraction throughout the population. However, it is not an optimal choice because it does not consider the variability, S_h^2, within each stratum. On the other hand, when the per-unit measurement cost, c_h, varies in every strata, the optimum allocation, n_h, that minimizes the variance in (4) is proportional to $W_h S_h / \sqrt{c_h}$ (see [1]). However, in many practical situations, the sampling costs are variable between the units in the stratum (i.e., the per unit cost is different within and between strata), and the surveyors may find it difficult to determine optimum stratification and allocation to the stratum.

Let us consider a simple cost function for a given budget C of the form:

$$C = c_0 + \sum_{h=1}^{L} c_h n_h, \tag{5}$$

where c_0 denotes the overhead cost, which typically represents the expense of administration, performing interviewer training, etc., and c_h is the average per-unit cost in h^{th} stratum defined by

$$c_h = \sum_{i=1}^{n_h} c_{hi} \tag{6}$$

The term c_{hi} is the cost of measuring i^{th} unit in h^{th} stratum. Note that c_{hi} will differ from unit to unit, which can be estimated from the past experience of the surveyors considering geographic location or statistical area, method or mode of gathering data, the distance between the administrative center and the respondent, and numerous other attributes. Once these are known, c_{hi} can be estimated based on the initial boundary points obtained by using the `stratifyR` package [32] or any other package such as `stratification` [33].

The lack of knowledge on per-unit costs related to various survey implementation factors, especially when a survey is being designed for the first time, is a significant restriction in survey design. While some consistency in costs between surveys can be helpful in designing a new survey, this is only applicable to qualitative cues about the relative sizes of various cost components. Moreover, costing is normally a highly confidential matter for any survey company or NSO, hence, all parameters pertaining to the costs related to a survey are not available. In such cases, one can only assume an apriori distribution of the per-unit measurement costs to estimate c_h. One must obtain all the information regarding the factors surrounding the varying stratum costs, especially during the data-collection exercise, in order to have a realistic distribution of the per-unit cost of the survey in the underlying population.

Let $f(c)$ be a probability density function of the per-unit measurement costs, the average stratum costs, c_h, can be obtained within its boundary points by

$$c_h = \frac{1}{W_h}\int c\, f(c)\, dc, \tag{7}$$

where W_h is the stratum weight in h^{th} stratum.

Then, to determine the optimum allocation to strata (n_h), we minimize $V(\bar{y}_{st})$ given in (4) subject to the total survey cost in (5). The optimum allocation is determined by equation (8), which applies the Lagrange multiplier technique to solve the problem.

$$n_h = \frac{C - c_0}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \frac{W_h S_h}{\sqrt{c_h}} \tag{8}$$

To compensate for heterogeneity, the method of optimum allocation ensures a selection of a greater sample size within a stratum if it accounts for a large part of the population, has a large within-stratum variance compared to other strata, or if sampling is not costly in a particular stratum. As a special case of optimal allocation, we use Neyman allocation when variances in the strata are different and the costs in the strata are approximately equal. With Neyman allocation, $n_h \propto N_h S_h$.

Furthermore, substituting (8) in (4), the minimum variance, given by equation (9), with optimum allocation is given by

$$V_{opt}(\bar{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)^2}{C - c_0} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} \tag{9}$$

Also, if finite population correction is ignored, for fixed C and c_0, minimizing the RHS of equation (9) is equivalently minimizing

$$\sum_{h=1}^{L} W_h S_h \sqrt{c_h} \tag{10}$$

In surveys, there could potentially be multiple variables targeted for estimation. A good stratification regime developed for one target variable may not necessarily be good for another. Hence, in the ensuing method, we consider a main study variable assuming that the stratification is made based on a single variable (x), which has a continuous probability density function $f(x), a \le x \le b$. Considering that the population is being partitioned into L strata where $x_0 = a$ and $x_L = b$ are the initial and final values of the density function. Then, the problem of computing the OSB is to cut the entire distance, d, that is,

$$x_L - x_0 = d \tag{11}$$

at the intermediary points $x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L$ such that the variance in (10) is minimum. With a known probability density function $f(x)$ of the stratification variable x, the estimated values of W_h and S_h^2 in (10) are obtained by

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx \tag{12}$$

$$S_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x)dx - \mu_h^2 \tag{13}$$

$$\text{where} \quad \mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x)dx \tag{14}$$

and (x_{h-1}, x_h) are the boundary points of h^{th} stratum. So, equation (10) can be represented in terms of boundary points (x_{h-1}, x_h). If we let $f(x_{h-1}, x_h) = W_h S_h \sqrt{c_h}$, then the problem of calculating the OSB can be converted to finding $x_1, x_2, ..., x_L$ from the following:

$$\begin{aligned} &\text{Minimize } f(x_{h-1}, x_h) = W_h S_h \sqrt{c_h} \\ &\text{subject to } a = x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b. \end{aligned} \tag{15}$$

With $l_h = x_h - x_{h-1}$ denoting the width of the h^{th} stratum, the distance of the probability density function in equation (11) can be represented in terms of the stratum width, and this is given by equation (16).

$$\sum_{h=1}^{L} l_h = \sum_{h=1}^{L} (x_h - x_{h-1}) = x_L - x_0 = d \tag{16}$$

Equation (17) below is then used to define the h^{th} stratification point $x_h : h = 1, 2, ..., L-1$.

$$x_h = x_0 + l_1 + l_2 + ... + l_h = x_{h-1} + l_h \tag{17}$$

The problem of computing the OSB can be considered as the problem of determining the optimum stratum widths (OSW) l_1, l_2, ..., l_L, and can be written as the following mathematical programming problem by adding (16) as a new constraint:

$$\begin{aligned} &\text{Minimize } \sum_{h=1}^{L} f_h(l_h, x_{h-1}) \\ &\text{subject to } \sum_{h=1}^{L} l_h = d \\ &\quad \text{and} \quad l_h \ge 0; h = 1, 2, ..., L \end{aligned} \tag{18}$$

With the known initial value of x_0, the first term in the objective function of (18), $f_1(l_1, x_0)$, is a function of l_1 only. After l_1 is determined, the second term $f_2(l_2, x_1) = f_2(l_2, x_0 + l_1)$ becomes a function of l_2 only and this pattern continues.

Due to its unique characteristics, the MPP (18) may be viewed as a function of l_h alone and is expressed as follows:

$$\begin{aligned} &\text{Minimize} \sum_{h=1}^{L} f_h(l_h) \\ &\text{subject to} \sum_{h=1}^{L} l_h = d \\ &\quad \text{and} \quad l_h \geq 0; h = 1, 2, ..., L \end{aligned} \tag{19}$$

3.2. Dynamic programming technique solution procedure

The MPP (19) is a multi-stage decision problem where the objective function and the constraint are sums of separable functions of l_h. As a result, due to this separable attribute and the nature of the MPP, it may be solved using the dynamic programming technique [18]. In order to find the best solution to a multi-variable problem, dynamic programming divides the problem into stages, each of which contains a single variable sub-problem. A DP model is a recursive solution procedure, based on Bellman's principle of optimality [34]. The recursive equation connects the many stages of the problem in a way that ensures that the optimum and feasible solution for each step is also the best feasible solution for the entire problem [19].

Considering a sub-problem of (19) for first $k < L$ strata:

$$\begin{aligned} &\text{Minimize} \sum_{h=1}^{k} f_h(l_h) \\ &\text{subject to} \sum_{h=1}^{k} l_h = d_k \\ &\quad \text{and} \quad l_h \geq 0; h = 1, 2, ..., L \end{aligned} \tag{20}$$

where $d_k < d$ is the total width or distance to be divided into k strata. Note that $d_k = d$ for $k = L$ and $d_k = l_1 + l_2 + ... + l_k$ are the transformation functions which reduce to $d_1 = l_1 = d_2 - l_1$.

Let $f_k(d_k)$ be the minimum value of the (20) objective function of

$$f_k(d_k) = \min \left[\sum_{h=1}^{k} f_h(l_h) \middle| \sum_{h=1}^{k} l_h = d_k, \text{ and } l_h \geq 0; h = 1, 2, ..., k \right].$$

With the above equation for $f_k(d_k)$, it is similar to finding $f_L(d)$ in MPP (19) recursively by finding $f_k(d_k)$ for $k = 1, 2, ..., L$ and $0 \leq d_k \leq d$. Thus, we can write:

$$f_k(d_k) = \min \left[f_k(l_k) + \sum_{h=1}^{k-1} f_h(l_h) \middle| \sum_{h=1}^{k-1} l_h = d_k - l_k, \text{ and } l_h \geq 0; h = 1, 2, ..., k \right].$$

For a fixed value of l_k; $0 \leq l_k \leq d_k$,

$$\Phi_k(d_k) = f_k(l_k) + \min \left[\sum_{h=1}^{k-1} f_h(l_h) \middle| \sum_{h=1}^{k-1} l_h = d_k - l_k, \text{ and } l_h \geq 0; h = 1, 2, ...k-1 \right].$$

Using Bellman's principle of optimality [34], we write a forward recursive equation of the DP technique as:

$$f(k, d_k) = \min_{0 \leq l_k \leq d_k} \left[f_k(l_k) + f(k-1, d_k - l_k) \right], k \geq 2. \tag{21}$$

For the initial stage, or for $k = 1$:

$$f(1, d_1) = f_1(d_1) \implies l_1^* = d_1, \tag{22}$$

where the optimum width of the first stratum is given by $l_1^* = d_1$. The relations (21) and (22) are solved logically in a forward manner for $k = 1, 2, ..., L$ and $0 \leq d_k \leq d$. After $f(L, d)$ is obtained, the optimum width of L^{th} stratum, l_L^*, is also obtained. From $f(L-1, d-l_1^*)$, the optimum width of $(L-1)^{th}$ stratum, l_{L-1}^* is obtained and this process goes on recursively until l_1^* is obtained.

After c_h have been estimated and OSB (y_h, y_{h-1}) have been constructed, the optimum sample sizes (OSS) in each stratum, $n_h; h = 1, 2, ..., L$, that minimizes the estimate's variance can be computed with relative ease. When the stratum variances vary, optimum allocation is preferable [35] due to the likelihood of larger units being more variable than smaller units. Then, as discussed in Section 3.1, the sample sizes n_h are obtained for a predetermined total budget C and overhead cost c_0 for $h = 1, 2, ..., L$ using the optimum allocation method given in equation (8).

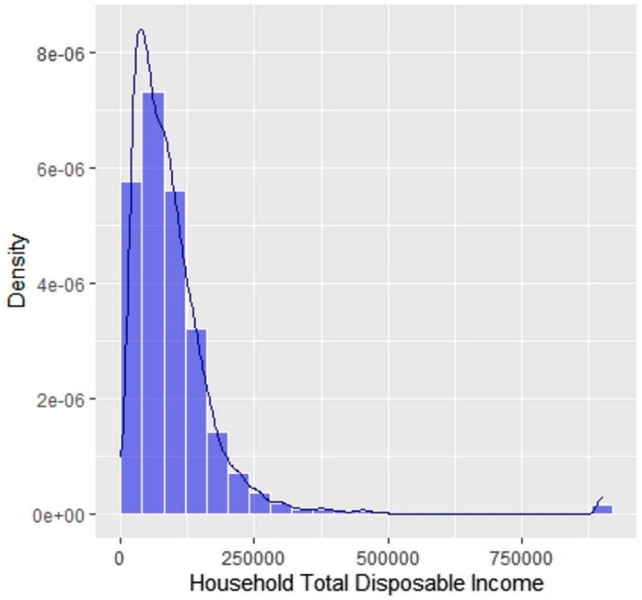


Fig. 1. Histogram with overlayed density curve of household total disposable income.

Do note that if the stratum costs are approximately equal for all study units, that is, $c_h = c$, the equation (8) can be reduced to what is termed as Neyman allocation, which is a special case of optimum allocation and is given by $n_h = n W_h S_h / \sum_{h=1}^{L} W_h S_h$, where $n = (C - c_0)/c$.

Using either of these allocation procedures, the OSB (y_h, y_{h-1}) are so obtained that n_h must fulfil the restrictions given in the inequality (23) below:

$$1 \leq n_h \leq N_h, \tag{23}$$

where $N_h = NW_h$. The restriction (23) specifies that the h^{th} stratum must avoid over-sampling and form with a minimum of one unit.

In the ensuing sections, we first present a numerical illustration of the application of the proposed method for an exponential distribution where the average per unit stratum costs are arbitrarily fixed. And then, secondly, we present an application of the method to real-world survey data where we investigate the study variable in terms of its distribution. Considering fixed average stratum costs estimated from an apriori distribution and the estimated distribution for the study variable, we formulate the problem of stratification into an MPP. A numerical illustration of the solution procedure is implemented and results are presented and deliberated upon.

4. Application using HILDA survey

We utilise the general release data from Wave 18 of the Household, Income and Labour Dynamics in Australia (HILDA) Survey to illustrate how the methodology is applied. [30,36], which is a household-based panel survey that gathers important data on economic and individual well-being, dynamics of the labour market, and family life. The HILDA Survey gathers data on many facets of Australian life, including household and family relationships, earnings and employment, as well as health and education, and offers policy-makers unique insights into Australia, empowering them to make informed decisions in a variety of policy areas, including social services, health, and education.

Having household unit-record data from such a survey is very critical in planning for future stratified surveys with budgetary constraints as one would aspire to stay within the budget without forgoing the precision of the estimates. Even though HILDA is a longitudinal study, data from the Wave 18 release of the HILDA survey is treated as a cross-sectional study that very well represents the Australian population. For the purposes of this application, assume that for a fixed budget, a stratified survey is being planned and the aim is to estimate a person's average income, which we refer to as the main study variable, y. Obtaining the characteristics of the population such as the stratum means of such a socio-economic variable would certainly assist policy-makers with sound decision-making.

The distribution, in the form of a frequency histogram, of income (or htdi - household total disposable income, henceforth referred to as income) is given in Fig. 1.

A desirably high level of estimation accuracy in any survey attempt calls for higher sample numbers, which may not be feasible. As a result, sample sizes are frequently determined by costs, and the level of accuracy may decrease slightly. A trade-off between cost and accuracy thus leads to this idea of having a clear objective of the design process, which is to minimize variance (or maximize

accuracy) at a fixed cost constraint. The costs may often differ among strata, so the most straightforward form of optimal allocation is to set the stratum sampling fraction (f_h) to be proportional to the stratum standard deviation (S_h), and inversely proportional to the square root of the stratum cost, that is, $f_h \propto S_h/\sqrt{c_h}$. This means that more heterogenous strata are sampled at a higher rate compared to a more homogenous one. And in cases where costs are the same between strata, the optimum allocation shortens to $f_h \propto S_h$, which is the Neyman allocation.

5. Results and discussion

Below we present the results and discussion in sections starting with the formulation of the stratification problem as an MPP and then on estimating the average stratum costs. Furthermore, the results and discussion are provided for computing the OSB and OSS for the Income variable from the HILDA survey. And finally, a comparison is made against other established methods in the literature.

5.1. Formulation of the problem as an MPP

Our main aim is to construct OSB on income for the whole population of Australia where income would be homogeneous within the constructed strata and could potentially span multiple states rather than just one. If one is interested in creating OSB for individual states then states could be taken as another stratification variable. This is particularly important in the planning of the stratified survey because the mode of data collection might be varied (possibly multi-modal to ensure a good response rate, improve coverage and shorten survey administration timeframes) and so would be the per-unit measurement costs in every state. This would also enable estimates for individual states and as well as to be combined for an overall estimate. The number of elements (OSS) ultimately selected from the finite target population depends on the boundaries that are created for the variable.

Ideally, to obtain the sampling frame (one that identifies the underlying survey population of Australia), one can obtain income or wealth data from a complete enumeration such as the Census or administrative data from the Australian Taxation Office (ATO) for accurate planning of a future stratified survey. In this application, we will use the HILDA survey data as a sampling frame that identifies the Australian population as it is a probability-based panel. The quality of stratum samples is only as good as the entire sampling frame from which it is drawn, however, for the purposes of planning a stratified survey, it is imperative to draw an optimal number of stratum samples based on the boundaries created as this will yield precise estimates for the population.

As depicted in Fig. 1, income (or htdi) is estimated to follow a Gamma (2P) distribution (i.e., $y \sim \Gamma(r,\theta)$) on the domain $[0,\infty)$, with the two-parameter probability density function given as:

$$f(y;r,\theta) = \frac{1}{\theta^r \Gamma(r)}\, y^{r-1} e^{-\frac{y}{\theta}}, \qquad y>0;\ \ r,\theta>0, \tag{24}$$

where r denotes the shape parameter and θ denotes the scale parameter with the Gamma function, $\Gamma(r)$, defined by equation (25).

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t}\, dt, \qquad r>0. \tag{25}$$

With a moderatley skewed profile, the Gamma distribution is frequently employed as a probability model for waiting times, such as the time until death and also in a variety of disciplines such as economics, modelling of rainfall, insurance claims, loan default amounts, wealth, and income. The formulation of the MPP is presented in Reddy et al. (2020) [32].

Computing the OSB with different strata costs as discussed later in Section 5.3, we illustrate the computational details of the DP solution procedure. In Fig. 1, the 2P Gamma distribution has the parameters: shape, $r = 1.7818$ and scale, $\theta = 53323.0953$ and the data spans on the domain $[x_0, x_L] = [\$58, \$899827]$, which implies that the range is $d = \$899769$.

With the Gamma density function presented in (24), using equations (12)-(14), the values for W_h, μ_h, and S_h^2 can be computed as a function of boundary points (y_{h-1}, y_h) where $y_h = y_{h-1} + l_h$. Equation (26) presents the stratum weight as:

$$W_h = \left[Q\left(r, \frac{y_{h-1}}{\theta}\right) - Q\left(r, \frac{y_{h-1}+l_h}{\theta}\right)\right]. \tag{26}$$

Similarly, equation (27) below presents the stratum mean, μ_h, as:

$$\mu_h = \frac{\theta\, r\left[Q\left(r+1, \frac{y_{h-1}}{\theta}\right) - Q\left(r+1, \frac{y_{h-1}+l_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{y_{h-1}}{\theta}\right) - Q\left(r, \frac{y_{h-1}+l_h}{\theta}\right)\right]}, \tag{27}$$

and equation (28) shows the stratum variance, S_h^2, as:

$$S_h^2 = \frac{\theta^2\, r(r+1)\left[Q\left(r+2, \frac{y_{h-1}}{\theta}\right) - Q\left(r+2, \frac{y_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{y_{h-1}}{\theta}\right) - Q\left(r, \frac{y_h}{\theta}\right)\right]} -$$

$$\theta^2 r^2 \frac{\left[Q\left(r+1, \frac{y_{h-1}}{\theta}\right) - Q\left(r+1, \frac{y_{h-1}+l_h}{\theta}\right)\right]^2}{\left[Q\left(r, \frac{y_{h-1}}{\theta}\right) - Q\left(r, \frac{y_{h-1}+l_h}{\theta}\right)\right]^2}. \tag{28}$$

Then, using the general forms in equations (15), (18) and (19), the formulated MPP could be derived as follows, where $d = y_L - y_0 = b - a$, θ and r are estimated parameters of the Gamma (2P) distribution, and c_h is the average per-unit stratum costs:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{h=1}^{L} \left\{ \begin{array}{l} SQRT\left\{ \theta^2 r(r+1)\left[Q\left(r, \frac{y_{h-1}}{\theta}\right) - Q\left(r, \frac{y_{h-1}+l_h}{\theta}\right)\right] \right. \\ \quad \times \left[Q\left(r+2, \frac{y_{h-1}}{\theta}\right) - Q\left(r+2, \frac{y_{h-1}+l_h}{\theta}\right)\right] \\ \quad \left. - \theta^2 r^2 \left[Q\left(r+1, \frac{y_{h-1}}{\theta}\right) - Q\left(r+1, \frac{y_{h-1}+l_h}{\theta}\right)\right]^2 \right\} \\ \quad \times c_h \end{array} \right\}, \\
\text{subject to} \quad & \sum_{h=1}^{L} l_h = d, \\
\text{and} \quad & l_h \geq 0; \;\; h = 1, 2, ..., L.
\end{aligned} \tag{29}$$

The function $Q(\cdot)$ is the Upper Regularized Incomplete Gamma function given by equation (30).

$$Q(r, y) = \frac{1}{\Gamma(r)} \int_{y}^{\infty} t^{r-1} e^{-t}\, dt, \qquad r, y > 0; \;\; \Gamma(r) \neq 0; \tag{30}$$

The process of obtaining the average stratum costs, c_h in equation (29), is presented in the next section.

5.2. Estimating the average stratum costs

Since costing is a highly confidential matter for this survey, all parameters pertaining to the costs related to such a survey are not available and can only be assumed apriori. For the purposes of applying the method to the unavailable cost information for the HILDA survey, we assume a prior distribution of cost to be lognormal in each of the 122 strata that were created, possibly based on the demographics of the survey and the geographic location (State and part of State), SA level (or statistical area), data collection mode, the distance between the administrative centre and the respondent, and numerous other stratification variables. The stochastic approach to fitting a random lognormal distribution takes into account our understanding—or lack thereof—of the relationship between the observed data and the unidentified parameters. Thus, we estimate 122 separate lognormal distributions with increasing *meanlog* and *sdlog* parameters, assuming that these increase as we go from Strata 1 to Strata 122.

The distribution of cost could obviously be a different one, such as normal, uniform, etc., depending on the type of survey and the evidence from past literature. The rationale behind the choice of the lognormal distribution is that we take evidence from the literature that in real-life, prices and costs often tend to follow such a distribution which suggests that healthcare costs are log-normal [37], also that exchange rates [38], price indices and stock market indices distributions typically exhibit such a distribution while Thompson (2005) [39] argues that the costs in a clinical trial best-fits a lognormal distribution. Overall, we could say that the distribution is discontinuous, whereby each stratum fits a different lognormal distribution based on some parameters that we supply for every stratum. It was noted by [29] that loss functions frequently have discontinuities. This can often be explained by the fact that surveys frequently make use of substantial and costly discrete administrative divisions. For instance, if a nationwide personal interview survey's data gathering effort expands, more field supervisors (or regional offices) are required. The cost per interview in the sample range in which the addition is made is significantly impacted by these additions because they incur significant financial outlays.

The two-parameter probability density function of the lognormal distribution of the per-unit measurement costs on the interval $(0, \infty)$ is given as

$$f(y; \mu, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{ln(y) - \mu}{\sigma}\right)^2 \right\}, \qquad y > 0 \tag{31}$$

where $\sigma > 0$ denotes scale parameter and μ denotes the location parameter.

To obtain the average stratum costs, c_h, the following steps could be undertaken:

1. Generate random variable (c_{hi}) with different estimated *meanlog* and *sdlog* parameters for each original strata specified in HILDA data.
2. Combine these into a stochastic variable to form the *cost* column.

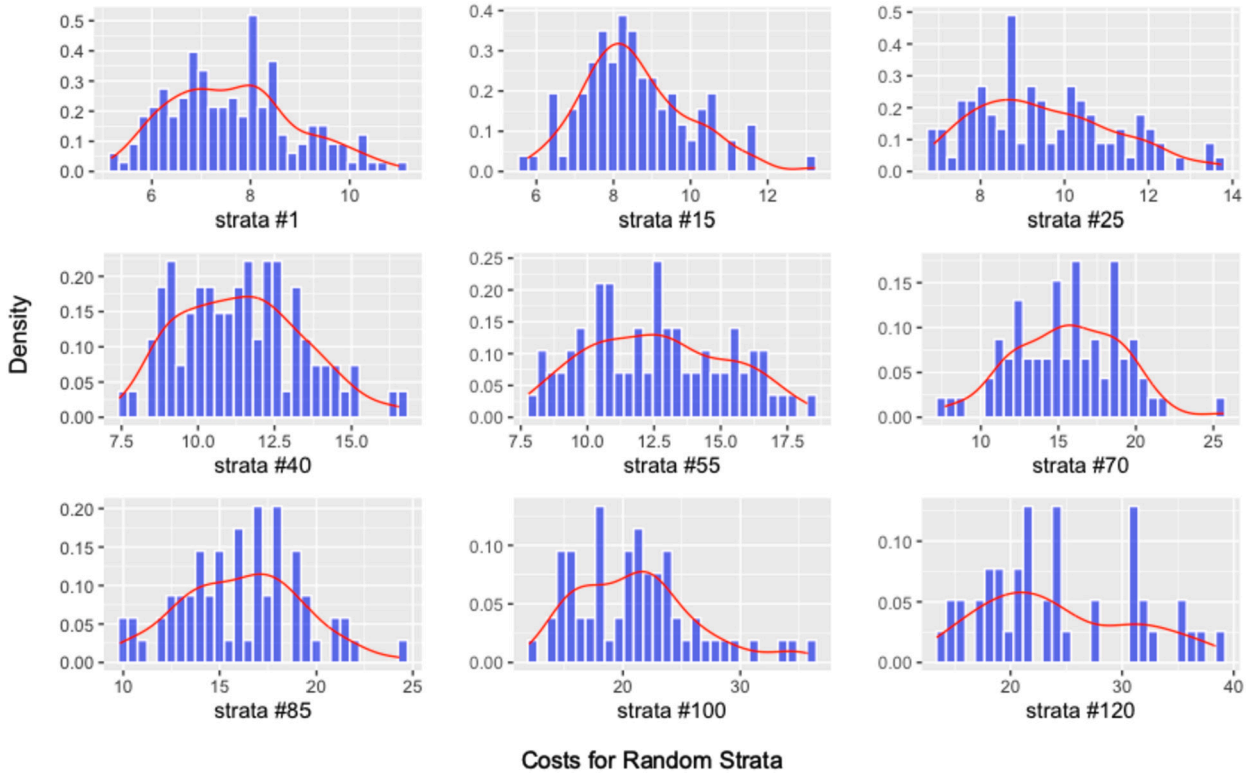


Fig. 2. Histograms with overlayed density curves for randomly-chosen strata.

3. Using *htdi* as the main variable, obtain the initial strata boundaries (OSB) using the DP method, which has been applied in the `stratifyR` package.
4. Using the OSB obtained in (iii) and equation (7), compute the average stratum costs. For e.g., if you have created 2 strata, your costs will be c_1 and c_2.
5. Repeat this process for $L = 2, 3, 4, 5$ and so on.

To present the above steps in a logical manner, we first assume separate lognormal distributions (with pdf given by equation (31)) of survey costs based on the original strata provided in the 2019 edition of the HILDA with the *meanlog* parameter ranging from 2 to 3.21 at increments of 0.01 and the *sdlog* parameter ranging from 0.15 to 0.27 at increments of 0.001. To visualise these hypothetical distributions, we present the histogram and density curves of only nine such random or stochastic cost data (for strata 1, 15, 25, 40, 55, 70, 85, 100, and 120) in Fig. 2. Note that we assumed that the higher strata costs would have larger minimum and maximum values with a slighter larger variation. This is evident in the figure as we go through the distributions for strata 1 to strata 120. The primary idea in constructing $h = 2, 3, ..., L$ strata for the main study variable, income, is that there exist h different average stratum costs which are then utilized in the computation of the OSB of the main study variable.

5.3. Computing the OSB and OSS for income

As indicated in Section 5.1, the main variable, *htdi*, follows a Gamma (2P) distribution. The corresponding MPP for (29) is solved using the DP solution procedure discussed in Section 3.2. Substituting the value of the $(h-1)^{th}$ stratification point as $x_{h-1} = d_h - l_h$, the recurrence relations given in equations (32) and (33) below are used for solving the MPP (29). For the initial stage, $k = 1$, at $l_1^* = d_1$:

$$\begin{aligned}\Phi_1 d_1 = \sqrt{\left\{ \theta^2 r(r+1) \left[Q\left(r, \frac{y_0}{\theta}\right) - Q\left(r, \frac{d_1 + y_0}{\theta}\right)\right]\right.} \\ \times \left[Q\left(r+2, \frac{y_0}{\theta}\right) - Q\left(r+2, \frac{d_1 + y_0}{\theta}\right)\right] \\ \left. -\theta^2 r^2 \left[Q\left(r+1, \frac{y_0}{\theta}\right) - Q\left(r+1, \frac{d_1 + y_0}{\theta}\right)\right]^2 \times c_1 \right\}\end{aligned} \tag{32}$$

and for stages $k \geq 2$:

Table 1
Results for stratification of income with respective average stratum costs.

Strata (L)	Cost (c_h in \$)	OSB (x_h in \$)	Weight (W_h)	VOF ($\sum W_h S_h$)	OSS (n_h)	Pop. (N_h)	Fraction (f_h)
2	13.69	115023.7	0.7	75003.3	181	6732	0.03
	13.45		0.3	132314.1	319	2887	0.11
3	13.79	76883.7	0.48	33666.2	119	4629	0.03
	13.52	166282.5	0.39	35886.3	127	3765	0.03
	13.26		0.13	72052.6	254	1225	0.21
4	13.79	59539.14	0.36	18839.2	88	3475	0.03
	13.59	116861.78	0.35	20592.2	96	3334	0.03
	13.51	202350.17	0.22	17943.1	84	2079	0.04
	13.23		0.08	49610.2	232	731	0.32
5	13.81	49141.41	0.29	12315.6	70	2775	0.03
	13.69	92101.86	0.29	13559.2	77	2825	0.03
	13.60	145955.58	0.24	13686.1	78	2277	0.03
	13.22	230047.41	0.13	10776.9	61	1236	0.05
	13.16		0.045	37550.9	214	506	0.42
6	13.83	42239.63	0.23	8512.3	57	2442	0.03
	13.76	77054.89	0.25	9390.3	63	2394	0.03
	13.45	117618.63	0.23	9665.7	65	2204	0.03
	13.56	170012.71	0.17	8982.2	60	1618	0.04
	13.37	252455.67	0.08	6549.4	44	763	0.06
	13.12		0.04	31135.1	210	398	0.53

$$\Phi_k d_k = \min_{0\le l_k\le d_k}\left\{\sqrt{\left\{\theta^2 r(r+1)\left[Q\left(r,\frac{d_k-l_k+y_0}{\theta}\right)\right.\right.\right.$$
$$\left.-Q\left(r,\frac{d_k+y_0}{\theta}\right)\right]\times\left[Q\left(r+2,\frac{d_k-l_k+y_0}{\theta}\right)\right.$$
$$\left.-Q\left(r+2,\frac{d_k+y_0}{\theta}\right)\right]-\theta^2 r^2\times\left[Q\left(r+1,\frac{d_k-l_k+y_0}{\theta}\right)\right.$$
$$\left.\left.\left.-Q\left(r+1,\frac{d_k+y_0}{\theta}\right)\right]^2\times c_k\right\}+\Phi_{k-1}(d_k-l_k)\right\}, \tag{33}$$

where c_1 and c_k $(k\ge 2)$ are computed in phase 1, as demonstrated in Section 4 5.2.

Substituting the quantities for r, θ, y_0 and d, the OSW (l_h^*) and the OSB ($y_h^* = y_{h-1}^* - l_h^*$) are determined by executing the updated **strata.data()** function of the `stratifyR` package [32]. The updated function, which has arguments such as the number of strata, income data, total sample size, and average stratum costs per unit, solves the recurrence relations (32) and (33) from the DP procedure. The stratification results pertaining to the estimated average stratum costs (c_h) from an apriori lognormal distribution of per unit stratum costs for $L = 2,3,..,6$ are presented in the form of OSB (x_h), OSS (n_h), the optimum Values of the Objective Function (VOF) and other important quantities. The given budget (C) and overhead cost (c_0) are not required in computing the OSB as equation (9) reduces to (10), however, they are required to compute the OSS, given by equation (8). The results for the stratification of income are presented in Table 1 where the OSS was computed by fixing $C = 25,000$ and $c_0 = \$15$.

With varying average stratum costs, the method would produce varying OSB and OSS, which means that every single sampling plan will be dependent on its unique costing paradigm. The availability of cost information enables accurate calculation of the OSS that guarantees precise estimation of the characteristic under study.

5.4. Comparison of efficiency

Comparisons of the proposed method's effectiveness are not possible against other methods since no other methods have since been used to construct strata boundaries based on average stratum per unit measurement costs. In fact, the other methods are not based on cost or assume the cost to be constant in all strata. Thus, the best possible way to make any comparison is by assuming that the per unit stratum costs are fixed at $c_h = 1$/unit. This comparison will be a basic test of the effectiveness of the proposed algorithm against other methods (implemented in the `stratification` package in R [32]), which are the following:

1. Dalenius and Hodges (1959) Cum $\sqrt{f}$ method [12];
2. Geometric method by Gunning and Horgan (2004) [14] and
3. Lavallee-Hidiroglou (1988) method with Kozak's (2004) algorithm [7].

Table 2 compares the suggested method's outcomes to the three approaches mentioned above.

Table 2
Results comparison for gamma study variable with constant stratum costs ($c_h = 1$).

Strata	DP Method		Cum $\sqrt{f}$		Geometric		L-H	
(L)	OSB	OSS	OSB	OSS	OSB	OSS	OSB	OSS
2	115681.4	182	108209.9	161	9492.2	1	143851	265
		318		339		499		235
3	77676.6	120	72173.3	101	2080.9	1	89654	175
	167607	126	162264.9	133	2080.9	14	227289.5	179
		253		266		485		146
4	60079.8	89	54155	70	974.3	1	65137.5	115
	117717.7	96	108209.9	89	9492.2	1	127776	112
	203791.8	83	198301.5	101	92480.3	108	268170.5	135
		231		240		390		138
5	49664	71	54155	87	618	1	56990	106
	93047.9	78	90191.6	53	3818.5	1	105537	98
	147514.3	77	144246.6	81	23596.1	3	171962	96
	231678.2	60	234338.2	69	145809.6	247	336313.5	96
		214		210		248		104
6	42672.1	58	36136.6	37	456.2	1	49587	99
	77874.8	63	72173.3	69	2080.9	1	87243	83
	118519.6	65	108209.9	55	9492.2	1	132601.5	84
	171391.3	60	162264.9	72	43300	21	213960.5	106
	254365.8	44	252356.5	57	197519.7	309	528390	120
		210		210		167		8

Table 3
Relative efficiency of proposed DP method over other methods.

Strata	Methods				Efficiency (%) of DP		
(L)	DP	CRF	GEO	L-H	CRF	GEO	L-H
2	41996.2	42113.6	72736.5	47635.2	100.28	172.71	113.43
3	29103.0	29176.1	56698.8	31139.2	100.25	194.33	107.00
4	22264.1	24451.3	42008.6	23639.4	109.82	171.80	106.18
5	18027.2	18184.4	37439.2	19615.2	100.87	205.89	108.81
6	15144.8	15303.0	35415.9	17784.3	101.04	231.43	117.43

The results indicate that the OSB and OSS vary slightly between the proposed method and CRF method. LH method produces somewhat larger values of the OSB while GEO method produces substantially lower OSB.

To compare the performance of the suggested DP method over other methods, the relative efficiency RE is calculated using the following formula [14]:

$$RE_{CRF,DP} = \frac{V_{CRF}(\bar{x}_{st})}{V_{DP}(\bar{x}_{st})} \tag{34}$$

The relative efficiency in equation (34) is then multiplied by 100% to give the percentage efficiency of the DP technique over the other three methods. Thus, Table 3 presents the objective function values and the relative efficiencies. It is seen that the proposed DP method's performance is on par with the CRF method, slightly more efficient than the L-H method, and far more efficient than the GEO method. This indicates that the proposed method works well compared to the three methods.

6. Conclusion

This paper presents a method for obtaining the OSB and OSS in a stratified sampling design while taking into account a fixed survey budget and varying per-unit measurement costs in each stratum. The method involves expressing the problem of stratification as an MPP, which is a function of the average stratum costs. The solutions are obtained by using the DP solution procedure, which has been implemented into the updated `stratifyR` package available on CRAN [40]. The procedure is illustrated with an application of the proposed method to the income variable from the HILDA survey in Australia, which best fits a 2-parameter Gamma distribution. Since the measurement costs are generally unknown, the survey design assumes this apriori, with the per-unit measurement costs following lognormal distribution. Furthermore, results for two more applications using hypothetical variables that follow exponential and right-triangular distributions respectively were also investigated (details of the results and discussion are presented in the Appendix section), where the average stratum costs were assumed. It was seen that similar to the real data, the proposed method is found to be promising with satisfactory results. The computed OSB and OSS are comparable with the three other most commonly used methods, where we find that the efficiency of the proposed methods is either on par or more efficient

than those methods. This implies that the proposed method is an efficient method for optimum stratification with survey cost as a constraint. This research incorporates cost into the determination of OSB and OSS, which is in itself an advantage as it considers survey cost while other methods treat the cost as a constant or fixed throughout all strata, which practically sounds counter-intuitive. Thus, when it comes to the actual planning of a survey where cost is involved, which consequently affects the sample allocation, this method can be used when some information regarding the survey cost is available to the surveyor. In the future, our research in this area could potentially involve adopting more innovative sampling strategies with regards to reducing data collection costs and developing methods to calculate optimal sample sizes by understanding the trade-offs between sample size, cost, survey mode, and statistical power. We can also explore adaptive survey designs that dynamically adjust the sample sizes which can reduce the overall survey costs. Bayesian approaches can also be used together with prior information to get better sample allocations - all of which are geared towards reducing sampling costs and increasing the precision of estimates.

CRediT authorship contribution statement

Karuna G. Reddy: Conceptualization, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **M.G.M. Khan:** Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data associated with this study has been deposited at https://ada.edu.au.

Appendix A

In this appendix, we provide further numerical investigations for two more hypothetical datasets with Exponential and Right-triangular distributions. Note that these are presented just as an illustration and we can apply the proposed method if our main study variables follow these two distributions:

A.1. Formulation of the MPP for exponential variable

Let x, the stratification variable, follow an exponential distribution with the rate (λ) parameter. Since the actual populations are frequently finite in practice, assuming the largest value of x in the population as D, the probability density function given can be presented as follows:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}; & 0 \le x \le D \\ 0; & elsewhere \end{cases} \tag{35}$$

with λ being the rate parameter and the pdf has a domain of $x_0 = 0$ and $x_L = D$. If D is sufficiently big enough, (35) could approximate an exponential density function, otherwise, the expression must utilise the truncated exponential density.

For the variable x that follows the exponential distribution represented by equation (35), using the equations (12), (13), and (14), the stratum weight W_h, stratum mean μ_h and the stratum variance S_h^2 can be obtained as a function of boundary points (x_{h-1}, x_h). These are provided below:

$$W_h = e^{-\lambda x_{h-1}}\left(1 - e^{-\lambda l_h}\right) \tag{36}$$

Stratum mean (μ_h) can be expressed by equation (37) below:

$$\mu_h = \frac{\left(x_{h-1} + \frac{1}{\lambda}\right)\left(-e^{-\lambda l_h} + 1\right) - l_h e^{-\lambda l_h}}{1 - e^{-\lambda l_h}} \tag{37}$$

Similarly, the stratum variance (S_h^2) is reduced and simplified to

$$S_h^2 = \frac{\frac{1}{\lambda^2}\left(1 - e^{-\lambda l_h}\right)^2 - l_h^2 e^{-\lambda l_h}}{\left(1 - e^{-\lambda l_h}\right)^2} \tag{38}$$

Thus, by substituting (36) and (38) into (10) gives $W_h S_h \sqrt{c_h}$ as:

$$\sqrt{\left(e^{-\lambda x_{h-1}}\right)^2 \left[\frac{1}{\lambda^2}\left(1 - e^{-\lambda l_h}\right)^2 - l_h^2 e^{-\lambda l_h}\right] \sqrt{c_h}} \tag{39}$$

Using the general form of MPP given in equation (20), in order to solve for the OSB and OSS for an exponential study variable, we express the MPP for equation (39) as follows:

$$\begin{aligned} &\text{Minimize} \sum_{h=1}^{L} \sqrt{\left(e^{-\lambda x_{h-1}}\right)^2 \left[\frac{1}{\lambda^2}\left(1-e^{-\lambda l_h}\right)^2 - l_h^2 e^{-\lambda l_h}\right] c_h} \\ &\text{subject to} \sum_{h=1}^{L} l_h = d \\ &\quad \text{and} \quad l_h \geq 0; h = 1, 2, ..., L \end{aligned} \tag{40}$$

with $d = x_L - x_0$ denoting the distance of the distribution, λ denotes the rate parameter of the exponential distribution and c_h are the average per unit stratum costs.

A.2. Numerical illustration of the solution procedure

Using different average per unit measurement stratum costs, the proposed solution procedure's computational details are illustrated for a study variable that follows an exponential distribution, using the R Statistical Software Package [40], we consider a randomly generated dataset of size $N = 10,000$ with a parameter of $\lambda = 1$, $x_0 = 2.03e^{-5}$ and $x_L = 8.50$ in Equation (35).

Thus, in order to solve for the MPP (40) where $d = 8.5035$, we substitute the value of the $(h-1)^{th}$ stratification point as $x_{h-1} = x_0 + l_1 + l_2 + ... + l_{h-1} = d_h - l_h$ since it is almost equal to zero. The recurrence relations given by (22) and (21) reduce to the following two equations given by (41) and (42):

For the initial stage, $(k = 1)$, at $l_1^* = d_1$ we have

$$f(1, d_1) = \sqrt{\left[\left(1-e^{-d_1}\right)^2 - d_1^2 e^{-d_1}\right] c_1}. \tag{41}$$

For the stage $k \geq 2$, $f(k, d_k)$ is given as:

$$\begin{aligned} &\min_{0 \leq l_k \leq d_k} \left\{ \sqrt{\left[\left(e^{-(d_k - l_k)}\right)^2 \left(1-e^{-d_k}\right)^2 - d_k^2 e^{-d_k}\right] c_k} \right. \\ &\left. + f(k-1, d_k - l_k) \right\} \end{aligned} \tag{42}$$

The OSB $(y_h^* = y_{h-1}^* - l_h^*)$ are calculated by solving the two recurrence relations (41) and (42) using the DP procedure implemented in the `stratifyR` package [40]. Assume arbitrary values of the estimated average stratum costs (c_h) and fixing a total sample size of $n = 500$, the results for the stratification of income, in the form of OSB (x_h), OSS (n_h), the optimum Values of the Objective Function (VOF) are presented in Table 4 below for $L = 2, 3, .., 6$.

A.3. Comparison with other methods and discussion

Similar to Section 5.4 for the income variable in the real data, we carry out a comparison of the effectiveness of the proposed method against the other three methods. For comparison purposes, we generate another set of exponential variables, and the stratum costs are again taken to be $c_h = 1; h = 1, 2, ..., 6$. The randomly generated exponential data was found to have a rate parameter of $\lambda = 1.005059$, a minimum value of $x_0 = 1.617348e-4$ and a maximum of $x_L = 9.562454$, which leads to the distance of the distribution being $d = x_L - x_0 = 9.562292$. The histogram of the data set is given below in Fig. 3, which shows a perfect exponential distribution. Table 5 presents the results for the DP method and the three methods: CRF, Geo, and L-H. It shows that the OSB and OSS only slightly vary between all methods except the Geometric method, which appears to be lower than the DP method. The OSS in Geometric method is also quite one-sided in how the sample sizes are being assigned among the strata.

To compare how the proposed DP method performs against other methods, as seen in section 5.4, the relative efficiencies are calculated by using equation (34). Table 6 presents the relative efficiencies of the DP method over the others. It is seen that the performance of the proposed DP method is almost on par (only very slightly more) with other methods and substantially better than the Geometric method. This is in line with what we found with the income variable.

A.4. Formulation of the MPP for right-triangular variable

The frequency function of the right-triangular distribution is given by

$$f(x; a, b) = \begin{cases} \dfrac{2(b-x)}{(b-a)^2}; & a \leq x \leq b \\ 0; & elsewhere \end{cases} \tag{43}$$

Let variable x exhibit a right-triangular distribution with density function given by (43), then W_h, μ_h, and S_h^2 can be obtained as a function of the boundary points (x_{h-1}, x_h), by using (12), (13), and (14) respectively as equations (44), (45) and (46) derived below.

Table 4
OSB, OSS, and VOF for exponential study variable with varying average stratum costs.

Strata (L)	Cost (c_h)	OSB (x_h)	OSS (n_h)	VOF $\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)$
2	2	1.47	262	0.837
	3		238	
	3	0.7	148	
3	2	2.17	184	0.59
	3		168	
	4	0.48	106	
4	2	1.43	151	0.487
	3	2.84	124	
	4		120	
	5	0.35	81	
	2	1.09	129	
5	3	1.99	105	0.419
	4	3.36	92	
	5		92	
	2	0.51	113	
	3	1.05	92	
6	4	1.67	79	0.366
	5	2.47	71	
	5	3.82	72	
	6		74	

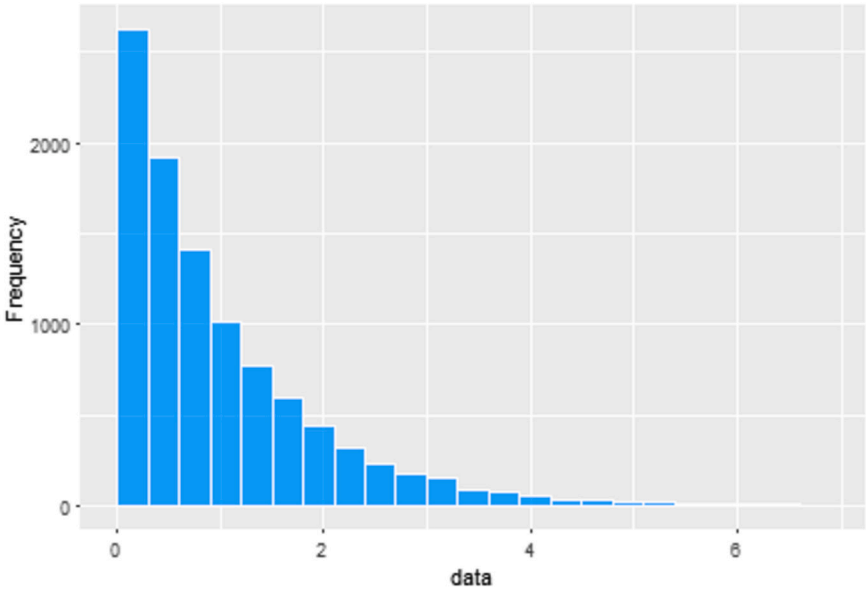


Fig. 3. Histogram of exponential distribution.

$$W_h = \frac{l_h(2a_h - l_h)}{(b-a)^2} \tag{44}$$

$$\mu_h = \frac{3b(l_h + 2x_{h-1}) - 2(l_h^2 + 3l_h x_{h-1}) + 3x_{h-1}^2}{3(2a_h - l_h)} \tag{45}$$

$$S_h^2 = \frac{l_h^2(l_h^2 - 6a_h l_h + 6a_h^2)}{18(2a_h - l_h)^2} \tag{46}$$

Thus, equation (47) presents the following MPP:

$$\text{Minimize} \quad \sum_{h=1}^{L} \frac{l_h(2a_h - l_h)}{(b-a)^2} \sqrt{\left[\frac{l_h^2(l_h^2 - 6a_h l_h + 6a_h^2)}{18(2a_h - l_h)^2}\right]} \sqrt{c_h}$$

Table 5
OSB and OSS for exponential study variable using different methods.

Strata	DP Method		Cum $\sqrt{f}$		Geometric		L-H	
(L)	OSB	OSS	OSB	OSS	OSB	OSS	OSB	OSS
2	1.25	236	1.34	257	0.04	1	1.22	226
		264		243		499		274
	0.76	160	0.77	162	0.01	1	0.75	156
3	2.01	161	2.10	174	0.25	9	1.96	152
		180		164		490		192
	0.55	121	0.57	131	0.00	1	0.53	113
4	1.31	121	1.34	118	0.04	1	1.27	117
	2.55	122	2.68	129	0.61	58	2.47	121
		136		122		440		149
	0.43	97	0.38	78	0.00	1	0.41	91
	0.97	97	0.96	109	0.01	1	0.94	91
5	1.73	98	1.72	98	0.12	3	1.68	95
	2.98	98	3.06	111	1.06	148	2.86	97
		109		104		347		126
	0.35	82	0.38	93	0.00	1	0.36	84
	0.78	82	0.77	64	0.01	1	0.80	82
6	1.33	82	1.34	88	0.04	1	1.36	84
	2.08	82	2.10	82	0.25	12	2.14	84
	3.33	82	3.44	92	1.53	236	3.38	80
		91		81		249		86

Table 6
Relative Efficiencies for exponential data.

Strata	Methods				Efficiency (%) of DP Over		
(L)	DP	CRF	GEO	L-H	CRF	GEO	L-H
2	0.529	0.530	0.953	0.529	100.2	180.2	100.0
3	0.361	0.362	0.786	0.361	100.1	217.7	100.0
4	0.274	0.275	0.605	0.274	100.2	220.8	100.1
5	0.221	0.222	0.487	0.221	100.4	220.4	100.1
6	0.185	0.186	0.423	0.185	100.3	228.7	100.0

$$\text{subject to} \quad \sum_{h=1}^{L} l_h = d$$
$$\text{and} \quad l_h \geq 0; \;\; h = 1, 2, ..., L, \tag{47}$$

where $d = x_l - x_0$ is the range of the distribution.

A.5. Computational details of the solution procedure

To illustrate the numerical details of the proposed solution procedure, assume that the initial and final values, $a = 1$ and $b = 2$, which gives $a_h = 2 - x_{h-1}$ and $d = 1$. This gives the following MPP:

$$\text{Minimize} \quad \sum_{h=1}^{L} \frac{l_h \sqrt{\left(l_h^4 - 6a_h l_h^3 + 6a_h^2 l_h^2\right) c_h}}{3\sqrt{2}}$$
$$\text{subject to} \quad \sum_{h=1}^{L} l_h = 1$$
$$\text{and} \quad l_h \geq 0; \;\; h = 1, 2, ..., L. \tag{48}$$

For solving the MPP (48), we use $d = 1$ and substitute the value of the $(h-1)^{th}$ stratification point as $x_{h-1} = x_0 + l_1 + l_2 + ... + l_{h-1} = 1 + d_h - l_h$ since $x_0 = 0$. The recurrence relations given by (22) and (21) shorten to:

For the initial stage, $k = 1$

$$f(1, d_1) = \frac{d_1 \sqrt{(d_1^4 - 6d_1^3 + 6l_1^2)c_1}}{3\sqrt{2}} \tag{49}$$

Table 7
OSB, OSS, and VOF for right-triangular study variable with varying average stratum costs.

Strata (L)	Cost (c_h)	OSB (x_h)	OSS (n_h)	VOF $\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)$
2	2	7.47	267	2.874
	3		233	
	2	5.68	194	
3	3	9.84	159	2.1
	4		147	
	4	3.73	107	
4	2	7.29	152	1.65
	3	10.83	125	
	4		116	
	5	3.16	81	
	4	5.14	90	
5	2	8.3	128	1.384
	3	11.46	105	
	4		97	
	6	2.72	60	
	4	4.3	73	
6	2	6.72	104	1.128
	3	8.96	85	
	5	10.95	66	
	2		112	

For the stage $k \geq 2$:

$$f(k, d_k) = \min_{0 \leq l_k \leq d_k} \left[d_k \, \text{sqrt}\left\{ \left(d_k^4 - 6\left(1 - d_h + l_h\right) d_k^3 + 6\left(1 - d_h + l_h\right) d_k^2 \right) c_k \right\} \right.$$
$$\left. \div 3\sqrt{2} + f\left(k - 1, d_k - l_k\right) \right] \tag{50}$$

The OSB ($y_h^* = y_{h-1}^* - l_h^*$) are computed by solving the recurrence relations (49) and (50) using the DP procedure implemented in the `stratifyR` package [32]. Assuming arbitrary values of the estimated average stratum costs (c_h) and fixing a total sample size of $n = 500$, the results for the stratification of income, in the form of OSB (x_h), OSS (n_h), the optimum Values of the Objective Function (VOF) are given in Table 7 for $L = 2, 3, .., 6$.

A.6. Comparison with other methods and discussion

As in Section 5.4, we investigate the effectiveness of the proposed method by comparing it against the other methods. The stratum costs are again taken to be $c_h = 1; h = 1, 2, ..., 6$. A data set of size $N = 10,000$ was randomly generated, which followed a right-triangular distribution with parameters $a = 1$ and $b = 2$. The minimum was found to be $x_0 = 1.007202$ and the maximum was $x_L = 1.999983$, which gives the range of the distribution as $d = x_L - x_0 = 0.992781$. The histogram of the right-triangular variable is given in Fig. 4.

The OSB, OSS and the VOF for the suggested DP method are compared to the three other methods. The OSB and the VOF are computed using the `stratifyR` package keeping all the stratum costs constant ($c_h = 1$). Table 8 presents the results for all the methods. In this example, it is seen that the OSB and OSS only slightly vary when compared to the Geometric method, while being quite different from the CRF and L-H methods.

To compare the performance of the suggested DP method over other methods, we generate Table 9, which presents the relative efficiencies. As seen previously in the other two datasets, in terms of the performance of the proposed DP method, it is found to be more efficient than all other methods, especially the CRF and L-H (Kozak) methods. However, in this example, we see that it is only slightly more efficient than the Geometric method.

References

[1] W.G. Cochran, Sampling techniques, 2007.
[2] R.M. Groves, Survey Errors and Survey Costs, John Wiley & Sons, 2005.
[3] Leslie Kish, Survey Sampling, John Wesley & Sons, New York, 1965.
[4] T. Zaman, An efficient exponential estimator of the mean under stratified random sampling, Math. Popul. Stud. 28 (2) (2021) 104–121.
[5] T. Dalenius, The problem of optimum stratification, Scand. Actuar. J. 1950 (3–4) (1950) 203–213.
[6] T. Dalenius, M. Gurney, The problem of optimum stratification. II, Scand. Actuar. J. 1951 (1–2) (1951) 133–148.
[7] P. Lavallie, Two-way optimal stratification using dynamic programming, Proc. Sect. Surv. Res. Methods (1988) 646–651.
[8] P. Mahalanobis, Some aspects of the design of sample surveys, Sankhya, Indian J. Stat. (1952) 1–7.

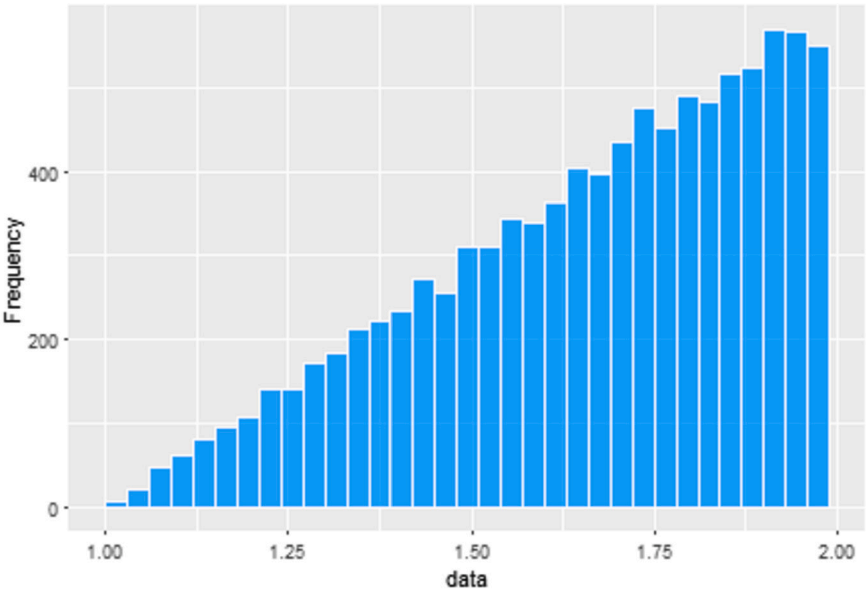


Fig. 4. Histogram of right-triangular distribution.

Table 8
OSB and OSS for right-triangular study variable using different methods.

Strata	DP Method		Cum $\sqrt{f}$		Geometric		L-H	
(L)	OSB	OSS	OSB	OSS	OSB	OSS	OSB	OSS
2	1.36	241	1.62	229	1.42	56	1.65	259
		259		271		444		241
	1.24	162	1.48	157	1.27	20	1.5	172
3	1.51	163	1.76	170	1.59	121	1.77	167
		175		173		359		161
	1.18	122	1.38	105	1.2	11	1.41	129
4	1.37	123	1.62	129	1.42	52	1.63	117
	1.59	123	1.82	132	1.68	137	1.82	125
		132		134		300		129
	1.14	98	1.34	95	1.16	6	1.37	116
	1.29	98	1.54	98	1.33	29	1.56	98
5	1.45	98	1.7	90	1.52	72	1.72	93
	1.65	99	1.86	115	1.74	141	1.86	94
		106		102		252		99
	1.12	82	1.31	78	1.13	4	1.31	84
	1.24	82	1.48	83	1.27	19	1.48	76
6	1.37	82	1.62	74	1.42	42	1.63	86
	1.51	82	1.76	94	1.59	79	1.76	87
	1.69	83	1.88	81	1.78	137	1.89	82
		89		90		219		85

Table 9
Relative efficiencies for right-triangular data.

Strata	Methods				Efficiency (%) of DP Over		
(L)	DP	CRF	GEO	L-H	CRF	GEO	L-H
2	0.122	0.159	0.124	0.166	130.3	101.6	136.1
3	0.082	0.117	0.085	0.122	142.7	103.7	148.8
4	0.062	0.090	0.064	0.095	145.2	103.2	153.2
5	0.050	0.076	0.052	0.082	152.0	104.0	164.0
6	0.042	0.066	0.043	0.066	157.1	102.4	157.1

[9] M.H. Hansen, W.N. Hurwitz, On the theory of sampling from finite populations, Ann. Math. Stat. 14 (4) (1943) 333–362.
[10] H. Aoyama, A study of the stratified random sampling, Ann. Inst. Stat. Math. 6 (1) (1954) 1–36.
[11] G. Ekman, et al., Approximate expressions for the conditional mean and variance over small intervals of a continuous distribution, Ann. Math. Stat. 30 (4) (1959) 1131–1134.
[12] T. Dalenius, J.L. Hodges Jr, The choice of stratification points, Scand. Actuar. J. 1957 (3–4) (1957) 198–203.
[13] D. Hedlin, A procedure for stratification by an extended Ekman rule, J. Off. Stat. 16 (1) (2000) 15.
[14] P. Gunning, J.M. Horgan, A new algorithm for the construction of stratum boundaries in skewed populations, Surv. Methodol. 30 (2) (2004) 159–166.
[15] M.G.M. Khan, K.G. Reddy, D.K. Rao, Designing stratified sampling in economic and business surveys, J. Appl. Stat. 42 (10) (2015) 2080–2099.
[16] K.G. Reddy, M.G. Khan, S. Khan, Optimum strata boundaries and sample sizes in health surveys using auxiliary variables, PLoS ONE 13 (4) (2018), https://doi.org/10.1371/journal.pone.0194787.
[17] K.G. Reddy, M. Khan, Optimal stratification in stratified designs using Weibull-distributed auxiliary information, Commun. Stat., Theory Methods (2018) 1–20.
[18] W. Bühler, T. Deutler, Optimal stratification and grouping by dynamic programming, Metrika 22 (1) (1975) 161–175, https://doi.org/10.1007/BF01899725.
[19] M.G.M. Khan, N. Nand, N. Ahmad, Determining the optimum strata boundary points using dynamic programming, Surv. Methodol. 34 (2) (2008) 205–214.
[20] E.A. Khan, M.G.M. Khan, M.J. Ahsan, Optimum stratification: a mathematical programming approach, Calcutta Stat. Assoc. Bull. 52 (2002) 323–333.
[21] J.L. Warner, J.J. Berman, J.M. Weyant, J.A. Ciarlo, Assessing mental health program effectiveness: a comparison of three client follow-up methods, Eval. Rev. 7 (5) (1983) 635–658.
[22] M.F. Weeks, R.A. Kulka, J.T. Lessler, R.W. Whitmore, Personal versus telephone surveys for collecting household health data at the local level, Am. J. Publ. Health 73 (12) (1983) 1389–1394.
[23] H.J. Sixma, J.J. Kerssens, C.v. Campen, L. Peters, Quality of care from the patients' perspective: from theoretical concept to a new measuring instrument, Health Expect. 1 (2) (1998) 82–95.
[24] J.R. Hochstim, A critical comparison of three strategies of collecting data from households, J. Am. Stat. Assoc. 62 (319) (1967) 976–989.
[25] A.H. Walker, J.D. Restuccia, Obtaining information on patient satisfaction with hospital care: mail versus telephone, Health Serv. Res. 19 (3) (1984) 291.
[26] C.A. McHorney, M. Kosinski, J.E. Ware Jr, Comparisons of the costs and quality of norms for the sf-36 health survey collected by mail versus telephone interview: results from a national survey, Med. Care (1994) 551–567.
[27] I. O'Toole Brian, D. Battittutta, A. Long, K. Crouch, A comparison of costs and data quality of three health survey methods: mail, telephone and personal home interview, Am. J. Epidemiol. 124 (2) (1986) 317–328.
[28] A. Ullah, J. Shabbir, Z. Hussain, B. Al-Zahrani, et al., Estimation of finite population mean in multivariate stratified sampling under cost function using goal programming, J. Appl. Math. 2014 (2014).
[29] I. Fellegi, A. Sunter, Balance Between Different Sources of Survey Errors: Some Canadian Experiences, Statistics, Canada, 1974.
[30] B. S. H. M. L. N. M. N. W. N. W. R., M. Summerfield, M. Wooden, Hilda user manual – release 18, 2019.
[31] S. Lohr, Sampling: Design and Analysis, Nelson Education, 2009.
[32] K. Reddy, M.G. Khan, stratifyr: an r package for optimal stratification and sample allocation for univariate populations, Aust. N. Z. J. Stat. 62 (3) (2020) 383–405.
[33] L.-P. Rivest, S. Baillargeon, Stratification: univariate stratification of survey populations, r package version 2.2-6, https://CRAN.R-project.org/package=stratification, 2017.
[34] B. Richard, Dynamic Programming, vol. 89, Princeton University Press, 1957, p. 92.
[35] W.G. Cochran, Comparison of methods for determining stratum boundaries, Bull. Int. Stat. Inst. 38 (2) (1961) 345–358.
[36] Department of Social Services, M. I. of Applied Economic, S. Research, The household, income and labour dynamics in Australia (hilda) survey, general release 19 (waves 1-19), https://doi.org/10.26193/3QRFMZ, 2021.
[37] E. French, J.B. Jones, On the distribution and dynamics of health care costs, J. Appl. Econom. 19 (6) (2004) 705–721.
[38] F. Black, M. Scholes, The pricing of options and corporate liabilities, J. Polit. Econ. 81 (3) (1973) 637–654.
[39] S.G. Thompson, R.M. Nixon, How sensitive are cost-effectiveness analyses to choice of parametric distributions?, Med. Decis. Mak. 25 (4) (2005) 416–423.
[40] K.G. Reddy, M.G.M. Khan, stratifyR: optimal stratification of univariate populations, r package version 1.0-2, https://CRAN.R-project.org/package=stratifyR, 2019.