
Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **62**(3), 2020, 383–405 doi: 10.1111/anzs.12301

stratifyR: An R Package for optimal stratification and sample allocation for univariate populations

K. G. Reddy[1*] and M. G. M. Khan[2]

Australian National University and The University of the South Pacific

Summary

This R package determines optimal stratification of univariate populations under stratified sampling designs using a parametric-based method. It determines the optimum strata boundaries (OSB), optimum sample sizes (OSS) and multiple other quantities for the study variable, y, using the best-fit probability density function of a study variable available from survey data. The method requires the parameters and other characteristics of the distribution of the study variable to be known, either from available data or from a hypothetical distribution if the data are not available. In the implementation, the problem of determining the OSB is formulated as a mathematical programming problem and solved by using a dynamic programming technique. If the data of the population (i.e. the study variable) are available to the surveyor, the method estimates its best-fit distribution and determines the OSB and OSS under Neyman allocation, directly. When the dataset is not available, stratification is made based on the assumption that the values of the study variable, y, are available as hypothetical realisations of proxy values of y from past/recent surveys. Thus, it requires certain distributional assumptions about the study variable. At present, the package handles stratification for the populations where the study variable follows a continuous distribution: namely, Pareto, Triangular, Right-triangular, Weibull, Gamma, Exponential, Uniform, Normal, Lognormal and Cauchy distributions. In this paper, applications of major functionalities in the package are illustrated with a number of real/simulated as well as some hypothetical populations.

Key words: dynamic programming; mathematical programming problem; optimum sample sizes; optimum strata boundaries; R project for statistical computing

1. Introduction

The main aim of stratification is to produce estimators with a small variance when a population characteristic (y) is under study. A simple method can be used to create strata for this population, if y itself is the stratification variable. The ideal situation is that the distribution of such a study variable is known and the optimum strata boundaries (OSB) can be determined by placing boundaries on the range of this distribution at suitable cut-points. This problem of determining the OSB, when both the estimation and stratification variables are the same, was first discussed by Dalenius (1950). He provided equations for the determination of stratum boundaries that minimise the variance of population estimates

*Author to whom correspondence should be addressed.
[1]Centre for Social Research & Methods, Australian National University, Canberra, ACT 2601, Australia. e-mail: karuna.reddy@anu.edu.au
[2]School of Computing, Information & Mathematical Sciences, The University of the South Pacific, Suva, Fiji.

under optimal allocation. Dalenius (1957) further proposed a solution to the problem by taking equal intervals of the cumulative square root of frequency scale of the stratification variable.

One of the many kinds of stratification methods that has been proposed in the literature is due to Bühler & Deutler (1975). They formulated the problem of determining the OSB as an optimisation problem and developed a computational technique to solve the problem by using dynamic programming (DP). A good review of this method can be found in Khan, Nand & Ahmad (2008). The DP procedure is applied in some of the following articles by Khan, Khan & Ahsan (2002); Khan, Reddy & Rao (2015); Reddy, Khan & Khan (2018) and Reddy & Khan (2019) for determining OSB for many different distributions. With the known frequency function of the study variable, they considered the problem of finding OSB as an equivalent problem of determining optimum strata width (OSW), which is formulated as a mathematical programming problem (MPP) and solved by using DP techniques. All the authors cited above applied the technique to several univariate populations where the study variables followed different probability distributions. The authors have established that the method certainly works with a variety of different populations, skewed and unskewed, giving precise and accurate results.

This called for an implementation of the idea into an R package that would be available to the surveyors an additional tool to create more accurate stratification boundaries. Another package from Rivest & Baillargeon (2017) called stratification solves a similar type of data-based stratification problem by implementing iterative or approximate methods. In the proposed R package, the univariate stratification technique implemented is primarily based on the probability distribution assumed by the stratification variable. With the objective of improving survey estimation efforts, the package implements the methods for various distributions, namely Uniform, Triangular, Right-triangular, Pareto, Exponential, Normal, Lognormal, Cauchy, Weibull and Gamma developed by the authors mentioned above. The package is able to determine the OSB and optimum sample sizes (OSS) for important study variables from available survey data; however, the key advantages of the method (hence, the package) is that it is able to construct OSB and OSS based on the distributional assumptions of a hypothetical dataset (i.e. when the study variable data are not available to the surveyor). The assumptions, such as initial value, range, estimated parameter values and best-fit distribution can easily be obtained as rough estimates from recent or past surveys.

Other advantages of the method are that it leads to substantial gains in the precision of the estimates over other available methods. Results from Khan *et al.* (2015); Reddy *et al.* (2018) and Reddy & Khan (2019) reveal that the variances get smaller with increasing number of strata (L), and they get much smaller at a much faster rate than other available methods. Once the OSB have been determined, the OSS can be easily calculated for each stratum using Neyman (1934) allocation.

There are two major functions which basically solve the two types of stratification problems: `strata.data`, which carries out univariate stratification for those populations, where the dataset is available and `strata.distr` which performs stratification when the dataset is not available prior to conducting the survey.

In the former case, data on the study variable, number of strata h, fixed sample size n and population size N are used as the input arguments to the `strata.data` function in the package. In the latter case, `strata.distr` function is called which requires the distribution to be assumed: its parameters, the initial value and the estimated range of the

distribution; fixed sample and population sizes. When executed, both functions output the OSB and OSS, among other quantities such as stratum weight (W_h), stratum variance (S_h^2), stratum objective function values (W_hS_h), stratum sample sizes (n_h), stratum population sizes (N_h) and stratum sampling fraction (f_h).

The following sections show the general formulation of the problem of stratification, the DP solution procedure, the concept of Neyman allocation as the method of determining stratum sample sizes and an overview of package functionalities. To illustrate, the package is applied to a Pareto Type II distributed study variable from a simulated dataset. The illustrations for three more distributions (Normal, Gamma and Lognormal) of the survey variables are presented in the Supplementary section.

2. General formulation of the problem, solution procedure and sample sizes

Khan *et al.* (2002) and Khan *et al.* (2008) presented a detailed description of the methodology of formulating the problem of stratification using MPP. To understand the problem at hand and the formulation, let the target population of the variable under study be stratified into L strata where the estimation of the mean of the study variable (y) is of interest. If a simple random sample of size n_h is to be drawn from hth stratum with sample mean $\bar{y}_h$, then the stratified sample mean, $\bar{y}_{st}$, is given by

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

where W_h (stratum weight) is the proportion of the population contained in the hth stratum.

When the finite population correction factors are ignored, under the Neyman (1934) allocation, the variance of $\bar{y}_{st}$ is given by

$$\text{var}(\bar{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n}, \tag{1}$$

where S_h^2 is the stratum variance for the study variable in the hth (where $h = 1, 2, \ldots, L$) stratum and n is the preassigned total sample size.

Let $f(y)$; $a \leqslant y \leqslant b$, be the frequency function of the study variable, y, on which OSB are to be constructed. If the population mean of this study variable is estimated under Neyman allocation, then the problem of determining OSB is to cut up the range, $d = b - a$, at $(L-1)$ intermediate points $a = y_0 \leqslant y_1 \leqslant y_2 \leqslant, \ldots, \leqslant y_{L-1} \leqslant y_L = b$ such that (1) is minimum. The lower and upper bounds of the study variable are denoted by a and b, respectively, and the cut-points $y_1, y_2, \ldots, y_{L-1}$ are the OSB. For a fixed sample size n, minimising the expression of the right hand side of (1) is equivalent to minimising

$$\sum_{h=1}^{L} W_h S_h. \tag{2}$$

If $f(y)$ is known and integrable, the stratum weight (W_h), stratum variance (S_h^2) and stratum mean (μ_h) can be obtained as a function of the boundary points y_h and y_{h-1} by using the following expressions:

$$W_h = \int_{y_{h-1}}^{y_h} f(y)\,dy, \tag{3}$$

$$S_h^2 = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 f(y)\,dy - \mu_h^2, \tag{4}$$

where

$$\mu_h = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y f(y)\,dy, \tag{5}$$

and where (y_{h-1}, y_h) are the boundaries of hth stratum.

Thus, the objective function in (2) could be expressed as a function of boundary points y_h and y_{h-1} only. We further define

$$l_h = y_h - y_{h-1};\ h = 1, 2, \ldots, L, \tag{6}$$

where $l_h \geqslant 0$ denotes the range or width of the hth stratum. Then, the range of the distribution, $d = b - a$, is expressed as a function of stratum width as:

$$\sum_{h=1}^{L} l_h = \sum_{h=1}^{L} (y_h - y_{h-1}) = b - a = y_L - y_0 = d. \tag{7}$$

The hth stratification point $y_h;\ h = 1, 2, \ldots, L$ is then expressed as $y_h = y_{h-1} + l_h$ and from (7), the problem can be treated as an equivalent problem of determining the OSW: $l_1, l_2, \ldots, l_L$. Due to the special nature of functions, the problem may be treated as a function of l_h alone and can be expressed as:

$$\begin{aligned} &\text{Minimise} && \sum_{h=1}^{L} \phi_h(l_h), \\ &\text{subject to} && \sum_{h=1}^{L} l_h = d, \\ &\text{and} && l_h \geqslant 0;\ h = 1, 2, \ldots, L. \end{aligned} \tag{8}$$

To solve the non-linear MPP (8), Khan *et al.* (2002) and Khan *et al.* (2008) presented a detailed description of the DP procedure, which was the brainchild of Richard (1957). As remarked by the authors, DP is a computational method well suited for solving an MPP that may be treated as a multistage decision problem. The solution is found by decomposing the problem into stages where each stage is comprised of a single variable sub-problem. The solution for n stages is obtained by adding the nth stage to the solution of $n-1$ stages. The solution procedure guarantees an optimal feasible solution for each stage and hence it is also optimum and feasible for the entire problem. Khan *et al.* (2008) presented a good account of the methodological developments together with an application of the DP method on a Normal population. Reddy & Khan (2019) presented these steps of the solution procedure in algorithmic form.

When the OSB (y_h, y_{h-1}) have been determined, the OSS $n_h;\ h = 1, 2, \ldots, L$ that minimises the variance of the estimate can easily be computed. The sample sizes n_h are obtained for a fixed total sample of size n under the Neyman allocation for $h = 1, 2, \ldots, L$, and given as follows:

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}, \tag{9}$$

where W_h and S_h are derived in terms of the optimum boundary points (y_h, y_{h-1}).

In Neyman allocation, the total sample size is proportional to the stratum size multiplied by the standard deviation of the stratum. If the variances are specified correctly, Neyman allocation will give an estimator with smaller variance compared to proportional allocation (Lohr 2009).

In (9), it is also worth noting that the OSB (y_h, y_{h-1}) are so obtained that n_h must satisfy the restrictions:

$$1 \leqslant n_h \leqslant N_h, \tag{10}$$

where $N_h = NW_h$. Thus, restriction (10) indicates that the hth stratum must form with at least one unit and also avoids the problem of over-sampling.

3. Overview of package functionalities

The package is available through the CRAN website: `https://cran.r-project.org/web/packages/stratifyR/index.html`. For the numerical illustrations and application of the package, some of the real datasets such as `sugarcane` of Khan *et al.* (2015), `anaemia` of Reddy *et al.* (2018), `hies` and `math` data are provided with the stratifyR package. The `quakes` and `Boston` data provided in the datasets package in R statistical software may also be used for illustration purposes. The stratifyR package has also been tested on some published and commonly used datasets such as `UScities` and `UScolleges` data from Cochran (1961), `Debtors` data of Gunning & Horgan (2004), `REV84` variable for 'Swedish municipalities' data from Särdnal, Swensson & Wretman (1992) and `MRTS` variable from 'Statistics Canada Monthly Retail Trade Survey' together with `SHS` data collected in 'Statistics Canada Survey of Household Spending'. For those distributions where real data are not found in the literature, data may be simulated to demonstrate the application of the package in this documentation.

For a user, there are two different routes available in the package and these are basically dependent on the type of stratification problem. It could either be a data-based (i.e. when the stratification variable dataset is available) or a distribution-based (i.e. when the dataset is not available but certain distributional assumptions are made) stratification problem. Whether stratification is based on data or not, the idea is that the problem is formulated as an MPP using the estimated (with available data) or assumed (with unavailable data) distribution of the dataset. There are numerous functions created in the package for various technical calculations; however, there are only two major functions to compute the solutions for the two different types of problems being studied under univariate stratification.

If it is a data-based problem, the function used is `strata.data` and the user has to provide as input arguments: the data, the number of strata (L) and the fixed sample size (n). For the distribution-based problem, the function used is `strata.distr` and the user has to provide the name of the assumed distribution, number of strata (L), possible range of data (distance), fixed sample size (n) and the population size (N). The following two subsections delve a little deeper into the workings surrounding the two functions: `strata.data` and `strata.distr`.

To provide support for fitting of statistical distributions, there are a few R packages that the stratifyR package is dependent on, which are fitdistrplus (Delignette-Muller, Dutang & et al. 2015), MASS (Venables & Ripley 2002), zipfR (Evert & Baroni 2007), actuar (Dutang *et al.* 2008), triangle (Carnell 2017) and mc2d (Pouillot & Delignette-Muller 2010). The following subsection presents and discusses the two major functions that exist in the package.

3.1. The function `strata.data`

This function computes the OSB and OSS, and other important quantities from univariate survey populations by employing the methodology proposed by Khan *et al.* (2002), (2015), Reddy *et al.* (2018) and Reddy & Khan (2019). Their method uses the estimated distribution of the data and formulates the problem of determining OSB as an MPP, which is an optimisation problem that is solved by the DP technique as discussed in Section 2. The OSB obtained are optimal solutions that are used to calculate the OSS under Neyman allocation. The function appears as follows:

```
strata.data(data, h, n, cost = FALSE, ch = NULL)
```

The key arguments are `data`, which is a vector data containing every unit of the survey population; h is the number of strata to be sampled (i.e. $h = 1, 2, \ldots, L$) and n is the fixed total sample size where $\sum_{h=1}^{L} n_h = n$.

The steps used by `strata.data` can be described as follows:

1. `strata.data` function is of class `strata` which needs the specification of the arguments: `data`, `h` and `n`. If the stratification problem considers sampling cost, two further arguments, a logical stratum `cost` (assigned `TRUE` to indicate it is a cost problem) and a vector of individual stratum costs `ch` also need to be specified. If it is not a cost problem, `cost=FALSE` and `ch=NULL` are taken as defaults. This step creates a new environment called `my_env` to store all the arguments and various computations that take place such as scaling of data, various evaluations computed from the data and also invokes the following functions to determine the best-fit distribution and the estimated parameters, OSB from the DP procedure, objective function values for the MPP, sample size allocations and then combines key outputs into a list.
2. The `get.dist` function takes the data and quantities stored in `my_env` as arguments. From a set of ten different distributions (`unif`, `triangle`, `rtriangle`, `gamma`, `weibull`, `norm`, `lnorm`, `exp`, `pareto` and `cauchy`), it chooses the best-fit distribution by looking at the lowest AIC. Parameter estimates for the best-fit distribution together with the smallest AIC are returned as a list.
3. The `create.mat` function creates 2D matrices from a set of defined constants, which depend on the range of data and required precision, to store computed values of the objective function.
4. The `data.optim` function then computes 3 dp and 6 dp solutions, respectively, for different values of the objective function at different incremental progressions of the y value on the scaled range of the study variable. It invokes the `data.root` function, which implements the methods for ten different distributions to calculate the objective function values.
5. The `data.alloc` function computes the sample sizes by using Neyman allocation. The OSB obtained in the previous steps are used to calculate the stratum weights

and stratum population sizes from the data – these are then used to obtain the stratum sample sizes. In case of oversampling problems, this function tries to adjust the sample sizes by the `realloc` function.

6. The `summary.strata` function prints important quantities and results obtained through the above procedures. This function defines the method for the `strata` class that has been created in the constructor function `strata.data` where all computed objects are collated and passed as a list. Using such an S3 method is a conventional way of being able to integrate the package within the analysis workflow.

To show the `strata.data` function call, an example of the command used from the package is given below. The problem uses the 'mag' variable from the 'quakes' data (with a population of $N = 1000$) available from the `datasets` package in R. As an example, to construct a 5-strata solution, with a fixed sample size of $n = 300$, we use the following R codes:

```
library("datasets") #load the 'datasets' package
data(quakes); head(quakes) #look at the quakes data
mag <- quakes[, "mag"] #extract the 'mag' variable
res <- strata.data(mag, h = 5, n = 300) #create a 5-strata solution
summary(res) #print out the results
```

The resulting output from the codes, above, are omitted here. An example illustrating stratification of a survey data with Pareto Type II distribution will be presented in the results section. Three other examples involving Normal, Gamma and Lognormal distributions are also provided in the section to illustrate how the functions are utilised.

3.2. The function `strata.distr`

This function is also used to compute the OSB, OSS and other important quantities from univariate survey populations by employing the methodologies proposed in various literature (for various distributions) provided earlier. The algorithm is quite similar to that of the `strata.data`; however, its functionality is applied to the case where the dataset of the population is not available and the distributional assumptions of the study variable are required, which could be based on recent or past surveys. If no prior information exists, assumptions could be based on a purely hypothetical distribution. Another caveat for such distribution-based stratification is that the `distr.alloc` function uses the probability density functions of the assumed distributions and integration rules presented by (3)–(5) to calculate the stratum sample sizes. It must be noted that this function works on ideal distributions that assumes the parameters chosen by the user. The function appears as follows:

```
strata.distr(h, initval = NULL, dist = NULL,
   distr = c("pareto", "triangle", "rtriangle", "weibull", "gamma",
   "exp", "unif","norm", "lnorm", "cauchy"), params = c(shape=0,
   scale=0, rate=0, gamma=0, location=0, mean=0, sd=0, meanlog=0,
   sdlog=0, min=0, max=0, mode=0), n, N, cost = FALSE, ch = NULL)
```

The arguments could be explained as follows:

`h` – numeric: number of strata to be sampled
`initval` – numeric: initial value of the assumed distribution

`dist` – numeric: distance or range of the assumed distribution
`distr` – character: the assumed distribution of the hypothetical population
`params` – list: parameters of the assumed distribution
`n` – numeric: fixed total sample size
`N` – numeric: fixed population size
`cost` – logical: `TRUE` if it is a cost problem, `FALSE` by default
`ch` – vector: individual stratum costs

The sequence of steps in the algorithm for `strata.distr` is quite similar to the `strata.data` for the construction of OSB. Apart from the fact that it uses the distributional properties to determine the OSB (i.e. since there are no data, these are normally provided), it is at the sample allocation (OSS) stage that this function is also different from the data-based method. This is where the `distr.alloc` function is utilised in the calculation of the stratum sample sizes. Once all results have been computed, the step where they are collated and organised in a list of class `strata` is the same as in `strata.data`.

The following code demonstrates the application of `strata.distr` function when the dataset of the stratification variable is not available. As an example, to construct a 4-strata, let us consider the `depth` variable from the `quakes` dataset (assuming that it was made available from a recent survey) from the `datasets` package. It has a Triangular distribution with parameters `min=39.99998, max=680, mode=39.99999` and starts at an initial value of `initval=40` and has a distance (range) of `d=640` with a fixed sample size of `n=300` from a population of `N=1000` seismic events. Thus, we use the following commands:

```
data(quakes) #load the quakes data from 'datasets' package
depth <- quakes[, "depth"] #extract the depth variable
min(depth); max(depth); d=max(depth)-min(depth);d #evaluations
res <- strata.distr(h=4, initval=40, dist=640, distr = "triangle",
   params = c(min=39.99998, max=680, mode=39.99999),
   n = 300, N = 1000) #4-strata solution
summary(res) #print the results
```

Again, the outputs from the above codes are omitted as the aim was to illustrate how the function is called. Using the functions from the package, an in-depth example illustrating the MPP formulation and solution procedure involving Pareto Type II distribution, which is a new addition to the list of distributions covered in the literature, is presented in the next section.

4. Stratification for a survey variable with Pareto Type II distribution

The Pareto distribution, named after Italian scientist Vilfredo Pareto, is a power law heavy-tail probability distribution used in description of social, socio-economic, scientific, actuarial and many other observable phenomena. One notable field of its application, as presented by Arnold (2015), is size distribution of income or wealth. Many different forms for Pareto distribution exist in literature but for the purpose of this research, a Type II Pareto distribution, also called a Lomax distribution, after Lomax (1954), will be utilised.

Hassan & Al-Ghamdi (2009) utilised Pareto II distribution for reliability modelling and life testing. Atkinson & Harrison (1978) used it for modelling business failure data, while Corbellini *et al.* (2010) used it to model firm size and queueing problems. Bryson (1974)

advised the usage of Pareto II as an alternative to the exponential distribution when the data are heavy-tailed.

If the study variable y follows the Pareto Type II (or Lomax) distribution on the domain $[0, \infty)$, its two-parameter probability density function with a state space $y \geqslant 0$ is given by:

$$f(y; s, a) = \frac{as^a}{(y+s)^{a+1}}, \quad a, s > 0, \tag{11}$$

where $a > 0$ is the shape parameter and $s > 0$ is the scale parameter of the distribution.

The MPP for the Pareto Type II variable, which has a general form given by (8), is obtained from (3), (4) and (5). The formulated MPP could be expressed as:

$$\begin{aligned}
\text{Minimise} \quad & \sum_{h=1}^{L} \sqrt{\left\{ as^{2a} \left[\frac{(y_{h-1}+l_h+s)^a - (y_{h-1}+s)^a}{(y_{h-1}+s)^a (y_{h-1}+l_h+s)^a} \right] \right.} \\
& \times \left[\frac{(y_{h-1}+l_h+s)^{2-a}}{2-a} - \frac{2s(y_{h-1}+l_h+s)^{1-a}}{1-a} - \frac{s^2(y_{h-1}+l_h+s)^{-a}}{a} \right. \\
& \left. - \frac{(y_{h-1}+s)^{2-a}}{2-a} + \frac{2s(y_{h-1}+s)^{1-a}}{1-a} + \frac{s^2(y_{h-1}+s)^{-a}}{a} \right] \\
& \left. \times \frac{s^{2a}}{(1-a)^2} \left[\frac{a(y_{h-1}+l_h)+s}{(y_{h-1}+l_h+s)^a} - \frac{ay_{h-1}+s}{(y_{h-1}+s)^a} \right]^2 \right\} \\
\text{subject to} \quad & \sum_{h=1}^{L} l_h = d, \\
\text{and} \quad & l_h \geqslant 0;\ h = 1, 2, \ldots, L,
\end{aligned} \tag{12}$$

where $d = y_L - y_0$, a and s are parameters of the Pareto Type II distribution.

4.1. DP solution for the Pareto Type II distribution

To solve the formulated MPP (12), we apply the algorithm using the DP technique discussed within Section 2. Substitution of the quantity $y_{h-1} = y_0 + d_h - l_h$ into the MPP results in the following recurrence relations that are used to determine the solutions:

For the first stage, $k = 1$, at $l_1^* = d_1$:

$$\begin{aligned}
\Phi_1 d_1 = & \sqrt{\left\{ as^{2a} \left[\frac{(d_1+y_0+s)^a - (y_0+s)^a}{(y_0+s)^a (d_1+y_0+s)^a} \right] \right.} \\
& \times \left[\frac{(d_1+y_0+s)^{2-a}}{2-a} - \frac{2s(d_1+y_0+s)^{1-a}}{1-a} \right. \\
& \left. - \frac{s^2(d_1+y_0+s)^{-a}}{a} - \frac{(y_0+s)^{2-a}}{2-a} + \frac{2s(y_0+s)^{1-a}}{1-a} + \frac{s^2(y_0+s)^{-a}}{a} \right] \\
& \left. \times \frac{s^{2a}}{(1-a)^2} \left[\frac{a(d_1+y_0)+s}{(d_1+y_0+s)^a} - \frac{ay_0+s}{(y_0+s)^a} \right]^2 \right\},
\end{aligned} \tag{13}$$

and for stages $k \geqslant 2$:

$$\begin{aligned}\Phi_k d_k = \min_{0\leqslant l_k\leqslant d_k}\Bigg\{\sqrt{\Bigg\{as^{2a}\left[\frac{(d_k+y_0+s)^a-(d_k+l_k+y_0+s)^a}{(d_k+l_k+y_0+s)^a(d_k+y_0+s)^a}\right]}\\ \times\left[\frac{(d_k+y_0+s)^{2-a}}{2-a}-\frac{2s(d_k+y_0+s)^{1-a}}{1-a}-\frac{s^2(d_k+y_0+s)^{-a}}{a}\right.\\ \left.-\frac{(d_k+l_k+y_0+s)^{2-a}}{2-a}+\frac{2s(d_k+l_k+y_0+s)^{1-a}}{1-a}+\frac{s^2(d_k+l_k+y_0+s)^{-a}}{a}\right]\\ \times\frac{s^{2a}}{(1-a)^2}\left[\frac{a(d_k+y_0)+s}{(d_k+y_0+s)^a}-\frac{a(d_k+l_k+y_0)+s}{(d_k+l_k+y_0+s)^a}\right]^2\Bigg\}+\Phi_{k-1}(d_k-l_k)\Bigg\}.\end{aligned} \tag{14}$$

Upon substitution of the values of a, s, y_0 and d, the OSW (l_h^*) and the OSB ($y_h^* = y_{h-1}^* - l_h^*$) are obtained by executing the `strata.distr` function.

4.2. A numerical example for the Pareto Type II distribution

To illustrate the application of the functions from the package, a dataset for a univariate population (one study variable) of size $N = 5000$ and one that follows Pareto Type II distribution (`pareto_dat`) was simulated using parameters `shape=5` and `scale=8`.

The number of strata (h) is usually chosen by the surveyor and depends on how many mutually exclusive subgroups one is interested in. It has been recommended by Cochran (1961) that constructing six strata for a continuous variable is ideal because the gain in precision is minimal after six strata. If we are interested to construct the 6-strata solution (i.e. `h = 6`) for the `pareto_dat` with a fixed total sample size of `n=500`, the following codes could be used:

```
library(stratifyR) #load the package
set.seed(8235411) #to reproduce the random object
#simulate Pareto II random variable
pareto_dat <- rpareto(5000, shape = 5, scale = 8)
res <- strata.data(pareto_dat, h = 6, n = 500) #6-strata solution
summary(res) #print results
```

In the results obtained in the R console, apart from information on the best-fit frequency distribution, fitted parameters, minimum, maximum values of the data (and distance), the solutions in the form of OSB, OSS, etc., are also obtained. These can be presented in Table 1 as follows:

Similarly, in order to find the OSB and other quantities, we can apply the `strata.distr` function to a hypothetical Pareto Type II population, which could be based either entirely on

Table 1 Results for the Pareto Type II distribution using `strata.data`.

Stratum (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	0.74	0.35	0.05	0.08	83	1773	0.05
2	1.73	0.26	0.08	0.08	83	1318	0.06
3	3.15	0.18	0.16	0.07	80	909	0.09
4	5.44	0.12	0.4	0.08	85	615	0.14
5	10.15	0.06	1.73	0.08	87	303	0.29
6	38.57	0.02	29.31	0.09	82	82	1
Total		1	31.73	0.47	500	5000	0.1

assumptions or on prior knowledge (past or recent surveys) regarding the distribution of the variable. Let us assume that we have some information from a population with a particular study variable that follows Pareto Type II distribution. Consider as an example, the case of previous data that was simulated. If such information was available, key attributes of the distribution could be used. The data exhibits a two-parameter Pareto Type II distribution with the MLE estimates of the parameters as `shape=5.018971` and `scale=8.177219`. The minimum and maximum values in the simulated data are $[y_0, y_L] = [0.0002193, 38.56871]$, which implies that $d = 38.56849$.

Below is a sequence of commands to obtain the characteristics of the study variable and compute the stratification boundaries (for a 6-strata solution) based entirely on distribution:

```
d <- max(pareto_dat) - min(pareto_dat); d #evaluations
#fit the distribution to estimate parameters
fit <- fitdist(pareto_dat, distr="pareto", method="mle"); fit
res <- strata.distr(h=6, initval=0.0002193, dist=38.56849,
   distr = "pareto", params = c(shape=5.018971,
   scale=8.177219), n=500, N=5000) #six-strata solution
summary(res) #print results
```

The results are presented in Table 2 below:

For both illustrations, since the data and the hypothetical assumptions were the same, similar results were expected with both functions. There are only slight deviations, where `strata.distr` gives slightly better results as the overall objective function value ($W_h S_h$) is somewhat smaller. The deviation is simply because once the OSB have been determined, the `strata.distr` calculates the other quantities from an ideal Pareto Type II density function while `strata.data` calculates from the actual data.

Due to rounding off errors, there are occasions when the totals might not match with what is presented in the final table as the individual stratum figures have been rounded off. The stratum weights must total to one, the stratum samples should total to 500 while the stratum population sizes should total to 5000 for this particular example.

To integrate the package into a data analysis framework for visualisation and reporting, the package employs S3 classes and methods where various items that were computed are contained in a list object, which could be accessed and used for reporting. As a result, the outputs from the two functions `strata.data` or `strata.distr` can be integrated into an analysis tool chain within the R Markdown document or R Notebook. For example, to

Table 2 Results for the Pareto Type II distribution using `strata.distr`.

Stratum (h)	OSB (y_i)	W_h	V_h	$W_h S_h$	n_h	N_h	f_h
1	0.74	0.35	0.05	0.075	83	1769	0.05
2	1.73	0.27	0.08	0.075	82	1327	0.06
3	3.15	0.19	0.16	0.075	82	932	0.09
4	5.44	0.12	0.41	0.075	83	586	0.14
5	10.15	0.06	1.61	0.076	83	299	0.28
6	38.57	0.02	21.15	0.079	87	87	1
Total		1	23.46	0.457	500	5000	0.1

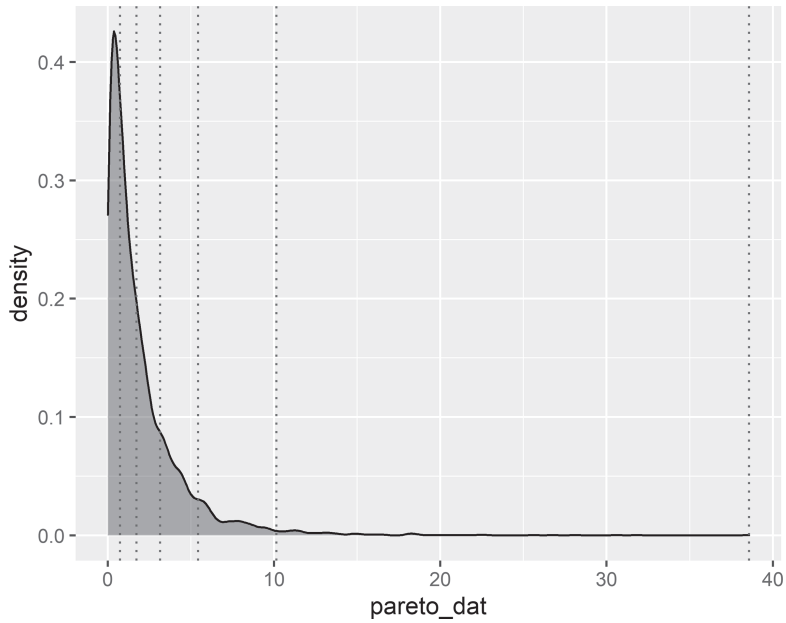

Figure 1. Visualisation of the density with overlay optimum strata boundaries.

illustrate how an end-user might dynamically utilise the output of `strata.data` function in the visualisation workflow: consider the example using the `strata.data` function where a 6-strata solution was constructed for `pareto_dat` data. The end-user is able to visualise the six strata created over the density plot with the following commands:

```
library(tidyverse) #load the tidyverse package
pareto_dat <- data.frame(pareto_dat) #convert to dataframe
pareto_dat % > % ggplot(aes(x = pareto_dat)) +
   geom_density(fill = "grey", colour = "black",
   alpha = 0.3, size=1) +
   geom_vline(xintercept = res$OSB, linetype =
   "dotted", color = "grey", size=1)
```

Figure 1 presents a density curve of the simulated `pareto_dat` variable, which follows Pareto Type II distribution, with the six stratification boundaries marked (in dots) as vertical lines on the curve.

5. Comparison with other methods

To compare the OSB obtained via the proposed DP method implemented in the stratifyR package, we use the stratification package to determine the OSB for the `pareto_dat` using the Cum $\sqrt{f}$, Geometric and L-H (Kozak) methods. Comparisons could also be made for $L = 2, 3, \ldots, 6$ strata; however, we will only compare 6-strata solutions. For comparison purposes, since the other methods work on data, we use the `strata.data` function which depends on data. The OSB and other important quantities for other methods are presented in Tables 3–5.

Table 3 Results for the Pareto Type II distribution using Cum $\sqrt{f}$ method.

Stratum (h)	OSB (y_i)	μ_h	V_h	N_h	n_h	f_h
1	0.77	0.34	0.05	1819	87	0.05
2	1.54	1.11	0.05	1062	51	0.05
3	3.09	2.18	0.19	1082	102	0.09
4	5.40	4.02	0.41	646	89	0.14
5	10.03	7.05	1.65	305	85	0.28
6	39.57	14.59	28.60	86	86	1.00
Total				5000	500	0.10

Table 4 Results for the Pareto Type II distribution using Geometric method.

Stratum (h)	OSB (y_i)	μ_h	V_h	N_h	n_h	f_h
1	0.00	0.00	0.00	4	1	0.25
2	1.01	1.01	0.00	37	1	0.03
3	0.09	0.05	0.00	245	1	0.00
4	0.69	0.37	0.03	1389	22	0.02
5	5.15	2.09	1.32	2898	312	0.11
6	39.57	8.42	16.78	427	163	0.38
Total				5000	500	0.10

Table 5 Results for the Pareto Type II distribution using L-H (Kozak) method.

Stratum (h)	OSB (y_i)	μ_h	V_h	N_h	n_h	f_h
1	0.60	0.28	0.03	1519	54	0.04
2	1.41	0.96	0.05	1215	56	0.05
3	2.58	1.92	0.11	998	67	0.07
4	4.24	3.32	0.22	639	60	0.09
5	7.40	5.38	0.73	429	74	0.17
6	39.57	11.12	21.66	200	189	0.94
Total				5000	500	0.10

Comparing Table 1 with Tables 3–5 reveal that the OSB obtained by the proposed DP method are very close to the Cum $\sqrt{f}$ method and the values of the objective function in the two methods are also quite similar. The OSB in the Geometric and L-H (Kozak) methods are quite different from the proposed method and their objective function values (given in Table 6) are also greater than the results from stratifyR package.

To compare the relative efficiency (RE) of the results from the stratifyR package with other methods, comparisons are made using the sum of objective function values, that is $\left(\sum_{h=1}^{L} W_h S_h\right)$. The results for other methods are provided in Table 6. It can be established that for the Pareto Type II simulated distribution, the package results in OSB are comparable to Cum $\sqrt{f}$ method (only slightly greater in RE) and are much more efficient than the other two methods (125% greater than Cum $\sqrt{f}$ method and 34% greater than L-H (Kozak) method).

Table 6 Relative efficiency of stratifyR results against other methods.

Stratum (h)	stratifyR	Cum $\sqrt{f}$	Geometric	L-H (Kozak)
1	0.076	0.081	0.000	0.068
2	0.076	0.047	0.000	0.054
3	0.074	0.094	0.000	0.087
4	0.077	0.083	0.048	0.081
5	0.080	0.078	0.666	0.110
6	0.089	0.092	0.350	0.231
$\sum_{h=1}^{L} W_h S_h$	0.472	0.475	1.064	0.631
Relative Efficiency		100.64%	225.42%	133.69%

6. Scope for future developments

The package stratifyR is limited to ten distributions which are primarily two-parameter (2P) distributions. This is because the stratifyR package uses many dependent packages available in R (particularly for parameter estimation) which are generally able to handle 2P distributions. A possible upgrade of the package will entail a multitude of optional distributions that would fit data with the best possible distribution. The possibility of including three- or four-parameter distributions can also be explored. Future versions can also consider including other allocation procedures like proportional and optimum allocations.

During the process of generating results using the stratifyR package, time-complexity issues arose because the solution procedure is quite time-consuming, especially for $h \geqslant 4$ onward. The program is still quite slow even in C++ computing environment; hence, faster convergence with increased computer processing power will be something to look at in future versions of the package. The possibility of using cluster computers or cloud computers will be explored to find out if the algorithm can execute faster. The very nature of the method of DP solution procedure introduces the 'curse of dimensionality' problem because the method is a brute force algorithm which is naturally very slow. Improved computer processing power will surely solve the problem.

7. Summary and discussion

This paper presents an R package called stratifyR, which deals with the concept of univariate stratification. The package, which is available through CRAN, successfully implements the DP technique based on a parametric method of optimum stratification of populations that follow any one of the ten standard statistical distributions, namely, Pareto, Triangular, Right-triangular, Weibull, Gamma, Exponential, Uniform, Normal, Lognormal and Cauchy. If particular data do not fit any one of the ten distributions, it will choose the next best-fit distribution.

In this research, the concept of constructing OSB from data directly was motivated by the fact that non-statistics people are able to do stratification without having to estimate the distribution. For those who have the know-how of estimating a distribution and its parameters, they would be able to estimate and confirm the best-fit distribution with Kolmogorov–Smirnov, Shapiro–Wilk and Anderson–Darling tests, together with exploratory and graphical analysis. Once done, the OSB and OSS could be computed much more efficiently if more accurate distributions are identified and fitted to the available survey data. Thus, the efficiency

of the OSB and OSS of the proposed method over others really depends on how well data are fitted to their ideal distribution.

The results presented in this paper, using the four distributions, illustrate that the stratified designs can be constructed with the proposed methodology in the stratifyR package. The OSB, OSS, etc., are presented and their performances against other established methods such as Cum $\sqrt{f}$, Geometric and L-H (Kozak) methods can also be compared. The real advantage is that when the data of the study variable are not available, which may occur in practice, the package is still able to construct the OSB and OSS based on the distributional assumptions of the data, which could be estimated or ascertained from recent, past or pilot surveys. These assumptions include some of the important quantities like the range of data, initial value of the data, a guesstimate distribution of the data and its associated parameters.

8. Supplementary Materials

To illustrate the use of stratifyR package with more examples, this supplementary section presents the results for the stratification of Normal, Lognormal and Gamma study variables. In total, as stated in the paper, the package is able to handle a total of ten continuous distributions that are quite commonly used in real-life situations. In the following sections, for each of the three distributions, a brief overview and an application of the relevant functions (using either real or simulated data) are given. Results obtained are not compared with other methods as these comparisons have already been done in the respective literature. Examples for hypothetical distributions (i.e. distribution-based stratification) are also presented under the three distributions. For the sake of brevity, only the recurrence relations of the DP solution procedure for determining OSB are provided.

8.1. Stratification of a survey variable with Normal distribution

The Normal distribution is commonly known for its bell shape and is considered to be the most widely known and used of all distributions. Many natural phenomena or biological, physical and psychological measurements/characteristics approximate normal distributions; hence, it is usually considered as a standard of reference for many probability problems in real situations. As an example, Lewis (1957) showed that characteristics such as height and intelligence are approximately normally distributed.

If the study variable y follows the Normal distribution on the interval $[y_0, y_L]$, it has the following two-parameter probability density function:

$$f(y;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\},\quad -\infty<y<\infty,$$

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

The following definitions of error function are worth noting since they are needed to simplify the integrations used to derive the stratum weight, mean and variance due to normal distribution.

$$\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z \exp\left\{-y^2\right\}dy.$$

It can also be written as

$$\frac{1}{\sqrt{2\pi}}\int_0^z \exp\left\{-\frac{1}{2}y^2\right\}dy = \frac{1}{2}\text{erf}\left(\frac{z}{\sqrt{2}}\right).$$

Then, using (3)–(5), as discussed in Section 4 for the Pareto Type II distribution, the MPP for determining the OSB for the normal study variable is obtained as seen in Khan *et al.* (2008).

8.2. DP solution for Normal distribution

To solve the MPP formulated for the Normal distribution, we apply the algorithm using the DP technique discussed in Section 2. The recurrence relations to determine the OSB for the Normal distribution are derived as follows:

For the first stage, $k=1$, at $l_1^* = d_1$:

$$\begin{aligned}\Phi_1 d_1 = \sqrt{\Bigg\{} & \frac{\sigma^2}{2\sqrt{2\pi}}\left[\text{erf}\left(\frac{d_1+y_0-\mu}{\sigma\sqrt{2}}\right)-\text{erf}\left(\frac{y_0-\mu}{\sigma\sqrt{2}}\right)\right] \\ & \times\left[\left(\frac{y_0-\mu}{\sigma}\right)\exp\left(-\left(\frac{y_0-\mu}{\sigma\sqrt{2}}\right)^2\right)-\left(\frac{d_1+y_0-\mu}{\sigma}\right)\exp\left(-\left(\frac{d_1+y_0-\mu}{\sigma\sqrt{2}}\right)^2\right)\right] \\ & +\frac{\sigma^2}{4}\left[\text{erf}\left(\frac{d_1+y_0-\mu}{\sigma\sqrt{2}}\right)-\text{erf}\left(\frac{y_0-\mu}{\sigma\sqrt{2}}\right)\right]^2 \\ & -\frac{\sigma^2}{2\pi}\left[\exp\left(-\left(\frac{y_0-\mu}{\sigma\sqrt{2}}\right)^2\right)-\exp\left(-\left(\frac{d_1+y_0-\mu}{\sigma\sqrt{2}}\right)^2\right)\right]^2\Bigg\},\end{aligned}$$

and for stages $k \geqslant 2$:

$$\begin{aligned}\Phi_k d_k = \min_{0\leqslant l_k\leqslant d_k}\Bigg\{\sqrt{\Bigg\{} & \frac{\sigma^2}{2\sqrt{2\pi}}\left[\text{erf}\left(\frac{d_k+y_0-\mu}{\sigma\sqrt{2}}\right)-\text{erf}\left(\frac{(d_k-l_k+y_0)-\mu}{\sigma\sqrt{2}}\right)\right] \\ & \times\Bigg[\left(\frac{(d_k-l_k+y_0)-\mu}{\sigma}\right)\exp\left(-\left(\frac{(d_k-l_k+y_0)-\mu}{\sigma\sqrt{2}}\right)^2\right) \\ & -\left(\frac{d_k+y_0-\mu}{\sigma}\right)\exp\left(-\left(\frac{d_k+y_0-\mu}{\sigma\sqrt{2}}\right)^2\right)\Bigg] \\ & +\frac{\sigma^2}{4}\left[\text{erf}\left(\frac{d_k+y_0-\mu}{\sigma\sqrt{2}}\right)-\text{erf}\left(\frac{(d_k-l_k+y_0)-\mu}{\sigma\sqrt{2}}\right)\right]^2 \\ & -\frac{\sigma^2}{2\pi}\left[\exp\left(-\left(\frac{(d_k-l_k+y_0)-\mu}{\sigma\sqrt{2}}\right)^2\right)-\exp\left(-\left(\frac{d_k+y_0-\mu}{\sigma\sqrt{2}}\right)^2\right)\right]^2\Bigg\} \\ & +\Phi_{k-1}(d_k-l_k)\Bigg\}.\end{aligned}$$

Upon substitution of the values of μ, σ, y_0 and d, the OSB ($y_h^* = y_{h-1}^* - l_h^*$) are obtained by executing the `strata.data` or `strata.distr` function.

Table 7 Results for the Normal distribution using `strata.data`.

Stratum (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	13.89	0.10	0.44	0.06	86	488	0.18
2	15.06	0.19	0.11	0.06	81	941	0.09
3	16.01	0.21	0.07	0.06	76	1062	0.07
4	16.97	0.22	0.08	0.06	82	1109	0.07
5	18.14	0.18	0.11	0.06	78	907	0.09
6	22.51	0.10	0.56	0.07	98	493	0.20
Total		1.00	1.37	0.38	500	5000	0.10

8.3. A numerical example for Normal distribution

A study variable dataset following a Normal distribution (herein called *data*), of size $N = 5000$ was simulated to demonstrate the application of the stratifyR package on a Normal population. The data exhibit a Normal distribution with the parameters $mean = 16.010776$ and $sd = 1.662357$. The minimum and maximum values in the simulated data are $[y_0, y_L] = [9.923816, 22.51267]$, which implies that $d = 10.62118$.

To construct the OSB for $h = 6$ (i.e. a 6-strata solution) using the simulated data with a fixed total sample size of 500, the commands below can be used:

```
set.seed(89821)
#simuate random normal variable
data <- rnorm(5000, mean = 16, sd = 1.65)
res <- strata.data(data, h = 6, n=500) #six-strata solution
summary(res) #print results
```

The results output in the R console has information on the best-fit frequency distribution and the fitted parameters, minimum, maximum values of the data, and the distance. Apart from these, the main results are presented in Table 7.

Similarly, in order to find the OSB and other quantities, we can apply the `strata.distr` function to a hypothetical Normal population, which could be based either entirely on assumptions or on prior knowledge regarding the distribution of the variable from past or recent surveys. Let us assume that we have some information from a population with a particular study variable that follows a Normal distribution with given attributes (such as the initial value, distance, parameters, etc.). Say, if the simulated data as presented above was available to us as a recent survey. Then, we can execute the following commands to obtain the characteristics of the study variable and the stratification boundaries (for a 6 strata solution) based entirely on distribution:

```
min(data); max(data); d=max(data)-min(data);d #useful quantities
fit <- fitdist(data, distr="norm", method="mle"); fit
res <- strata.distr(h=6, initval=9.923816, dist=12.58885,
   distr = "norm", params = c(mean=16.010776,
   sd=1.662357), n=500, N=5000) #six-strata solution
summary(res) #print results
```

The results of the stratification for a Normal study variable using `strata.distr` are presented in Table 8.

Table 8 Results for a Normal distribution using `strata.distr`.

Stratum (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	13.89	0.1	0.45	0.068	91	506	0.18
2	15.06	0.18	0.11	0.06	80	909	0.09
3	16.01	0.22	0.08	0.06	79	1087	0.07
4	16.97	0.22	0.08	0.06	79	1086	0.07
5	18.14	0.18	0.11	0.06	80	908	0.09
6	22.51	0.1	0.46	0.068	91	503	0.18
Total		1	1.28	0.376	500	5000	0.1

8.4. Stratification for a survey variable with Gamma distribution

The Gamma distribution is frequently used as a probability model for waiting times; for instance, in life testing, the waiting time until death is a random variable that is usually modelled with a Gamma distribution. It has a moderately skewed profile and due to its versatile nature in fitting a variety of distributions, it is a flexible life distribution model and also useful in risk analysis modelling. Stacy *et al.* (1962) and Chakraborti & Patriarca (2008) showed that it can also be used as a model in a range of disciplines, including climatology and economics, where it can be used for modelling of rainfall and various economic data such as insurance claims or risk, the size of loan defaults, wealth, income, etc.

If the study variable y follows the Gamma distribution (i.e. $y \sim \Gamma(r, \theta)$) on the interval $[y_0, \infty)$, it has the following two-parameter probability density function:

$$f(y; r, \theta) = \frac{1}{\theta^r \Gamma(r)} y^{r-1} e^{-\frac{y}{\theta}}, \quad y > 0;\ r, \theta > 0, \tag{15}$$

where r is a shape parameter and θ is the scale parameter and $\Gamma(r)$ is a Gamma function defined by

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt, \quad r > 0. \tag{16}$$

The function in (16) is also defined by an upper incomplete gamma function $\Gamma(r, x)$ and a lower incomplete gamma function $\gamma(r, x)$, respectively, as follows:

$$\Gamma(r, y) = \int_y^\infty t^{r-1} e^{-t} dt;$$

$$\gamma(r, y) = \int_0^y t^{r-1} e^{-t} dt.$$

There also exist regularised/normalised incomplete Gamma functions which give a value restricted between 0 and 1 and can be stated as:

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

$$P(r, y) = \frac{1}{\Gamma(r)} \int_0^y t^{r-1} e^{-t} dt, \quad r, y > 0;\ \Gamma(r) \neq 0,$$

where $Q(r, y)$ denotes the Upper Regularised Incomplete Gamma function while $P(r, y)$ denotes regularised Lower Incomplete Gamma function (Abramowitz & Stegun 1972). Note that $Q(r, y) = 1 - P(r, y)$.

8.5. DP solution for Gamma distribution

To solve the MPP formulated for Gamma distribution (15), we apply the algorithm using the DP technique discussed in Section 2. The recurrence relations used to determine the OSB are given by:

For the first stage, $k=1$, at $l_1^* = d_1$:

$$\begin{aligned}\Phi_1 d_1 = \sqrt{\left\{\theta^2 r(r+1)\left[Q\left(r, \frac{y_0}{\theta}\right) - Q\left(r, \frac{d_1+y_0}{\theta}\right)\right]\right.} \\ \times \left[Q\left(r+2, \frac{y_0}{\theta}\right) - Q\left(r+2, \frac{d_1+y_0}{\theta}\right)\right] \\ \left. - \theta^2 r^2 \left[Q\left(r+1, \frac{y_0}{\theta}\right) - Q\left(r+1, \frac{d_1+y_0}{\theta}\right)\right]^2\right\},\end{aligned} \tag{17}$$

and for stages $k \geqslant 2$:

$$\begin{aligned}\Phi_k d_k = \min_{0 \leqslant l_k \leqslant d_k} \left\{\sqrt{\left\{\theta^2 r(r+1)\left[Q\left(r, \frac{d_k - l_k + y_0}{\theta}\right)\right.\right.\right.} \\ \left. - Q\left(r, \frac{d_k+y_0}{\theta}\right)\right] \times \left[Q\left(r+2, \frac{d_k - l_k + y_0}{\theta}\right)\right. \\ \left. - Q\left(r+2, \frac{d_k+y_0}{\theta}\right)\right] - \theta^2 r^2 \times \left[Q\left(r+1, \frac{d_k - l_k + y_0}{\theta}\right)\right. \\ \left.\left. - Q\left(r+1, \frac{d_k+y_0}{\theta}\right)\right]^2\right\} + \Phi_{k-1}(d_k - l_k)\right\}.\end{aligned} \tag{18}$$

The recurrence relations (21) and (22) are solved using the DP technique to determine the OSB.

8.6. A numerical example for Gamma distribution

A health data of size $N=724$, derived from the 'National Nutritional Survey' on the 'Micronutrient Status of Women in Fiji', which is provided with the package, is used to demonstrate the application of the stratifyR package on Gamma population. In this example, the variable Folate is used since it exhibits a two-parameter Gamma distribution with the shape and scale parameters as $r=6.9922$ and $\theta=2.5785$ respectively. The minimum and maximum values are $[y_0, y_L]=[4.9, 45.4]$, which implies that $d=40.5$.

To construct the OSB ($h=2$) for the Folate data with a fixed total sample size of 500, we use the following codes and the results are shown in Table 9.

```
data(anaemia) #load the data
folate <- anaemia$Folate #extract 'folate' variable
res <- strata.data(folate, h = 6, n = 500) #6-strata solution
summary(res) #print results
```

Similarly, in order to find the OSB and other stratification results, we can apply the `strata.distr` function to a hypothetical Gamma population. Based on the assumption from past knowledge that the study variable in the population follows Gamma distribution with `initial value=0.5`, `distance=50` and parameters (`shape=3.835768`, `rate=0.340328`). If a fixed sample of $n=500$ is to be selected from a total population

Table 9 Results for the Gamma distribution using `strata.data`.

Stratum (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	11.43	0.17	2.69	0.276	103	122	0.84
2	15.18	0.22	1.32	0.251	93	158	0.59
3	18.84	0.22	1.12	0.237	88	162	0.54
4	23.02	0.18	1.4	0.217	81	133	0.61
5	28.83	0.13	2.74	0.217	81	95	0.85
6	45.4	0.07	23.57	0.362	54	54	1
Total		1.00	32.83	1.560	500	724	0.69

Table 10 Results for the Gamma distribution using `strata.distr`.

Stratum (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	2.8	0.18	0.29	0.099	88	2207	0.04
2	4.17	0.23	0.15	0.09	80	2749	0.03
3	5.59	0.22	0.17	0.089	80	2634	0.03
4	7.3	0.18	0.24	0.089	80	2195	0.04
5	9.81	0.13	0.49	0.09	80	1535	0.05
6	50.5	0.06	3.25	0.102	91	680	0.13
Total		1.00	4.59	0.559	500	12000	0.04

size of $N = 12000$, we can execute the following command to obtain the results as presented in Table 10.

```
#obtain a 6-strata solution of a hypothetical gamma variable
res <- strata.distr(h=6, initval=0.5, dist=50, distr = "gamma",
   params = c(shape=3.835768, rate=0.340328), n=500, N=12000)
summary(res) #print results
```

8.7. Stratification for a survey variable with Lognormal distribution

The Lognormal distribution is a continuous distribution in which the logarithm of a variable has a normal distribution. It is also widely used to describe many natural phenomena. As examples, Limpert, Stahel & Abbt (2001) suggested that phenomena such as milk production by cows, amounts of rainfall and the volume of gas in a petroleum reserve, etc., can all be modelled with a Lognormal distribution. If the study variable y follows the Lognormal distribution on the interval $(0, \infty)$, it has the following two-parameter probability density function:

$$f(y;\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{ln(y)-\mu}{\sigma}\right)^2\right\}, \quad y>0 \tag{19}$$

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

8.8. DP solution for Lognormal distribution

To solve the MPP formulated for a Lognormal distribution (19), we apply the algorithm using the DP technique discussed in Section 2. The recurrence relations used to determine the OSB are given by:

For the first stage, $k=1$, at $l_1^* = d_1$:

$$\Phi_1 d_1 = \sqrt{\left\{ \frac{1}{4}\exp\left(2\mu+2\sigma^2\right)\left[\operatorname{erf}\left(\frac{\log(d_1+y_0)-\mu-2\sigma^2}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log(y_0)-\mu-2\sigma^2}{\sigma\sqrt{2}}\right)\right]\left[\operatorname{erf}\left(\frac{\log(d_1+y_0)-\mu}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log(y_0)-\mu}{\sigma\sqrt{2}}\right)\right] - \frac{1}{4}\exp\left(2\mu+\sigma^2\right)\left[\operatorname{erf}\left(\frac{\log(d_1+y_0)-\mu-\sigma^2}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log(y_0)-\mu-\sigma^2}{\sigma\sqrt{2}}\right)\right]^2 \right\}}, \tag{20}$$

and for stages $k \geqslant 2$:

$$\Phi_k d_k = \min_{0\leqslant l_k \leqslant d_k} \left\{ \sqrt{\left\{ \frac{1}{4}\exp\left(2\mu+2\sigma^2\right)\left[\operatorname{erf}\left(\frac{\log(d_k+y_0)-\mu-2\sigma^2}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log(d_k-l_k+y_0)-\mu-2\sigma^2}{\sigma\sqrt{2}}\right)\right]\left[\operatorname{erf}\left(\frac{\log(d_k+y_0)-\mu}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log(d_k-l_k+y_0)-\mu}{\sigma\sqrt{2}}\right)\right] - \frac{1}{4}\exp\left(2\mu+\sigma^2\right) \times \left[\operatorname{erf}\left(\frac{\log(d_k+y_0)-\mu-\sigma^2}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log(d_k-l_k+y_0)-\mu-\sigma^2}{\sigma\sqrt{2}}\right)\right]^2 \right\} + \Phi_{k-1}(d_k-l_k) \right\}. \tag{21}$$

Upon substitution of the values of μ, σ, y_0 and d, the OSB ($y_h^* = y_{h-1}^* - l_h^*$) are obtained by executing the `strata.data` or `strata.distr` function.

8.9. A numerical example for Lognormal distribution

The `hies` data of size $N = 3566$ is used to demonstrate the application of the stratifyR package on a Lognormal population. The `hies` data, which accompanies the package, comes from the HIES survey conducted in Fiji in the year 2010. The data contains only two aspects of the survey, namely Income and Expenditure. In this example, the variable Expenditure is used since it exhibits a 2-parameter Lognormal distribution with the shape and scale parameters as `meanlog = 9.2804934` and `sdlog = 0.6917842` respectively. The minimum and maximum values are $[y_0, y_L] = [991.24, 136539.1]$, which implies that $d = 135547.8$.

To construct the OSB ($h=6$) for the Expenditure data with a fixed total sample size of 500, we use the following codes. Table 11 presents the stratification results.

```
data(hies) #load data
Expenditure <- hies$Expenditure #extract main variable
res <- strata.data(Expenditure, h = 6, n=500) #6-strata solution
summary(res) #print results
```

If data of the study variable are not available but can be assumed to follow a Lognormal distribution, we can apply the `strata.distr` function to obtain the OSB and other stratification results as presented in Table 12 by executing the following commands:

Table 11 Results for the Lognormal distribution using `strata.data`.

Strata (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	7033.13	0.28	1771772.82	374.763	83	1004	0.08
2	11491.13	0.28	1632728.76	362.982	80	1013	0.08
3	17143.84	0.2	2716245.8	336.923	75	729	0.1
4	25531.44	0.12	5296567.94	274.932	61	426	0.14
5	41367.24	0.08	20652444.55	341.538	76	268	0.28
6	136539.06	0.04	462208769	759.641	126	126	1
Total		1	494278529	2450.779	500	3566	0.14

Table 12 Results for Lognormal distribution using `strata.distr`.

Strata (h)	OSB (y_i)	W_h	V_h	W_hS_h	n_h	N_h	f_h
1	7033.13	0.27	1978332.72	380.588	90	966	0.09
2	11491.13	0.27	1628962.51	343.018	81	958	0.08
3	17142.48	0.21	2602729.74	341.03	81	754	0.11
4	25530.07	0.14	5612494.87	341.053	81	513	0.16
5	41364.5	0.08	18607532.55	342.905	81	283	0.29
6	136539.04	0.03	209845971.7	368.083	87	91	0.96
Total		1	240276024	2116.677	500	3566	0.14

```
#get important measures
length(Expenditure); d=max(Expenditure)-min(Expenditure);d
#obtain parameter estimates
fit <- fitdist(Expenditure, distr="lnorm", method="mle"); fit
#6-strata solution for a hypothetical lognormal study variable
res <- strata.distr(h=6, initval=10, dist=188, distr = "lnorm",
  params = c(meanlog=3.23, sdlog=0.65), n=500, N=3566)
summary(res) #print results
```

References

ABRAMOWITZ, M. & STEGUN, I.A. (1972). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

ARNOLD, B.C. (2015). *Pareto Distributions*. London: Chapman and Hall/CRC.

ATKINSON, A.B. & HARRISON, A.J. (1978). *Distribution of Personal Wealth in Britain*. Cambridge: Cambridge University Press.

BRYSON, M.C. (1974). Heavy-tailed distributions: Properties and tests. *Technometrics* **16**, 61–68.

BÜHLER, W. & DEUTLER, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika* **22**, 161–175.10.1007/BF01899725.

CARNELL, R. (2017). *triangle: Provides the Standard Distribution Functions for the Triangle Distribution*. https://cran.r-project.org/package=triangle.

CHAKRABORTI, A. & PATRIARCA, M. (2008). Gamma-distribution and wealth inequality. *Pramana* **71**, 233–243.

COCHRAN, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute* **38**, 345–358.

CORBELLINI, A., CROSATO, L., GANUGI, P. & MAZZOLI, M. (2010). Fitting Pareto II Distributions on Firm Size: Statistical Methodology and Economic Puzzles. In *Advances in Data Analysis*. Boston: Birkhäuser Boston, pp. 321–328. 10.1007/978-0-8176-4799-5_26.

DALENIUS, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal* **1950**, 203–213.

DALENIUS, T. (1957). *Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice*. Stockholm: Almqvist and Wiksell.

DELIGNETTE-MULLER, M.L., DUTANG, C. *et al.* (2015). fitdistrplus: An R Package for fitting distributions. *Journal of Statistical Software* **64**, 1–34.

DUTANG, C., GOULET, V., PIGEON, M. *et al.* (2008). actuar: An R Package for actuarial science. *Journal of Statistical Software* **25**, 1–37.

EVERT, S. & BARONI, M. (2007). *zipfR*: Word Frequency Distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*. Prague, Czech Republic, pp. 29–32. (R package version 0.6-10 of 2017-08-17).

GUNNING, P. & HORGAN, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology* **30**, 159–166.

HASSAN, A.S. & AL-GHAMDI, A.S. (2009). Optimum step stress accelerated life testing for lomax distribution. *Journal of Applied Sciences Research* **5**, 2153–2164.

KHAN, E.A., KHAN, M.G.M. & AHSAN, M.J. (2002). Optimum stratification: A mathematical programming approach. *Calcutta Statistical Association Bulletin* **52**, 323–333.

KHAN, M.G.M., NAND, N. & AHMAD, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology* **34**, 205–214.

KHAN, M.G.M., REDDY, K.G. & RAO, D.K. (2015). Designing stratified sampling in economic and business surveys. *Journal of Applied Statistics* **42**, 2080–2099.

LEWIS, D. (1957). Normal distribution of intelligence: A critique. *British Journal of Psychology* **48**, 98–104.

LIMPERT, E., STAHEL, W.A. & ABBT, M. (2001). Log-normal Distributions Across the Sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. *BioScience* **51**, 341–352.

LOHR, S. (2009). *Sampling: Design and Analysis*. Toronto: Nelson Education Ltd.

LOMAX, K. (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association* **49**, 847–852.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–625.

POUILLOT, R. & DELIGNETTE-MULLER, M.L. (2010). Evaluating variability and uncertainty in microbial quantitative risk assessment using two R Packages. *International Journal of Food Microbiology* **142**, 330–40.

REDDY, K.G. & KHAN, M. (2019). Optimal stratification in stratified designs using weibull-distributed auxiliary information. *Communications in Statistics-Theory and Methods*, **48**, 3136–3152.

REDDY, K.G., KHAN, M.G. & KHAN, S. (2018). Optimum strata boundaries and sample sizes in health surveys using auxiliary variables. *PloS One* **13**, 1–34.

RICHARD, B. (1957). *Dynamic Programming*. Princeton: Princeton University Press.

RIVEST, L.P. & BAILLARGEON, S. (2017). *stratification: Univariate Stratification of Survey Populations*. https://CRAN.R-project.org/package=stratification. R package version 2.2-6.

SÄRDNAL, C.E., SWENSSON, B. & WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. Berlin: Springer-Verlag.

STACY, E.W. *et al.* (1962). A generalization of the Gamma distribution. *The Annals of Mathematical Statistics* **33**, 1187–1192.

VENABLES, W.N. & RIPLEY, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.