

## Optimal stratification in stratified designs using weibull-distributed auxiliary information

Karuna G. Reddy & M.G.M. Khan

To cite this article: Karuna G. Reddy & M.G.M. Khan (2018): Optimal stratification in stratified designs using weibull-distributed auxiliary information, Communications in Statistics - Theory and Methods, DOI: [10.1080/03610926.2018.1473609](https://doi.org/10.1080/03610926.2018.1473609)

To link to this article: <https://doi.org/10.1080/03610926.2018.1473609>



Published online: 16 Dec 2018.



Submit your article to this journal [↗](#)



Article views: 25



View Crossmark data [↗](#)



# Optimal stratification in stratified designs using weibull-distributed auxiliary information

Karuna G. Reddy<sup>a\*</sup> and M. G. M. Khan<sup>b</sup>

<sup>a</sup>Office of Deputy Vice-Chancellor, Research, International & Innovation, The University of South Pacific, Suva, Fiji;

<sup>b</sup>School of Computing, Information and Mathematical Sciences, The University of South Pacific, Suva, Fiji

## ABSTRACT

Sampling has evolved into a universally accepted approach for gathering information and data mining as it is widely accepted that a reasonably modest-sized sample can sufficiently characterize a much larger population. In stratified sampling designs, the whole population is divided into homogeneous strata in order to achieve higher precision in the estimation. This paper proposes an efficient method of constructing optimum stratum boundaries (OSB) and determining optimum sample size (OSS) for the survey variable. The survey variable may not be available in practice since the variable of interest is unavailable prior to conducting the survey. Thus, the method is based on the auxiliary variable which is usually readily available from past surveys. To illustrate the application as an example using a real data, the auxiliary variable considered for this problem follows Weibull distribution. The stratification problem is formulated as a Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter under Neyman allocation. The solution procedure employs the dynamic programming technique, which results in substantial gains in the precision of the estimates of the population characteristics.

## ARTICLE HISTORY

Received 1 November 2017

Accepted 27 April 2018

## KEYWORDS

Optimal stratification;  
mathematical programming  
problem; dynamic  
programming technique;  
stratified random sampling;  
Weibull distribution

## 1. Introduction

Businesses, organizations and government departments that rely on analytics to understand the population have been employing the sampling phenomenon for decades. Extracting data from a database for the purpose of data mining and statistical analyses is based on the sampling techniques routinely used in surveys (Christopher and Blaxton 1998). For example, Stratified sampling is an important sampling technique used in health surveys to estimating the prevalence of diseases, diabetes, anemia, obesity hypertension, smoking and in many other parameter estimations. It is also a common phenomenon in the disciplines of business and sciences.

In stratified sampling, the sampling-frame is divided into non overlapping groups or strata in such a way that the strata constructed are internally homogeneous with respect to the main study variable that maximizes the precision of its estimate. More often the surveyors stratify the population in most convenient manners such as the use of geographical/administrative regions, provinces, districts, etc.) or other natural criteria such as gender and age. However,

**CONTACT** Karuna G. Reddy  [reddy\\_k@usp.ac.fj](mailto:reddy_k@usp.ac.fj)  University of the South Pacific, Suva, Fiji.

\*ANU Centre for Social Research and Methods, Australian National University.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/lssta](http://www.tandfonline.com/lssta).

© 2018 Taylor & Francis Group, LLC

stratification by convenience manner is not always a reasonable criterion as the strata so obtained may not be internally homogeneous with respect to a variable of interest. Thus, one has to look for the OSB that maximizes the precision of the estimates.

The problem of determining OSB for a variable, when its frequency distribution is known, is well known in the sampling literature. In order to achieve maximum precision in determining OSB, the stratum variances  $\sigma_h^2$  should be as small as possible. When a single variable is of interest and the stratification is made based on this study variable, then the OSB can be determined by cutting the range of its known distribution at suitable points. This problem of determining the OSB was first discussed by Dalenius (1950). He presented a set of minimal equations which are usually difficult to solve because of their implicit nature. When the frequency distribution of the auxiliary variable,  $x$ , is known, several approximation methods of determining OSB using the auxiliary variable have also been suggested and discussed by many authors such as Sethi (1963), Dalenius (1957), Dalenius and Hodges (1959), Taga (1967), Serfling (1968), Singh and Sukhatme (1969, 1972, 1973), Singh (1971), Singh and Dev Prakash (1975), Cochran (1977), Mehta, Singh, and Kishore (1996), Rizvi, Gupta, and Bhargava (2002), Gupta, Singh, and Mahajan (2005), Jurina and Gligorova (2017), Danish et al. (2017), Khan, Reddy, and Rao (2015), and Hidirolou and Kozak (2017).

Attempts have also been made to determine the global OSB by many authors, such as, Lavalée and Hidirolou (1988) who proposed an algorithm to construct stratum boundaries for a power allocated stratified sample. Later, Hidirolou and Srinath (1993) presented a more general form of the algorithm. Lavallée and Hidirolou's algorithm was reviewed by Sweet and Sigman (1995) and Rivest (2002) and they proposed a modified algorithm that incorporates the different relationships between the stratification and study variables. There are several other algorithms available in the literature, for example, Niemi (1999) proposed a random search method and Nicolini (2001) suggested Natural Class Method. Later on, Kozak (2004) presented a modified random search algorithm while Gunning and Horgan (2004) proposed an alternative method to approximate stratification based on a geometric progression. Horgan (2006) compared this approach with Dalenius and Hodges (1959), Ekman (1959), and Lavalée and Hidirolou (1988) and confirmed that the geometric progression method is more efficient. However, Kozak and Verma (2006) studied the usefulness of Gunning and Horgan's geometric progression method and found out a different result that the geometric progression approach is less efficient than Lavallée and Hidirolou's algorithm (see Kozak, Verma, and Zielinski 2007).

Another kind of stratification method that has been proposed in the literature is due to Bühler and Deutler (1975) and later by Khan, Khan, and Ahsan (2002), Khan (2005), Khan, Ahmad, and Sabiha (2009), Khan et al. (2015), Nand and Khan (2009), Khan and Sushita (2015), Khan, Reddy, and Rao (2015), and Reddy, Khan, and Rao (2016) who formulated the problems of determining OSB as optimization problems, which are solved by developing computational techniques using dynamic programming. Bühler and Deutler's approach was also used by Lavallée (1988) and Lavalée and Hidirolou (1988) for determining the OSB which would divide the population domain of two stratification variables into distinct subsets such that the precision of the variables of interest is maximized.

This paper proposes a procedure of determining OSB and OSS for each stratum for the purpose of data mining a variable of interest. Since stratification based on the survey variable ( $y$ ), is not feasible in practice (as the variable is unavailable prior to conducting the survey), the optimum stratification is made based on an auxiliary variable ( $x$ ), if  $y$  holds a regression

model (see Yong et al. 2016). The problem of determining OSB is formulated as an MPP that seeks minimization of the variance of the estimated population parameter under Neyman allocation (see Neyman 1934; De Gruijter, Minasny, and Mcbratney 2015). The formulated MPP, being a multistage decision problem, is solved using dynamic programming technique. The data set is obtained from a national nutrition survey aiming to estimate the mean of the study variable “hemoglobin” using the auxiliary variable “iron”.

## 2. General formulation of the problem

Let the population be stratified into  $L$  strata based on an auxiliary variable  $x$ , when the estimation of the mean of a study variable  $y$  is of interest. If a simple random sample of size  $n_h$  is to be drawn from  $h$ th stratum with sample mean  $\bar{y}_h$ ; ( $h = 1, 2, \dots, L$ ), then the stratified sample mean,  $\bar{y}_{st}$ , is given by

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

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

$$\text{Var}(\bar{y}_{st}) = \frac{\left( \sum_{h=1}^L W_h \sigma_{hy} \right)^2}{n} \quad (2)$$

where  $W_h$  and  $\sigma_{hy}^2$  are the stratum weight and the stratum variance in  $h^{th}$  stratum;  $h = 1, 2, \dots, L$  respectively and  $n$  is the preassigned total sample size.

Consider that the study variable has the regression model of the form:

$$y = \lambda(x) + \epsilon \quad (3)$$

where  $\lambda(x)$  is a linear or a nonlinear function of  $x$  and  $\epsilon$  is an error term such that  $E(\epsilon|x) = 0$  and  $V(\epsilon|x) = \phi(x) > 0$  for all  $x$ .

Under model Equation (3), the stratum mean  $\mu_{hy}$  and the stratum variance  $\sigma_{hy}^2$  of  $y$  can be expressed as (see Singh and Sukhatme 1969):

$$\mu_{hy} = \mu_{h\lambda} \quad (4)$$

$$\text{and } \sigma_{hy}^2 = \sigma_{h\lambda}^2 + \mu_{h\phi} \quad (5)$$

where  $\mu_{h\lambda}$  and  $\mu_{h\phi}$  are the expected values of functions  $\lambda(x)$  and  $\phi(x)$ , respectively, and  $\sigma_{h\lambda}^2$  denotes the variance of  $\lambda(x)$  in the  $h$ th stratum.

If  $\lambda$  and  $\epsilon$  are uncorrelated, from model Equation (3),  $\sigma_{hy}^2$  can also be expressed as (see Dalenius and Gurney 1951):

$$\sigma_{hy}^2 = \sigma_{h\lambda}^2 + \sigma_{h\epsilon}^2 \quad (6)$$

where  $\sigma_{h\epsilon}^2$  is the variance of  $\epsilon$  in the  $h$ th stratum. It can be verified that the expressions Equations (5) and (6) are equivalent.

Let  $f(x)$ ;  $a \leq x \leq b$  be the frequency function of the auxiliary variable  $x$  that is used for the stratification. If the population mean of the study variable  $y$  is estimated under (1), then the problem of determining the strata boundaries is to cut up the range,  $d = b - a$ , at  $(L - 1)$

intermediate points  $a = x_0 \leq x_1 \leq x_2 \leq \dots \leq x_{L-1} \leq x_L = b$  such that Equation (3) is minimum.

For a fixed sample size  $n$ , minimizing the expression of the right-hand side of Equation (2) is equivalent to minimizing  $\sum_{h=1}^L W_h \sigma_{hy}$ . Thus, from Equation (6), we minimize

$$\sum_{h=1}^L W_h \sqrt{\sigma_{h\lambda}^2 + \mu_{h\phi}} \quad (7)$$

If  $f(x)$ ,  $\lambda(x)$  and  $\phi(x)$  are known and integrable, then the quantities  $W_h$ ,  $\sigma_{h\lambda}^2$  and  $\mu_{h\phi}$  can be obtained as a function of the boundary points  $x_h$  and  $x_{h-1}$  by using the following expressions:

$$W_h = \int_{x_{h-1}}^{x_h} f(x) dx \quad (8)$$

$$\sigma_{h\lambda}^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \lambda^2(x) f(x) dx - \mu_{h\lambda}^2 \quad (9)$$

and

$$\mu_{h\phi} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \phi(x) f(x) dx \quad (10)$$

where

$$\mu_{h\lambda} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \lambda(x) f(x) dx \quad (11)$$

and  $(x_{h-1}, x_h)$  are the boundaries of the  $h^{th}$  stratum.

Thus, the objective function in Equation (7) could be expressed as a function of boundary points  $(x_h, x_{h-1})$  only.

Let  $\phi_h(x_h, x_{h-1}) = W_h \sigma_{hy} = W_h \sqrt{\sigma_{h\lambda}^2 + \mu_{h\phi}}$ . Then, the problem of determination of OSB can be expressed as the following optimization problem: Find  $x_1, x_2, \dots, x_L$  that

$$\begin{aligned} & \text{Minimize} \quad \sum_{h=1}^L \phi_h(x_h, x_{h-1}) \\ & \text{subject to} \quad a = x_0 \leq x_1 \leq x_2 \leq \dots \leq x_{L-1} \leq x_L = b \end{aligned} \quad (12)$$

We further define

$$l_h = x_h - x_{h-1}; \quad h = 1, 2, \dots, L \quad (13)$$

where  $l_h \geq 0$  denotes the range or width of the  $h^{th}$  stratum.

Obviously, with this definition of  $l_h$ , the range of the distribution,  $d = b - a$ , is expressed as a function of stratum width as

$$\sum_{h=1}^L l_h = \sum_{h=1}^L (x_h - x_{h-1}) = b - a = x_L - x_0 = d \quad (14)$$

The  $h^{th}$  stratification point  $x_h$ ;  $h = 1, 2, \dots, L$  is then expressed as

$$\begin{aligned} x_h &= x_0 + \sum_{i=1}^h l_i \\ \text{or, } x_h &= x_{h-1} + l_h \end{aligned}$$

Adding Equation (14) as a constraint, the problem (12) can be treated as an equivalent problem of determining optimum strata widths (OSW),  $l_1, l_2, \dots, l_L$ , and is expressed as the following MPP:

$$\begin{aligned} & \text{Minimize} \quad \sum_{h=1}^L \phi_h(l_h, x_{h-1}) \\ & \text{subject to} \quad \sum_{h=1}^L l_h = d \\ & \text{and} \quad l_h \geq 0; h = 1, 2, \dots, L \end{aligned} \quad (15)$$

Initially,  $x_0$  is known. Therefore, the first term, that is,  $\phi_1(l_1, x_0)$  in the objective function of the MPP Equation (15) is a function of  $l_1$  alone. Once  $l_1$  is known, the second term  $\phi_2(l_2, x_1)$  will become a function of  $l_2$  alone and so on. Due to the special nature of functions, the MPP Equation (15) may be treated as a function of  $l_h$  alone and can be expressed as

$$\begin{aligned} & \text{Minimize} \quad \sum_{h=1}^L \phi_h(l_h) \\ & \text{subject to} \quad \sum_{h=1}^L l_h = d \\ & \text{and} \quad l_h \geq 0; h = 1, 2, \dots, L \end{aligned} \quad (16)$$

### 3. The solution procedure using dynamic programming technique

The problem Equation (16) is a multi-stage decision problem in which the objective function and the constraint are separable functions of  $l_h$ , which allows us to use a dynamic programming technique (see Khan, Nand, and Ahmad 2008). Dynamic programming determines the optimum solution of a multi-variable problem by decomposing it into stages, each stage compromising a single variable sub-problem. A dynamic programming model is basically a recursive equation based on Bellman's principle of optimality (see Bellman 1957). This recursive equation links the different stages of the problem in a manner which guarantees that each stage's optimal feasible solution is also optimal and feasible for the entire problem (see Taha 2007, Chapter 10).

Consider the following subproblem of Equation (16) for first  $k (< L)$  strata:

$$\begin{aligned} & \text{Minimize} \quad \sum_{h=1}^k \phi_h(l_h) \\ & \text{subject to} \quad \sum_{h=1}^k l_h = d_k \\ & \text{and} \quad l_h \geq 0; h = 1, 2, \dots, k \end{aligned} \quad (17)$$

where  $d_k < d$  is the total width available for division into  $k$  strata or the state value at stage  $k$ . Note that  $d_k = d$  for  $k = L$ .

The transformation functions are given by

$$d_k = l_1 + l_2 + \dots + l_k$$

$$\begin{aligned}
d_{k-1} &= l_1 + l_2 + \cdots + l_{k-1} = d_k - l_k \\
d_{k-2} &= l_1 + l_2 + \cdots + l_{k-2} = d_{k-1} - l_{k-1} \\
&\vdots \quad \quad \quad \vdots \\
d_2 &= l_1 + l_2 = d_3 - l_3 \\
d_1 &= l_1 = d_2 - l_2
\end{aligned}$$

Let  $\Phi_k(d_k)$  denote the minimum value of the objective function of Equation (17), that is,

$$\begin{aligned}
\Phi_k(d_k) = \min \left[ \sum_{h=1}^k \phi_h(l_h) \right] &\left| \sum_{h=1}^k l_h = d_k, \text{ and } l_h \geq 0; \right. \\
&\left. h = 1, 2, \dots, k \text{ and } 1 \leq k \leq L \right]
\end{aligned} \quad (18)$$

With the above definition of  $\Phi_k(d_k)$ , the MPP Equation (16) is equivalent to finding  $\Phi_L(d)$  recursively by finding  $\Phi_k(d_k)$  for  $k = 1, 2, \dots, L$  and  $0 \leq d_k \leq d$ .

We can write

$$\begin{aligned}
\Phi_k(d_k) = \min \left[ \phi_k(l_k) + \sum_{h=1}^{k-1} \phi_h(l_h) \right] &\left| \sum_{h=1}^{k-1} l_h = d_k - l_k, \right. \\
&\left. \text{and } l_h \geq 0; h = 1, 2, \dots, k \right]
\end{aligned} \quad (19)$$

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

$$\begin{aligned}
\Phi_k(d_k) = \phi_k(l_k) + \min \left[ \sum_{h=1}^{k-1} \phi_h(l_h) \right] &\left| \sum_{h=1}^{k-1} l_h = d_k - l_k \right. \\
&\left. \text{and } l_h \geq 0; h = 1, 2, \dots, k-1 \text{ and } \right. \\
&\left. 1 \leq k \leq L \right]
\end{aligned} \quad (20)$$

Using Bellman's principle of optimality, we write a forward recursive equation of the dynamic programming technique as

$$\Phi_k(d_k) = \min_{0 \leq l_k \leq d_k} [\phi_k(l_k) + \Phi_{k-1}(d_k - l_k)], \quad k \geq 2 \quad (21)$$

For the first stage, that is, for  $k = 1$ :

$$\Phi_1(d_1) = \phi_1(d_1) \implies l_1^* = d_1 \quad (22)$$

where  $l_1^* = d_1$  is the optimum width of the first stratum. The relations Equations (21) and (22) are solved recursively for each  $k = 1, 2, \dots, L$  and  $0 \leq d_k \leq d$ , and  $\Phi_L(d)$  is obtained. From  $\Phi_L(d)$  the optimum width of  $L^{\text{th}}$  stratum,  $l_L^*$ , is obtained. From  $\Phi_{L-1}(d - l_L^*)$  the optimum width of  $(L-1)^{\text{th}}$  stratum,  $l_{L-1}^*$ , is obtained and so on until  $l_1^*$ , optimum width of 1<sup>st</sup> stratum, is obtained.

#### 4. Constructing OSB for Weibull auxiliary variable

The Weibull distribution is a two-parameter family of continuous probability distributions. Because of its versatility in the fitting of a variety of distributions, it is one of the most widely used distributions in applied statistics, especially in survival analysis, mortality or failure analysis, reliability, engineering to model manufacturing and delivery times, in extreme value

**Table 1.** ANOVA for regression model.

Source	SS	df	MS	<i>f</i>	<i>p</i> -Value
Regression	461.92	1	461.92	299.95	0.000
Residual	1050.61	682	1.54		
Lack of fit	236.40	204	1.16	0.68	0.890
Pure error	814.21	478	1.70		
Total	1512.54	683			

**Table 2.** Summary of model parameters.

Predictor	Coefficient	SE coef	<i>t</i>	<i>p</i> -Value
$\alpha$	10.9449	0.1245	87.89	0.000
$\beta$	0.114115	0.009548	11.95	0.000

theory and weather forecasting. Due to its moderately skewed profile, it also characterizes well a wide range of health data, including health monitoring data, Epidemiological data such as episode durations of depression and gene expressions data (see Patten 2006; Wahed, Luong, and Jeong 2009; and Wang et al. 2011).

If an auxiliary variable  $x$  follows the Weibull distribution on the interval  $[x_0, x_L]$ , its two-parameter probability density function with a state space  $x \geq 0$  is given by:

$$f(x; \theta, r) = \frac{r}{\theta} \left( \frac{x}{\theta} \right)^{r-1} e^{-(x/\theta)^r}, \quad x \geq 0 \quad (23)$$

where  $r > 0$  is the shape parameter and  $\theta > 0$  is the scale parameter of the distribution.

The shape parameter gives the Weibull distribution its flexibility. By changing the value of the shape parameter, the distribution can model a wide variety of data that follows the exponential distribution, the Rayleigh distribution, the normal distribution or even the approximate log-normal distribution.

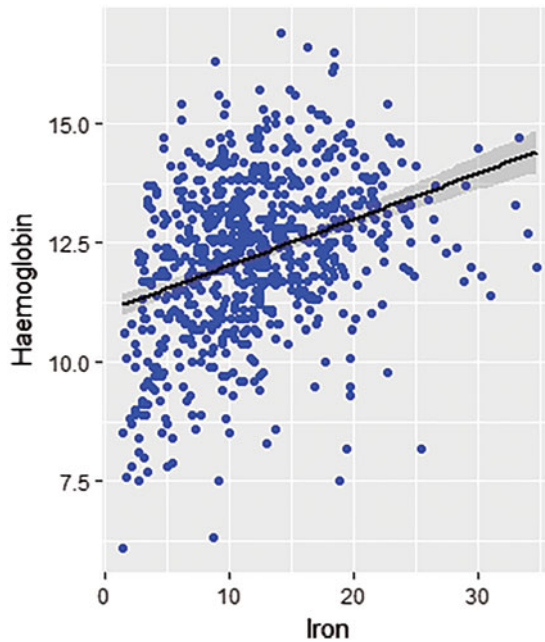
#### 4.1. Estimating the linear regression model

To illustrate the formulation of the problem of determining OSB as an MPP for a population with Weibull auxiliary variable, we use a set of health data of size  $N = 724$  obtained from 2004 Fiji National Nutrition Survey on “Micronutrient Status of Women in Fiji.” The data in this problem have the characteristics: Level of Iron and Level of Hemoglobin for each woman.

Suppose that a survey on iron deficiency anaemia is to be conducted in a country, where a sample will be collected using stratified random sampling and hemoglobin ( $y$ ) will be the variable of interest. That is, the hemoglobin will be the main stratification variable. Then, the level of iron collected in some previous study may be a reasonable choice for an auxiliary variable ( $x$ ).

To estimate the hemoglobin content ( $y$ ) in women, we fit a regression model given in Equation (5) for the survey mentioned above. We observed that the data significantly fit a linear regression model with iron level ( $x$ ). Table 1 presents the analysis of variance (ANOVA) and Table 2 depicts the summary of the estimates of the model parameters. From these tables, the computational results reveal that the fitted regression model and estimated parameters are highly significant with  $p$ -value  $< 0.001$ .

The coefficient of determination or correlation coefficient,  $R^2 = \frac{SSR}{SST}$ , with a value of 30.54% obtained from Table 1 indicates a moderate strength of linear relationship between



**Figure 1.** Scatterplot of iron vs hemoglobin.

the two variables.  $R^2$  is found to be one of the highest for the linear model when compared with the model summary of all the other non linear models available in statistical package. Table 1 also reveals that there is no significant lack of fit in the linear regression with  $p$ -value = 0.890. Thus, the model fits the data well and gives us no reason to consider an alternative model.

Figure 1 depicts the linear association through the scatterplot for the Iron versus the Hemoglobin. It indicates a moderately positive linear relationship.

Therefore, the hemoglobin content ( $y$ ) and the iron level ( $x$ ) are fairly assumed to follow a linear regression model with the following equation

$$\lambda(x) = \alpha + \beta x \quad (24)$$

and the least-squares estimates of the parameters are given by

$$\hat{\alpha} = 10.9449 \quad \text{and} \quad \hat{\beta} = 0.1141 \quad (25)$$

#### 4.2. Estimating the distribution

To determine the distribution,  $f(x)$ , for the auxiliary variable, we construct a relative frequency histogram of iron level ( $x$ ). Figure 2 shows that the distribution of  $x$  is a right skewed distribution that matches the Weibull distribution.

The probability plot ( $q-q$ ) of  $x$  was also obtained to determine whether the distribution of  $x$  matches Weibull distribution. Figure 3 reveals that the points are clustered around the straight line, thus,  $x$  is assumed to follow Weibull distribution with a probability density function given by Equation (23).

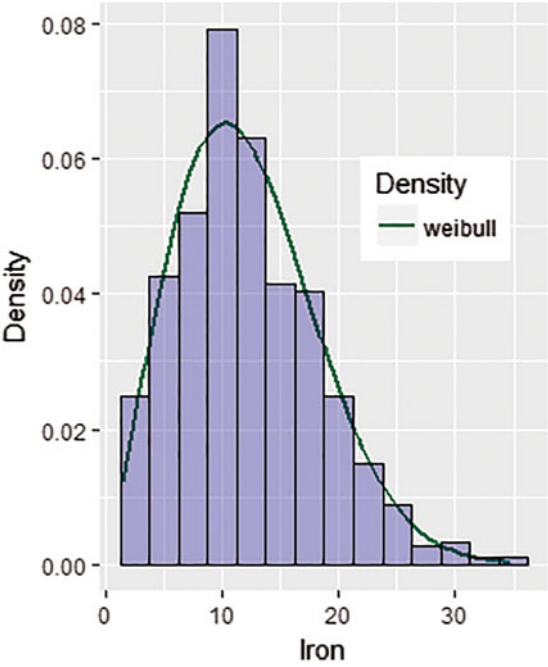


Figure 2. Frequency histogram of the iron level ( $x$ ).

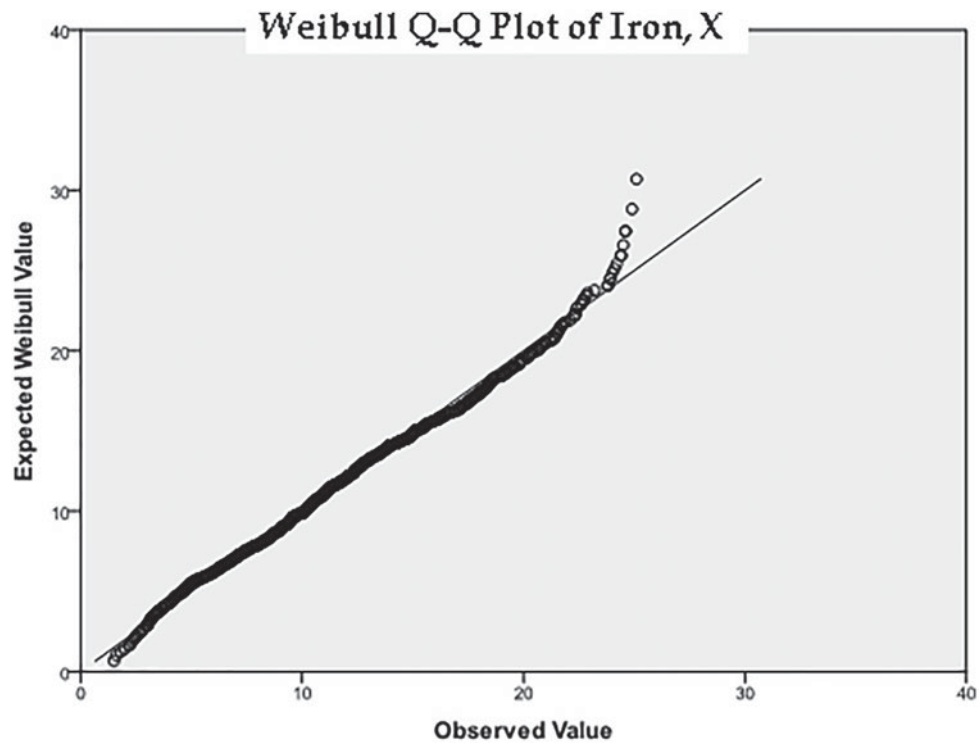


Figure 3. Weibull Q–Q plot of the iron level ( $x$ ).

The maximum likelihood estimate (MLE) of the parameters for Weibull distribution is found to be

$$\text{Shape}, r = 2.34318488 \quad \text{and} \quad \text{Scale}, \theta = 13.40282496$$

Using the Kolmogorov-Smirnov test, the maximum difference between the observed distribution and the Weibull distribution is found to be non significant ( $D = 0.0328$  and  $p\text{-value} = 0.452$ ), which supports that  $x$  follows Weibull distribution with the indicated parameters.

### 4.3. Formulating the problem of OSB as an MPP

Let the auxiliary variable  $x$  follow Weibull distribution (i.e.,  $x \sim W(r, \theta)$ ) with density function given by Equation (23). By using Equations (8), (9), (11) and (23), the quantities  $W_h$ ,  $\mu_{h\lambda}$ , and  $\sigma_{h\lambda}^2$  can be obtained as a function of boundary points  $(x_{h-1}, x_h)$  as follows:

$$W_h = -e^{-\left(\frac{x_h}{\theta}\right)^r} - \left(-e^{-\left(\frac{x_{h-1}}{\theta}\right)^r}\right) \quad (26)$$

Using Equation (14), that is, substituting  $x_h = x_{h-1} + l_h$  in Equation (25),  $W_h$  is obtained as:

$$W_h = \left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1}+l_h}{\theta}\right)^r} \right] \quad (27)$$

$\mu_{h\lambda}$  can be expressed as

$$\mu_{h\lambda} = \alpha + \frac{\beta\theta}{W_h} \left[ \int_{\left(\frac{x_{h-1}}{\theta}\right)^r}^{\infty} t^{\frac{1}{r}} e^{-t} dt - \int_{\left(\frac{x_h}{\theta}\right)^r}^{\infty} t^{\frac{1}{r}} e^{-t} dt \right] \quad (28)$$

Let  $\Gamma(r, x)$  and  $Q(r, s)$  denote the upper incomplete gamma function and the regularized/normalized incomplete gamma function, respectively, given by

$$\Gamma(r, x) = \int_x^{\infty} t^{r-1} e^{-t} dt \quad (29)$$

$$Q(r, x) = \frac{1}{\Gamma(r)} \int_x^{\infty} t^{r-1} e^{-t} dt, \quad r, x > 0; \quad \Gamma(r) \neq 0 \quad (30)$$

Then, using Equations (28) and (29),  $\mu_{h\lambda}$  given in Equation (27) is derived to be

$$\begin{aligned} \mu_{h\lambda} &= \alpha + \frac{\beta\theta \Gamma\left(1 + \frac{1}{r}\right)}{W_h} \left[ Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(1 + \frac{1}{r}, \left(\frac{x_h}{\theta}\right)^r\right) \right] \\ &= \alpha + \frac{\beta\theta \Gamma\left(1 + \frac{1}{r}\right)}{\left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_h}{\theta}\right)^r} \right]} \left[ Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(1 + \frac{1}{r}, \left(\frac{x_h}{\theta}\right)^r\right) \right] \\ &= \alpha + \frac{\beta\theta \Gamma\left(1 + \frac{1}{r}\right) \left[ Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1}+l_h}{\theta}\right)^r\right) \right]}{\left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1}+l_h}{\theta}\right)^r} \right]} \quad (31) \end{aligned}$$

Similarly, the quantity  $\sigma_{h\lambda}^2$  is reduced to

$$\sigma_{h\lambda}^2 = \frac{\beta^2 \theta^2 \Gamma\left(1 + \frac{2}{r}\right) \left[ Q\left(1 + \frac{2}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(1 + \frac{2}{r}, \left(\frac{x_{h-1} + l_h}{\theta}\right)^r\right) \right]}{\left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_h}{\theta}\right)^r} \right]} - \frac{\beta^2 \theta^2 \Gamma^2\left(1 + \frac{1}{r}\right) \left[ Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1} + l_h}{\theta}\right)^r\right) \right]^2}{\left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1} + l_h}{\theta}\right)^r} \right]^2} \quad (32)$$

Then, the formulated MPP given in Equation (17) could be expressed using Equations (8), (26) and (31) as:

$$\begin{aligned} \text{Minimize} \quad & \sum_{h=1}^L \left\{ \text{Sqrt} \left\{ \beta^2 \theta^2 \Gamma\left(1 + \frac{2}{r}\right) \left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1} + l_h}{\theta}\right)^r} \right] \right. \right. \\ & \times \left[ Q\left(1 + \frac{2}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(1 + \frac{2}{r}, \left(\frac{x_{h-1} + l_h}{\theta}\right)^r\right) \right] \\ & - \beta^2 \theta^2 \left[ \Gamma\left(1 + \frac{1}{r}\right) \left[ Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1}}{\theta}\right)^r\right) \right. \right. \\ & \left. \left. - Q\left(1 + \frac{1}{r}, \left(\frac{x_{h-1} + l_h}{\theta}\right)^r\right) \right] \right]^2 \\ & \left. + \mu_{h\phi} \left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1} + l_h}{\theta}\right)^r} \right]^2 \right\} \\ \text{subject to} \quad & \sum_{h=1}^L l_h = d \\ \text{and} \quad & l_h \geq 0; \quad h = 1, 2, \dots, L \end{aligned} \quad (33)$$

where  $d = x_L - x_0 = b - a$ ,  $\beta$  is the regression coefficient,  $\theta$  and  $r$  are parameters of the Weibull distribution,  $\Gamma(\cdot)$  is the upper incomplete gamma function and  $Q(\cdot)$  is the upper regularized incomplete gamma function. Whereas,  $\mu_{h\phi}$  is the expected variance given in Equation (6) for the error term in the regression model Equation (4), which can be estimated as discussed in the following section.

#### 4.4. Estimating the variance of the error term

In the regression model given in Equation (24), it is assumed that the variance of the error term is  $V(\epsilon|x) = \phi(x)$  for all  $x$  in the range  $(a, b)$  and the expected value of the function  $\phi(x)$  given by  $\mu_{h\phi}$  is obtained by Equation (10).

Many authors have assumed that  $\phi(x)$  may be of the form:

$$\phi(x) = cx^g; \quad c > 0, \quad g \geq 0 \quad (34)$$

where  $c$  and  $g$  are constants and in many populations  $0 \leq g \leq 2$  (see Singh and Sukhatme 1969; Singh 1971; and Rizvi, Gupta, and Bhargava 2002).

Thus, from Equations (10), (23) and (35), we may compute  $\mu_{h\phi}$  as a function of boundary points as follows:

$$\begin{aligned} \mu_{h\phi} = & rc \Gamma(r+g) \left[ Q\left(r+g, \left(\frac{x_{h-1}}{\theta}\right)^r\right) \right. \\ & \left. - Q\left(r+g, \left(\frac{x_{h-1}+l_h}{\theta}\right)^r\right) \right] \\ & \div \theta^r \left[ e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1}+l_h}{\theta}\right)^r} \right] \end{aligned} \quad (35)$$

Therefore, one can determine the expected value of the stratum variance of the error term using Equation (34), if the values of the constants  $c$  and  $g$  are known. However, for our sample data, when a common regression model holds across the strata, we obtain the expected stratum variance of the error as

$$\mu_{h\phi} = \frac{SS_{Res}}{N-p} = MS_{Res} \quad (36)$$

where  $SS_{Res}$  and  $MS_{Res}$  are the sum of squares of residuals and mean square of residuals respectively, and  $p$  is the number of parameters in the regression model.

## 5. Results and discussion

This section presents the results by using the proposed method whereby the OSB of a population with Weibull auxiliary variable is computed. Considering that the estimation of the hemoglobin level for the population is of interest, the minimum and the maximum values of  $x$  (iron), are  $x_0 = 1.5$  and  $x_L = 25.1$ , respectively. This implies that the range of the distribution of iron level is  $d = x_0 - x_L = 23.6$ .

The problem of determining the OSB given in MPP Equation (34) is solved by reducing it into two stages (for  $k = 1$  and  $k \geq 2$ ) using the recurrence equations in Equations (21) and (22). These equations are solved to obtain optimum strata widths  $l_h^*$  and the optimum strata boundaries  $x_h^* = x_{h-1}^* - l_h^*$  by implementing the dynamic programming solution procedure via a C++ computer program.

Numerical investigations are also undertaken in this section to study the effectiveness of the proposed method compared to the following methods available in the literature:

1. Cum  $\sqrt{f}$  method of Dalenius and Hodges (1959).
2. Geometric method of Gunning and Horgan (2004).
3. Lavallée-Hidiroglou method Lavallee and Hidiroglou (1988) with Kozak's algorithm Kozak (2004).

The stratification package recently developed by Baillargeon and Rivest (2011) in the *R* statistical software is used to determine the OSBs for the methods mentioned above. These OSB are then used to compute the sample size of each stratum and the variance of the estimated mean (or the values of the objective function) so that a comparative analysis could be carried out.

**Table 3.** OSW, OSB and OFV for proposed method.

Strata	OSW	OSB	OFV
(L)	$(l_h^*)$	$(x_h^* = x_{h-1}^* + l_h^*)$	$\sum_{h=1}^L W_h \sigma_h$
2	$l_1^* = 10.72$	$x_1^* = 12.22$	1.3658
3	$l_2^* = 12.88$ $l_1^* = 7.79$ $l_2^* = 6.15$ $l_3^* = 9.66$	$x_1^* = 9.29$ $x_2^* = 15.44$	1.3462
4	$l_1^* = 6.22$ $l_2^* = 4.60$ $l_3^* = 4.98$ $l_4^* = 7.81$	$x_1^* = 7.72$ $x_2^* = 12.31$ $x_3^* = 17.29$	1.3384
5	$l_1^* = 5.20$ $l_2^* = 3.78$ $l_3^* = 3.75$ $l_4^* = 4.30$ $l_5^* = 6.57$	$x_1^* = 6.70$ $x_2^* = 10.48$ $x_3^* = 14.23$ $x_4^* = 18.53$	1.3346

**Table 4.** Optimum strata boundaries for the different methods.

L	CSRF		GEO		L-H Kozak		DP	
	OSB	OFV	OSB	OFV	OSB	OFV	OSB	OFV
2	12.12	1.366	6.14	1.404	8.1	1.384	12.22	1.366
3	9.76 15.66	1.346	3.84 9.81	1.369	5.55 9.15	1.372	9.29 15.44	1.346
4	7.40 12.12 16.84	1.339	3.03 6.14 12.41	1.353	5.55 9.15 15.55	1.342	7.71 12.31 17.29	1.338
5	6.22 9.76 13.30 18.02	1.335	2.64 4.63 8.13 14.21	1.345	5.55 9.15 12.65 17.00	1.335	6.70 10.48 14.23 18.53	1.335

Table 3 presents the OSW and OSB obtained by the proposed method (DP) together with the objective function values  $\sum_{h=1}^L \phi_h(l_h) = \sum_{h=1}^L W_h \sqrt{\sigma_{h\lambda}^2 + \mu_{h\phi}}$  (indicated as (OFV) in the tables) for  $L = 2, 3, 4, 5$ .

For comparison purposes, the OSB determined for cum  $\sqrt{f}$  method (CSRF), geometric method (GEO), Lavallée and Hidirolou's method (K-H (Kozak's algorithm)) using the stratification package with  $CV = 0.4575$  (obtained from the data) and the proposed dynamic programming method (DP) are presented in Table 4 for  $L = 2, 3, 4, 5$ . The optimum values of the objective function of the estimate are also presented (OFV). The optimum sample size (OSS) for each stratum using these OSB for the different methods are presented in Table 5.

Upon careful examination of Tables 4 and 5, it is noted that the OSB and the sample sizes obtained by the cum  $\sqrt{f}$  method are by far the closest to the proposed dynamic programming method. These values in the other two methods, namely geometric and Lavallée and Hidirolou's methods, differ vastly from that of the proposed method. It can also be seen that geometric method produces the larger sample size towards the tailer stratum as compared to others. Thus, it can be concluded that there seems to be a difference between the OSB

**Table 5.** Optimum sample size for the different methods with  $n = 500$ .

L	h	CSRf		GEO		L-H Kozak		DP	
		$n_h$	OFV	$n_h$	OFV	$n_h$	OFV	$n_h$	OFV
2	1	274		69		128		278	
	2	226	1.366	431	1.403	372	1.384	222	1.366
3	1	190		23		56		173	
	2	195		165		107		206	
	3	115	1.346	312	1.369	337	1.372	121	1.346
4	1	109		12		57		119	
	2	166		59		110		163	
	3	139		211		215		141	
	4	86	1.339	218	1.353	118	1.342	77	1.338
5	1	75		8		58		88	
	2	115		29		110		128	
	3	125		95		125		129	
	4	122		211		124		101	
	5	63	1.335	157	1.345	83	1.335	54	1.335

**Table 6.** Optimum stratum boundary for survey variable ( $y$ ).

No. of Strata	OSB for $x$	OSB for $y$	OFV of $y$
$L$	$(x_h)$	$\hat{y}_h = \hat{\alpha} + \hat{\beta}x$	$\sum_{h=1}^L W_h \sigma_h$
2	$x_1^* = 12.22$	$\hat{y}_1^* = 12.34$	1.366
3	$x_1^* = 9.29$ $x_2^* = 15.44$	$\hat{y}_1^* = 12.01$ $\hat{y}_2^* = 12.71$	1.346
4	$x_1^* = 7.71$ $x_2^* = 12.31$ $x_3^* = 17.29$	$\hat{y}_1^* = 11.82$ $\hat{y}_2^* = 12.35$ $\hat{y}_3^* = 12.92$	1.338
5	$x_1^* = 6.70$ $x_2^* = 10.48$ $x_3^* = 14.23$ $x_4^* = 18.53$	$\hat{y}_1^* = 11.71$ $\hat{y}_2^* = 12.14$ $\hat{y}_3^* = 12.57$ $\hat{y}_4^* = 13.06$	1.335

and the sample size obtained using the different methods including the proposed dynamic programming method.

By looking at the variances in [Tables 4 and 5](#), it can be seen that the proposed method yields the smallest variance for all  $L = 2, 3, 4$  and  $5$  as compared to all the other methods. Although the values of the objective function for the DP method are very close to the cum  $\sqrt{f}$  method, the other two methods produce a greater variance than the dynamic programming technique. Thus, the study reveals that the proposed dynamic programming technique is more efficient than the other stratification methods.

Finally, the OSB points of the survey variable,  $y$ , is obtained by using the OSB for the auxiliary variable (Iron) and applying the regression model Equation (24). These results are presented in [Table 6](#).

## 6. Conclusion

A well-designed sampling plan and efficient data mining strategies can greatly enhance the information that can be produced from a survey. In this paper, an optimal algorithm is

presented for data mining using stratified sampling. The proposed technique uses auxiliary information in the absence of the main study variable in designing the sampling plan.

The numerical example in the paper uses a real data set to illustrate the application of the method. The results reveal that the construction of strata using an auxiliary variable for a health population, which follows Weibull distribution, leads to substantial gains in the precision of the estimates of the main study variable. It is also evident from the results that, compared to other commonly used methods, the proposed technique performs much more efficiently.

The proposed method, unlike other classical methods, does not require any initial approximate solution and is able to obtain optimum solutions. With the main variable not available to us, the method uses the auxiliary variable and parametric assumptions of the main variable in order to understand the characteristics of the main variable. The proposed method can surely be extended to other statistical distributions that characterize the auxiliary information.

## References

- Baillargeon, S., and L. P. Rivest. 2011. The construction of stratified designs in R with the package Stratification. *Survey Methodology* 37 (1):53–65.
- Bellman, R. E. 1957. *Dynamic programming*. Princeton, NJ: Princeton University Press.
- Bühler, W., and T. Deutler. 1975. Optimal stratification and grouping by dynamic programming. *Metrika*. 22 (1):161–75.
- Christopher, W., and T. Blaxton. 1998. *Data mining solutions: methods and tools for solving real-world problems*. New York: John Wiley and Sons.
- Cochran, W. 1977. *Sampling techniques*. Wiley and Sons. New York: 98:259–61.
- Dalenius, T. 1950. The problem of optimum stratification. *Scandinavian Actuarial Journal* (3–4): 203–13.
- Dalenius, T. 1957. *Sampling in Sweden: Contributions to the methods and theories of sample survey practice*. Stockholm: Almqvist and Wiksell.
- Dalenius, T., and M. Gurney. 1951. The problem of optimum stratification. II. *Scandinavian Actuarial Journal* (1–2):133–48.
- Dalenius, T., and J. L. Hodges. 1959. Minimum variance stratification. *Journal of the American Statistical Association* 54 (285):88–101.
- Danish, F., S. E. H. Rizvi, M. I. Jeelani, and J. A. Reashi. 2017. Obtaining strata boundaries under proportional allocation with varying cost of every unit. *Pakistan Journal of Statistics and Operation Research* 13 (3):567–74.
- De Gruijter, J. J., B. Minasny, and A. B. Mcbratney. 2015. Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. *Journal of Survey Statistics and Methodology* 3(1):19–42.
- Ekman, G. 1959. An approximation useful in univariate stratification. *The Annals of Mathematical Statistics* 30 (1):219–29.
- Gunning, P., and J. M. Horgan. 2004. A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology* 30 (2):159–66.
- Gupta, R. K., R. Singh, and P. K. Mahajan. 2005. Approximate optimum strata boundaries for ratio and regression estimators. *Aligarh Journal of Statistics* 25:49–55.
- Hidiroglou, M. A., and M. Kozak. 2017. Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*.
- Hidiroglou, M. A., and K. P. Srinath. 1993. Problems associated with designing subannual business surveys. *Journal of Business & Economic Statistics* 11 (4):397–405.
- Horgan, J. M. 2006. Stratification of skewed populations: A review. *International Statistical Review* 74 (1):67–76.

- Jurina, I., and L. Gligorova. 2017. Determination of the optimal stratum boundaries in the monthly retail trade survey in the Croatian Bureau of Statistics. *Romanian Statistical Review* 65 (4):41–56.
- Khan, M. G. M., N. Ahmad, and K. Sabiha. 2009. Determining the optimum stratum boundaries using mathematical programming. *Journal of Mathematical Modelling and Algorithms* 8 (4):409–23. doi:10.1007/s10852-009-9115-3.
- Khan, M. G., K. G. Reddy, and D. K. Rao. 2015. Designing stratified sampling in economic and business surveys. *Journal of Applied Statistics* 42 (10):2080–99.
- Khan, M. G. M., N. Nand, and N. Ahmad. 2008. Determining the optimum strata boundary points using dynamic Programming. *Survey Methodology* 34 (2):205–14.
- Khan, E. A., M. G. M. Khan, and M. J. Ahsan. 2002. Optimum stratification: A mathematical programming approach. *Calcutta Statistical Association Bulletin* 52:323–33.
- Khan, M. G., N. Sehar, and M. J. Ahsan. 2005. Optimum stratification for exponential study variable under Neyman allocation. *Journal of the Indian Society of Agricultural Statistics* 59 (2): 146–50.
- Khan, M. G., D. Rao, A. H. Ansari, and M. J. Ahsan. 2015. Determining optimum strata boundaries and sample sizes for skewed population with log-normal distribution. *Communications in Statistics-Simulation and Computation* 44 (5):1364–87.
- Khan, M. G. M., and S. Sushita. 2015. Determining optimum strata boundaries and optimum allocation in stratified sampling. *Aligarh Journal of Statistics* 35:23–40.
- Kozak, M. 2004. Optimal stratification using random search method in agricultural survey. *Statistics in Transition* 6 (5):797–806.
- Kozak, M., and M. R. Verma. 2006. Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology* 32 (2):157–163.
- Kozak, M., M. R. Verma, and A. Zielinski. 2007. Modern approach to optimum stratification: Review and perspectives. *Statistics in Transition* 8 (2):223–50.
- Lavallée, P. 1988. *Some contributions to optimal stratification*. Ottawa, Canada: Carleton University.
- Lavallee, P., and M. Hidiroglou. 1988. On the stratification of skewed populations. *Survey Methodology* 14 (1):33–43.
- Mehta, S. K., R. Singh, and L. Kishore. 1996. On optimum stratification for allocation proportional to strata totals. *Journal of Indian Statistical Association* 34:9–19.
- Nand, N., and M. G. M. Khan. 2009. Optimum Stratification for Cauchy and power type study variable. *Journal of Applied Statistical Science* 16 (4):453–462.
- 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 (4):558–625.
- Nicolini, G. 2001. A method to define strata boundaries. Università di Milano, Dipartimento di economia politica e aziendale, Milano.
- Niemiro, W. 1999. Optimal construction of strata using random search method. *Wiadomosci statystyczne* 10:1–9.
- Patten, S. B. 2006. A major depression prognosis calculator based on episode duration. *Clinical Practice and Epidemiology in Mental Health* 2 (1):1–13.
- Reddy, K. G., M. G. M. Khan, and D. K. Rao. 2016. A Procedure for computing optimal stratum boundaries and sample sizes for multivariate surveys. *Journal of Software* 11 (8):816–32. doi: 10.17706/jsw.11.8.816–32.
- Rivest, L. P. 2002. A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology* 28 (2):191–8.
- Rizvi, S. E. H., J. P. Gupta, and M. Bhargava. 2002. Optimum stratification based on auxiliary variable for compromise allocation. *Metron* 60 (3–4):201–15.
- Serfling, R. J. 1968. Approximately optimal stratification. *Journal of the American Statistical Association* 63 (324):1298–309.
- Sethi, V. K. 1963. A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics* 5 (1):20–33.
- Singh, R. 1971. Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association* 66 (336):829–33.

- Singh, R., and D. Dev Prakash. 1975. Optimum stratification for equal allocation. *Annals of the Institute of Statistical Mathematics* 27 (1):273–80.
- Singh, R., and B. V. Sukhatme. 1969. Optimum stratification. *Annals of the Institute of Statistical Mathematics* 21 (1):515–28.
- Singh, R., and B. V. Sukhatme. 1972. Optimum stratification in sampling with varying probabilities. *Annals of the Institute of Statistical Mathematics* 24 (1):485–94.
- Singh, R., and B. V. Sukhatme. 1973. Optimum stratification with ratio and regression methods of Estimation. *Annals of the Institute of Statistical Mathematics* 25 (1):627–33.
- Sweet, E. M., and R. S. Sigman. 1995. Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Section on Survey Research Methods* 1:491–6.
- Taga, Y. 1967. On optimum stratification for the objective variable based on concomitant variables using prior information. *Annals of the Institute of Statistical Mathematics* 19 (1):101–29.
- Taha, H. A. 2007. *Operations research: An introduction* 8th edition, New Jersey: Pearson Education, Inc.
- Wang, H., Z. Wang, X. Li, B. Gong, L. Feng, and Y. Zhou. 2011. A robust approach based on Weibull distribution for clustering gene expression data. *Algorithms for Molecular Biology* 6 (1):1–14.
- Wahed A. S., T. M. Luong, and J. H. Jeong. 2009. A new generalization of Weibull distribution with application to a breast cancer data set. *Statistics in Medicine* 28 (16):2077–94.
- Yong, F. H., L. Tian, S. Yu, T. Cai, and L. J. Wei. 2016. Optimal stratification in outcome prediction using baseline information. *Biometrika* 103 (4):817–28.