Research Article

Open Access CrossMark

Carrying out single-stage probability sampling designs using functions in R software

Abstract

The aim of this paper is to demonstrate how to implement some single-stage sampling designs as well as design-based estimation of finite population parameters and their variances using available R packages. A simulated data set has been used to show how the codes work.

Keywords: household income survey, optimum allocation, probability sampling design, simulation study

Volume 11 Issue 1 - 2022

Hamid Ghorbani
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

Correspondence: Hamid Ghorbani, Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran,
Email hamidghorbani@kashanu.ac.ir

Received: December 20, 2021 | **Published:** January 22, 2022

Introduction

In gathering data about a group of individuals or items, rather than conducting a full census, very often a sample is taken from a larger population in order to save time and resources. These samples can be classified into two major groups describing the way in which they were chosen: probability samples and non-probability samples. Probability sampling involves random selection, allowing one to make statistical inferences about the whole group. In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included. It is worth mentioning that in non-probability sampling the chance of any member being selected for a sample cannot be calculated. It's the opposite of probability sampling, where you can calculate the chances. Various forms of random sampling include simple, stratified, cluster, systematic, and multi-stage random sampling and the ones for non-sampling techniques include quota, snowball, judgment, convenience, haphazard, purposive, expert, diversity, and modal instance sampling.[1] One major disadvantage of non-probability sampling is that it's impossible to know how well the sample is representing the population.[2,3] The lack of an underlying mathematical theory does not permit unbiased estimation of the population mean along with measurable sampling error because its theory is not based on design-based inference. Therefore, the confidence intervals and margins of error cannot be calculated. On the other hand, a major advantage with non-probability sampling is that, compared to probability sampling, it's very cost and time-effective.[4] Most internet surveys today benefit from these advantages, in which large numbers of volunteers encourage by survey companies to fulfill the questionnaires in exchange for cash or gifts.[5] In sum, non-probability sampling is not ideal for quantitative research because results from non-probability samples cannot be generalized to the population, the probability sampling may on the other hand be less appropriate for qualitative studies in which the goal is to describe a very specific group of people and generalizing the results to population is not the focus of the study.

This paper provides a brief introduction of how to use two packages in statistical software R for making inferences for survey data regarding different probability sampling designs. We demonstrate the estimation of the population mean, proportion, total, domain means, and totals and their associated variances and confidence intervals. In addition, the problem of usual and optimal sample size determination is also discussed.

Sampling analysis with R

R is an open-source, multi-platform and excellent programming environment for statistical computing and graphics developed by Ross Ihaka and Robert Gentleman in 1993.[6,7] The availability of R packages makes it stand different from the other languages. There are thousands of packages available which perform all sorts of exceptional tasks. Nowadays, the exists a large expansion of R packages dedicated to surveying sampling methods. A comprehensive list of all packages dedicated to surveying sampling techniques and official statistics is given at https://cran.r-project.org/web/views/OfficialStatistics.htmlmaintained by M.Templ. sampling[8] and survey[9] packages are two main R packages addressed the survey sampling methods. While, the main concern of the sampling package is performing sample selection according to various with or without replacement sampling designs, the survey package concerned mainly with a design-based estimation of finite population parameters and their variance[10] which should be installed and be loaded using the following commands:

```
>install.packages(c('sampling','survey'), dep=T)

>library(sampling); library(survey).
```

Analyzed data

In the following sections, different probability sampling designs are analyzed using simulated data with the following hypothesized meaningful structure.

Suppose in the rural areas of four neighbor cities (called A to D) located in a particular county region, having different socio-economics conditions, a survey was conducted to estimate the mean yearly income of households. The tax administration needs also an estimate of total household income in four cities. In our simulation, we assume that the cities have 400, 700, 600, and 650 households each reported with a predominance of agricultural or non-agricultural workers. Also, the number of working adults in the household was recorded as one, two, or three workers. A list of all 2350 households and *their household income*, let's say, last 12 months' incomes and receipts are available. It was assumed that the household's income has a Pareto distribution with the following distribution function,[11]

$$F_X\left(x\right)=1-(\frac{1000}{x})^2, x>1000.$$

Therefore, the mean income in the population is 2000 and the standard error is infinite. The data were generated using the following codes. For simulated data, the mean and standard error of income variable is 2122 and 3624.70, respectively.

```
>set.seed(731313126)
>simuldata = rbind(matrix(rep("A", 400), 400, 1, byrow =
TRUE),matrix(rep("B", 700),
700,1, byrow = TRUE),matrix(rep("C",600), 600, 1, byrow = TRUE),
matrix(rep("D", 650), 650, 1, byrow = TRUE))
>agricpredomworkerA<-rep(c(0,1),c(100,300))
>agricpredomworkerB<-rep(c(0,1),c(150,550))
>agricpredomworkerC<-rep(c(0,1),c(300,300))
>agricpredomworkerD<-rep(c(0,1),c(300,350))
>agrpredomworker<-c(agricpredomworkerA,agricpredomworkerB,
agricpredomworkerC,agricpredomworkerD)
>householdwsize<-NA
>x<-runif(nrow(simuldata))
for (i in 1:length(x)){
if (x[i] < 0.3) householdwsize[i] = 1
else if ((x[i] >= 0.3) && (x[i] <= 0.5)) householdwsize[i] = 2
else householdwsize[i] = 3}
>install.packages('Pareto',dep=T); library(Pareto); income<-
rPareto(2350, 1000, 2)
>simuldata = cbind.data.
frame(simuldata,agrpredomworker,income,householdwsize)
>names(simuldata)<-c("cities", "agrpredomworker", "income",
"householdwsize")
```

Simple random sampling

Simple random sampling (SRS) is a method of selection of a sample comprising of *n* number of sampling units from the population having *N* number of units such that every sampling unit has an equal chance of being chosen. The samples can be drawn in two possible ways, namely sampling with or without replacement.

[examp1] Suppose in a country region having four cities with different socio-economics statuses a survey was conducted to estimate the regional average income of the household. Regarding the tax issues, the administration is also interested to have an estimate of total household income in this region. A list of all 2350 households' income is available, along with the household's predominance of agricultural or non-agricultural workers identification and the number of the households workers. Suppose an SRS without replacement (srswor) of n = 200 is required for the survey. The following codes show how an srswor design is conducted in order to estimate the population parameters, mean and total.

```
> N<-nrow(simuldata);  n<-200;  swor<- srswor(n,N);  s
<- simuldata[swor==1,]
>srs<- svydesign(id=~1,data=s,fpc=~rep(N,n))
```

The id=~1 says that individual households were sampled (there is one row for each household in the data set). The variable 'fpc' contains the population size.

```
> summary(srs)
## Independent Sampling design
## svydesign(id = ~1, data = s, fpc = ~rep(N, n))
## Probabilities:
##  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08511 0.08511 0.08511 0.08511 0.08511 0.08511
## Population size (PSUs): 2350
## Data variables:
## [1] "cities"   "agrpredomworker"    "income"      "householdwsize"
## Independent Sampling design
> m<-svymean(~income, srs); m;  confint(m, level = 0.95)
##       mean    SE
##income 2060.9 163.1
##          2.5 %   97.5 %
##income 1741.195 2380.543
> t<-svytotal(~income, srs); t;  confint(t, level = 0.95)
##        total     SE
##income 4843043 383290
##        2.5 %  97.5 %
##income 4091809 5594277
The following codes show how the SRS design with replacement
might be implemented.
>N<-nrow(simuldata);  n<-200; swr <- srswr(n,N); s <-
simuldata[swr!=0,]
>s$freq<-swr[swr!=0]; m<-sum(s$income*s$freq)/200
>v<-var(rep(s$income,s$freq))/200 s.e<-sqrt(v)
##m=1793.917 , s.e= 88.822
```

[examp2] Consider again [examp1]. Suppose we are now interested to estimate the mean income of the households with three workers for which their income is less than the median income of all households using an srswor with the sample fraction equal to 0.1.

for estimating the desired parameter first the target population, i.e. list of all households with three workers for which their income is less than the median income of all households is extracted from a list of 2350 households. Doing this using R codes below the new population size is 588. From this population, a random sample of size 60, is drawn according to ansrswor design.

```
>simuldata$p<-NA
for (i in 1:nrow(simuldata))
simuldata$p[i]<-
ifelse((householdwsize[i]==3
&simuldata$income[i]<median(simuldata$income)),1,0)
>table(simuldata$p)
##  0    1
##1762  588
>newsimuldata<-simuldata[simuldata$p==1,]
>newN<-nrow(newsimuldata) ; n<-60
>swor <- srswor(n,newN); s <- newsimuldata[swor==1,]
>srsh <- svydesign(id=~1,data=s,fpc=~rep(newN,n))
>m<-svymean(~income, srsh); m;  confint(m, level = 0.95)
##        mean     SE
##income  1180.5 15.037
##          2.5 %    97.5 %
##income  1151.025  1209.968
```

[examp3] Consider again [examp2]. Suppose we are now interested to estimate the proportion of the households with three workers for which their income is less than the median income of all households.

The desired parameter value in the population is $P=\frac{588}{2350}\approx 0.23$. A random sample of size 60 is drawn using srswor design to estimate *P*. The following codes might be used for doing this practice.

```
>N<-nrow(simuldata); n<-60; swor <- srswor(n,N); s <-
simuldata[swor==1,]
>srsp <- svydesign(id=~1,data=s,fpc=~rep(N,n))
>m<-svymean(~p, srsp); m; confint(m, level = 0.95)
## mean    SE
##p 0.25 0.0556
##      2.5 %    97.5 %
##p 0.1409297 0.3590703
```

Sampling with varying probabilities

The simple random sampling scheme provides a random sample where every unit in the population has an equal probability of selection. Under certain circumstances, more efficient estimators are obtained by assigning unequal probabilities of selection to the units in the population. This type of sampling is known as varying probability sampling or probability proportional to size (PPS) sampling design. Unequal selection probabilities are often based on auxiliary variable values which are measures of the sizes of population units. Assume that the auxiliary variable is $x=\left(x_1,x_2,\cdots,x_N\right)$. Then the first-order inclusion probability, $\pi_j=\frac{nx_j}{\sum_{i=1}^{N}x_j}$, of each population unit $j=1,2,\cdots,N$ to be included in the sample is proportional to x_j, i.e., the corresponding value of an auxiliary variable. Units for which the probability is larger than one are selected with certainty, while the inclusion probabilities for the remainder of the units are calculated after excluding the large ones.

[examp4] Consider again [examp1]. Suppose we are again interested to estimate the mean and total household income using a PPS sampling design with households number of workers as an auxiliary variable.

For carrying out the PPS sampling design, using the following codes, first the (first-order) inclusion probabilities π_j are calculated according to the number of household workers.[12] Then the minimum support method[13] is implemented to select 200 samples of household units. Many other unequal sampling designs, amongst them to name, the balanced sampling, the Brewer, Sampford, Tillé are also implemented in the sampling package.[14]

```
>Tot=simuldata$householdwsize
>pj=inclusionprobabilities(Tot,200)
>s=UPminimalsupport(pj)
>insample<-NA
>insample<-simuldata[s==1,]
>simuldata$w <- NA
for (i in 1:3)
simuldata$w[simuldata$householdwsize== i] <- (sum(simuldata$householdwsize == i))
>insample$psi<-insample$w/2350
>dppswr<- svydesign(id=~1, probs=~psi, data=insample)
>m<-svymean(~income, dppswr); m; confint(m, level = 0.95)
##        mean    SE
##income 2074.5 217.16
##          2.5 %   97.5 %
##income 1648.911 2500.147
t<-svytotal(~income, dppswr); t; confint(t, level = 0.95)
##        total    SE
##income 1224393 134912
##          2.5 %  97.5 %
##income 959970.8 1488814
```

Stratified random sampling

If the population is non-homogeneous with respect to the characteristic under study then the SRS design does not provide a representative sample since each possible sample is equally likely to occur and the sample variance would not be able to represent the heterogeneity in the population. Stratified random sampling involves dividing the entire population into homogeneous sub-groups called stratum such that the sampling units within each stratum are homogeneous with respect to the characteristic under study and heterogeneous among these strata. Samples are then selected from each stratum using a srswor design in each stratum and combined to form the full sample. Since the variance of the sample means not only depends not on the sample size but also on the population variance, the stratified sampling scheme increases the precision of the estimator by reducing the heterogeneity in the population.

[examp5] Regarding the different number of working adults in the household, makes it reasonable to stratify households into three strata St_1, St_2 and St_3 according to their number of workers. The number of samples taken out of each stratum using srswor design is proportional to its size. The observed frequencies of households with one, two, and three workers are 692, 494 and 1164, receptively. Using the following codes first the $n_1=61$, $n_3=99$, and $n_3=99$ samples are drawn randomly from the corresponding stratum to build strsample data frame. Before specifying the design, it is necessary to include the stratum sizes in the data frame.

```
>insample<-c(sample(1:692,61),sample(693:1187,40),sample(1188:nro
w(simuldata),99))
>strsample<-NA
>newdata <- simuldata[order(householdwsize),]
>strsample<-newdata[insample,]
>strsample$stratasize<- NA
>strsample$stratasize[strsample$householdwsize==1] <- 692
>strsample$stratasize[strsample$householdwsize==2] <- 494
>strsample$stratasize[strsample$householdwsize==3] <- 1164
>stratadesign <- svydesign(id = ~1, strata = ~householdwsize, data =
strsample,
fpc = ~stratasize)
>summary(stratadesign)
##Stratified Independent Sampling design
##svydesign(id = ~1, strata = ~householdwsize, data = strsample,
##    fpc = ~stratasize)
##Probabilities:
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##0.08454 0.08454 0.08555 0.08511 0.08555 0.08584
##Stratum Sizes:
##            1  2  3
##obs        61 40 99
##design.PSU 61 40 99
##actual.PSU 61 40 99
##Population stratum sizes (PSUs):
##   1    2    3
## 692  494 1164
##Data variables:
##[1] "cities"          "agrpredomworker" "income"
"householdwsize"
##[5] "stratasize"
>m<-svymean(~income, stratadesign); m; confint(m, level = 0.95)
##         mean     SE
##income 2178.6 227.01
##          2.5 %   97.5 %
##income 1733.712 2623.588
>t<-svytotal(~income, stratadesign); t; confint(t, level = 0.95)
##         total     SE
##income 5119827 533481
##         2.5 %  97.5 %
##income 4074224 6165431
```

Sample size determination

Consider an unbiased estimator $\hat{\theta}$ for θ. The associated, $100(1-\alpha)\%$, confidence interval can be expressed as

$$P(\left|\hat{\theta}-\theta\right|<e)=1-\alpha, \tag{1}$$

Where e is called the margin of error and is a function of $\operatorname{var}\left(\hat{\theta}\right)$. Therefore, for determining the sample size, level of precision, level of confidence, and degree of variability of measured variable need to be specified.[13]

Simple random sampling

If we are estimating the mean μ, in a population with size σ^2 and measurement variability σ^2, using n samples drawn randomly according to a srswor sampling design, for large enough $\bar{y}$ if $\bar{y}$ can be treated as being normally distributed a $100(1-\alpha)\%$ confidence interval for is given by:

$$\bar{y}\pm z_{1-\alpha/2}\sqrt{\widehat{\operatorname{var}}(\bar{y})},$$

(where $\widehat{\operatorname{var}}(\bar{y})=\left(1-\frac{n}{N}\right)\frac{\hat{\sigma}^2}{n}$ and the term $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution). The following form of the confidence interval formula

$$P(\left|\bar{y}-\mu\right|<z_{1-\alpha/2}\sqrt{\widehat{\operatorname{var}}(\bar{y})})=1-\alpha, \tag{2}$$

is equivalent to set the margin of error in (1) to, $e=z_{1-\alpha/2}\sqrt{\widehat{\operatorname{var}}(\bar{y})}$, specified as the half-width of the normal approximation confidence interval for the population mean. Solving for, $\pm e$, the sample size required to provide an interval estimate with precision $\pm e$, gives the required sample size as

$$n=\frac{z_{1-\alpha/2}^2\hat{\sigma}^2}{e^2+\frac{z_{1-\alpha/2}^2\hat{\sigma}^2}{N}}. \tag{3}$$

As mentioned in,[15] for some applications like household surveys for which the value of sampling fraction is negligible, compared with the school surveys where the population size is also small, we obtain

$$n=(\frac{z_{1-\alpha/2}s}{e})^2 \tag{4}$$

The R function, nContMoe in thePracTools package,[16] will compute a sample size using the formula (4). The parameters used by the function are shown in the example below:

```
>library(PracTools)
>nContMoe(moe.sw=1, e=0.1, alpha=0.05, S2=4),
```

which yields $n=1536.584\simeq 1537$. The default value of N is infinity, but a user-specified value can also be provided to compute a sample size using the formula (3). For example

```
>nContMoe(moe.sw=1, e=0.05, alpha=0.1, S2=4, N=10000),
```

which yields $n=3021.082\simeq 3022$.

In addition, it is worth mentioning that the large sample approximation used to make the normality assumption for the distribution of the estimator may lead us to underestimate the required sample sizes, see[17]and[18] for details. On the other hand, if we require to bound the relative error of estimation, i.e.

$$P\left(\left|\frac{\bar{y}-\mu}{\mu}\right|<r\right)=1-\alpha, \tag{5}$$

we set in formula (3), $e=r\mu$ and solve for the sample size n to obtain

$$n=\frac{z_{1-\alpha/2}^2\hat{\sigma}^2}{r^2\hat{\mu}^2+\frac{z_{1-\alpha/2}^2\hat{\sigma}^2}{N}}.$$

In the above equations, where necessary, we need either prior knowledge to replace the unknown parameters or a preliminary sample of some sort to estimate them.

Stratified random sampling

Using stratified random sampling requires that we decide how best to allocate effort among strata so that the sampling process will provide the most efficient balance of effort, cost, and estimate precision. The overall sampling efficiency depends on the allocation strategies which are based on information like the variability within each stratum, the relative cost of obtaining and measuring a sample unit from each stratum, and the number of sample units in each stratum. If there is no information available about the variability of units within strata, the cost of sampling is similar for all strata, and strata are of similar size, the simplest allocation strategy is the uniform allocation. On the other hand, the number of sample units to select from each stratum can be made proportional to the number of units within each stratum. Since the variation in a stratum often increases with the size of a stratum, so in some cases, this can be considered a rough approach for allocating more effort to strata that are likely to be more variable strata. In the following additional sampling strategies along with their implementation in R software is presented.

Optimal allocation, equal sampling costs

Neyman allocation is a method used to allocate samples to strata based on the strata variances and similar sampling costs in the strata. A Neyman allocation scheme provides the most precision for estimating the population mean given a fixed total sample size.[19]

$$n_h = n\frac{N_hS_h}{\sum_h N_hS_h}, \tag{6}$$

in which, N_h and h are the total number of the units and the true standard error of the interesting variable related to these units within the stratum h, respectively.

This formula says that if N_hS_h is large, then the corresponding stratum should be sampled heavily. This is very natural since large N_h means that the stratum contains a large portion of the population and large S_h means that the population values in the stratum are quite non homogeneous and, therefore, to estimate the stratum mean accurately a relatively large sample size is needed.

[examp6] Remember the households of the rural areas of four neighboring cities. A survey was conducted to estimate the mean yearly income of households. Regarding different numbers of worker adults in the household, makes it reasonable to stratify households into three strata St_1, St_2 and St_3 according to their number of workers. The corresponding standard deviations of the households incomes within these strata are 4987.271, 2612.676, and 2974.448, respectively. The optimum Neyman allocations for each stratum are calculated according to the formula (6) using the following codes for the total sample of, let say, 400 households.

```
>St1<-subset(simuldata, householdwsize == 1)
>St2<-subset(simuldata, householdwsize == 2)
>St3<-subset(simuldata, householdwsize == 3)
>Nh<-c(dim(St1)[1], dim(St2)[1], dim(St3)[1])
>Sh <-c(sd(St1$income), sd(St2$income), sd(St3$income))
>library(PracTools)
>strAlloc(n.tot = 400, Nh = Nh, Sh = Sh, alloc = "neyman")
```

This will result the below R output.

```
 allocation = neyman
Nh = 692, 494, 1164
Sh = 4987.271, 2612.676, 2974.448
nh = 168.26645, 62.92757, 168.80598
nh/n = 0.4206661, 0.1573189, 0.4220150
anticipated SE of estimated mean = 157.7261.
```

Optimal allocation, unequal sampling costs

The simplest form of the cost function used in a stratified sample survey is composed of an overhead cost C_0, for instance, costs of preparing the sampling frame, and a variable cost, which is written as a linear combination of the stratum sample sizes, i.e., $C = c_0 + \sum_h n_hc_h$. In stratified random sampling with the mentioned linear cost function, let $\bar{y}_h$ be a mean estimate in stratum $\bar{y}_{st}$ and $\bar{y}_{st}$ be the total mean estimate. Then, if the sample size n_h in stratum h is proportional to $\frac{N_hS_h}{c_h}$, the $\widehat{\text{var}}(\bar{y}_{st})$ is minimum for a specified cost C, and the total cost is a minimum for a specified variance $\text{var}(\bar{y}_{st})$ [20]. This yields the n in terms of n,

$$n_h = n\frac{N_hS_h/\sqrt{c_h}}{\sum_h(N_hS_h/\sqrt{c_h})}, \tag{7}$$

but we do not yet know what value *n* has. The optimal sample size that minimize $\widehat{\text{var}}(\bar{y}_{st})$ for a specified total cost C gives

$$n = \frac{(C-c_0)\sum_h(N_hS_h/\sqrt{c_h})}{\sum_h(N_hS_h\sqrt{c_h})}. \tag{8}$$

If the variance (V) is fixed, the optimal sample size that minimized the total cost is computed as

$$n = \frac{\left(\sum_h W_hS_h\sqrt{c_h}\right)\left(\sum_h W_hS_h/\sqrt{c_h}\right)}{V + \frac{1}{N}\sum_h W_hS_h^2}, \tag{9}$$

where $N = \sum_h N_h$ and $W_h = \frac{N_h}{N}$.

Two examples of optimal allocations using R, based on cost-constrained, and variance constrained are as follows:

[examp7] Refer to [examp6] and reconsider the values of the S_h and S_h. A survey was conducted to estimate the mean yearly income of households. Cost per household in three stratum St_1, St_2 and St_3 is 8, 8 and 16 per unit.

How would you distribute a sample that minimize the total field cost to make $\sqrt{\widehat{\text{var}}(\bar{y}_{st})} = 100$?

```
>Nh<-c(dim(St1)[1], dim(St2)[1], dim(St3)[1])
>Sh <-c(sd(St1$income), sd(St2$income), sd(St3$income))
>ch <-c(8, 8,16)
# fixed variance
```

>strAlloc(Nh = Nh, Sh = Sh, V0 = (100)^2, ch = ch, alloc = “totvar”)

allocation = totvar

Nh = 692, 494, 1164

Sh = 4987.271, 2612.676, 2974.448

nh = 386.3007, 144.4671, 274.0317

nh/n = 0.4799962, 0.1795069, 0.3404969

anticipated SE of estimated mean = 100

If the object is to find the sample size required to minimize $\widehat{\text{var}}\left(\bar{y}_{st}\right)$ for specified total cost value, $C = 10000$, how should the sample be distributed?

# fixed total cost C

>strAlloc(Nh = Nh, Sh = Sh, cost = 10000, ch = ch, alloc = “totcost”)

allocation = totcost

Nh = 692, 494, 1164

Sh = 4987.271, 2612.676, 2974.448

nh = 447.5917, 167.3884, 317.5099

nh/n = 0.4799962, 0.1795069, 0.3404969

anticipated SE of estimated mean = 88.68425.

It is worth mentioning when using strAlloc to compute the optimized value for n_h , in some cases achieving a specified variance or cost is impossible because the number of available strata is too small (i.e., $n_h > N_h$ for some h). In these cases, a warning will be given.

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgments

None.

References

1. A Field, J Miles, Z Field. Discovering Statistics Using R. *SAGE Publications Ltd*, London. 2012.
2. B MacInnis, JA Krosnick, AS Ho, et al. The accuracy of measurements with probability and nonprobability survey samples replication and extension. *Public Opinion Quarterly.* (2018);82(4):707–744.
3. N Malhotra, JA Krosnick. The effect of survey mode and sampling on inferences about political attitudes and behavior: comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis*. 2007;15:286–324.
4. A Wiśniowski, JW Sakshaug, DAP Ruiz, et al. Integrating probability and non-probability samples for survey inference. *Journal of Survey Statistics and Methodology*. 2020;8(1):120–147.
5. M Callegaro, RP Baker, J Bethlehem, et al. Online panel research: a data quality perspective. *John Wiley & Sons*, New York, 2014;512.
6. B S Everitt , A Skrondal. The cambridge dictionary of statistics. *Cambridge University Press*, New York. 2010.
7. AT Arnholt, AF Militino, MD Ugarte. Probability and statistics with R. *CRC Press*, Boca Raton. 2016.
8. Y Tillé, A Matei. sampling: Survey Sampling. *R package version. 2.9*. (2021).
9. TS Lumley. survey: analysis of complex survey samples. *R package version 4.0*. (2020).
10. T Lumley. Complex surveys: a guide to analysis using R. *John Wiley & Sons*. New Jersey. 2010.
11. C Kleiber, S Kotz. Statistical size distributions in economics and actuarial sciences. *John Wiley & Sons,*Hoboken. 2003.
12. A Gelman, T C Little. Improving on probability weighting for household size. *Public Opinion Quarterly*. 1998;62(3):398−404.
13. FG Miaoulis, RD Michener. An introduction to sampling. *Kendall Hunt Pub, Iowa*. 1976.
14. C Goga. Brief overview of survey sampling techniques with R. *Romanian Statistical Review*. 66(1):83−94.
15. R Valliant, JA Dever, F Kreuter. Practical tools for designing and weighting survey samples. *Springer*, New York. 2018.
16. R Valliant, JA Dever, F Kreuter. Prac tools: tools for designing and weighting survey samples. *R package version 1.2.2*. (2020).
17. LL Kupper, KB Hafner. How appropriate are popular sample size formulas?. *The American Statistician*. 1989;43(2):101−105.
18. TG Gregoire, DLR.Affleck. Estimating desired sample size for simple random sampling of a skewed population. *The American Statistician*. 2018;72(2):184−190.
19. WG Cochran. Sampling techniques. *John Wiley & Sons*, New York, 1977.
20. JJ de Gruijter, MFP Bierkens, DJ Brus, et al. Sampling for natural resource monitoring. *Springer*, Heidelberg. 2006.