On design and analysis of sample surveys:

Sampling strategies for elements with equal probability sampling designs

By:
Andrés Gutiérrez

Contents

Aim:

To realize the advantages and disadvantages of the main sampling strategies where sampling frames of elements are used

Contenido:

This presentation focuses on the practical development of a survey with the LUCY population and the sampling frame MARCO:

1. R and TeachingSampling
2. Marco and Lucy
3. Strategies for Bernoulli sampling
4. Strategies for simple random sampling
5. Strategies for simple random sampling with replacement
6. Strategies for Systematic sampling

Bibliography :

- Estrategias de muestreo. Gutiérrez (2009). USTA.
- Model Assisted Survey Sampling. Sarndal (1992). Springer.

¿Why R?

If you're using softwares such as SAS, SPSS, Stata or Systat why use R?

1. It's free. If you are a teacher or a student, the benefits are obvious. If you work in a company, your boss will assess it more when he finds that he should not paid for any annual licence to perform their statistical analysis.

2. Is enforceable in a variety of platforms including Windows, Unix and MacOS.

3. It provides a programming platform for new statistical methods with a simple mannered.

4. Provides advanced statistical routines that are not yet available in other packages.

5. Generates powerful graphics up-to-date with the state of the art.

6. The routines created in R and can be loaded and executed in other important software such as SAS and SPSS.

The TeachingSampling package

```
TeachingSampling:     Sampling designs and parameter
estimation in finite population

Foundations of inference in survey sampling

Version:       1.4.9
Depends:       R (≥ 2.6.0)
Published:     2010-03-11
Author: Hugo Andres Gutierrez Rojas
Maintainer:    Hugo Andres Gutierrez Rojas
               <hugogutierrez at usantotomas.edu.co>
License:       GPL (≥ 2)

URL:    http://www.predictive.wordpress.com/stats/
```

In R: Menu → Packages → Install Package → choose your preferred server → search and click on TeachingSampling. Load the package with the following instruction:

```
> library(TeachingSampling)
```

Marco and Lucy

• Lucy refers to a population of companies in the industrial sector.
• Marco refers to the sampling frame that is required to design a survey in order to asses probabilistic inference about Lucy.

The target population comprise all companies whose main activity is linked to the industrial sector. The measurement process will be based on: revenue in the last fiscal year, tax declared in the last fiscal year and number of employees. Additionally, it is required to know if the company sends periodically some kind of advertising material via email.

To recollect the information, an interviewer will visit the facilities of the company and will take the following questions:

1. In the last fiscal year, how much revenue were amounted by this business?
2. In the last fiscal year, how much taxes were declared by this company?
3. Currently, how many employees are working for this company?
4. Does this company use to send publicity material for e-mail to its customers or potential customers?

Marco and Lucy

To address the selection of a sample that allow statistical inference about the economic growth of the industrial sector, it must be provided a sampling frame with the following features for each company of the population.

1. ID: it is an alphanumeric string of two letters and three digits. This identification number is given to each company at the time of legal establishment to the relevant registration entity.
2. Location: is the address that is recorded in the statement of the taxes.
3. Area: the city is made up of neighborhoods or geographical areas. Depending of the address, the company belongs to one and only one geographical area of the city.
4. Level: according to tax records, businesses are categorized into three groups:

   1. Large: companies taxed $ 49 million a year or more.
   2. Medium: companies are taxed more than 11 million and less than 49 million dollars a year.
   3. Small: companies taxed $ 11 million a year or less

Information concerning the first 10 companies in the sampling frame is displayed with the following code in R:

```
> data(Marco)
> Marco[1:10,]
      ID Ubication Level Zone
1  AB001      c1k1 Small    A
2  AB002      c1k2 Small    A
3  AB003      c1k3 Small    A
4  AB004      c1k4 Small    A
5  AB005      c1k5 Small    A
6  AB006      c1k6 Small    A
7  AB007      c1k7 Small    A
8  AB008      c1k8 Small    A
9  AB009      c1k9 Small    A
10 AB010     c1k10 Small    A

> names(Marco)
[1] "ID" "Ubication" "Level" "Zone"
> dim(Marco)
[1] 2396    4
```

Marco and Lucy

Information for all of the characteristics of interest concerning the first 10 companies of the LUCY population is revealed by the code below:

```
> data(Lucy)
> Lucy[1:10,]
      ID Ubication Level Zone Income Employees Taxes SPAM
1  AB001      c1k1 Small    A    281        41   3.0   no
2  AB002      c1k2 Small    A    329        19   4.0  yes
3  AB003      c1k3 Small    A    405        68   7.0   no
4  AB004      c1k4 Small    A    360        89   5.0   no
5  AB005      c1k5 Small    A    391        91   7.0  yes
6  AB006      c1k6 Small    A    296        89   3.0   no
7  AB007      c1k7 Small    A    490        22  10.5  yes
8  AB008      c1k8 Small    A    473        57  10.0  yes
9  AB009      c1k9 Small    A    350        84   5.0  yes
10 AB010     c1k10 Small    A    361        25   5.0   no
```

Marco and Lucy

The statistics concerning the variables in the population are displayed easily by applying the `summary` function to the dataset in Lucy.

Can all of them be considered as parameters?

```
> summary(Lucy)

       ID           Ubication        Level        Zone        Income
 AB001    :   1   c10k1   :   1   Big   :  83   A:307   Min.   :   1.0
 AB002    :   1   c10k10  :   1   Medium: 737   B:727   1st Qu.: 230.0
 AB003    :   1   c10k11  :   1   Small :1576   C:974   Median : 390.0
 AB004    :   1   c10k12  :   1                 D:223   Mean   : 432.1
 AB005    :   1   c10k13  :   1                 E:165   3rd Qu.: 576.0
 AB006    :   1   c10k14  :   1                         Max.   :2510.0
 (Other):2390     (Other):2390


Employees            Taxes          SPAM
 Min.   :  1.00   Min.   :  0.50   no : 937
 1st Qu.: 38.00   1st Qu.:  2.00   yes:1459
 Median : 63.00   Median :  7.00
 Mean   : 63.42   Mean   : 11.96
 3rd Qu.: 84.00   3rd Qu.: 15.00
 Max.   :263.00   Max.   :305.00
```

Marco and Lucy

An important parameter (which is complete with the objectives of the research) is the total population of continuous features :

```
> total <- function(x){length(x)*mean(x)}
> attach(Lucy)
> total(Income); total(Employees); total(Taxes)
[1] 1035217
[1] 151950
[1] 28653.5

> tapply(Income,Level,total)
   Big Medium  Small
103706 487351 444160

> table(SPAM,Level)
     Level
SPAM  Big Medium Small
  no   26    291   620
  yes  57    446   956
```

Almost always, in most surveys, one want to obtain estimates by subgroups, in this case estimates of the total income per industrial level.

In this case the number of companies that send SPAM is discriminated by industrial level.

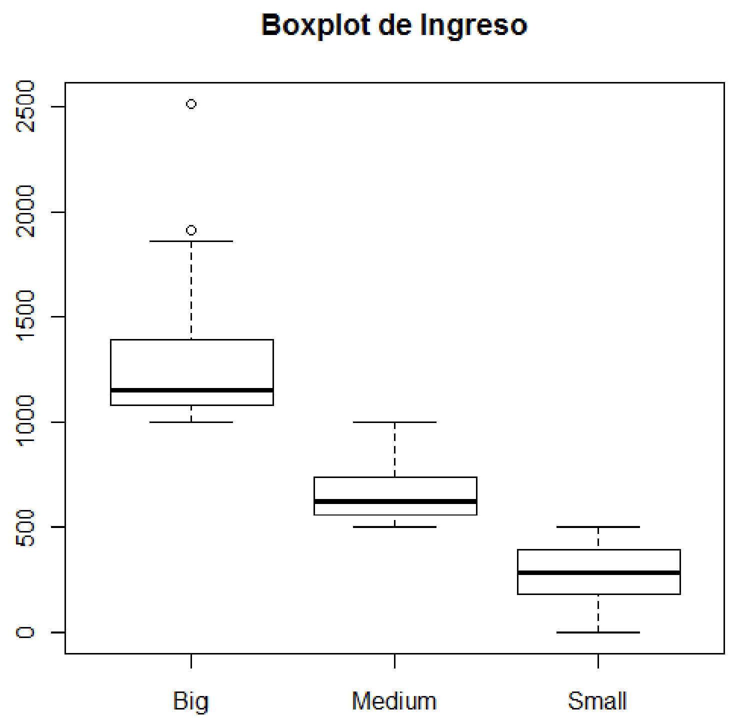
Industrial sector has a high income amounting to US 1,035,217 million, provides taxes by U.S. $ 28.653 million and employs to 151,950 people.

Note that most of the industrial sector income is acquired by the small and medium companies. However, on average, big firms bend the income of the medium companies, which in turn is three times the income of small companies. In absolute terms, the advertising strategy of sending spam to customers or potential customers deploy more often in small companies.

```
> xtabs(Income~Level+SPAM)
        SPAM
Level        no    yes
  Big     31914  71792
  Medium 190852 296499
  Small  175186 268974
```

The income of the companies that use SPAM as advertising strategy doubles the income of companies that do not use SPAM in almost all levels. That is because there exist more companies sending spam; then more companies implies more income.

```
> boxplot(Income  ~ Level,main=c("Boxplot de Ingreso"))
```



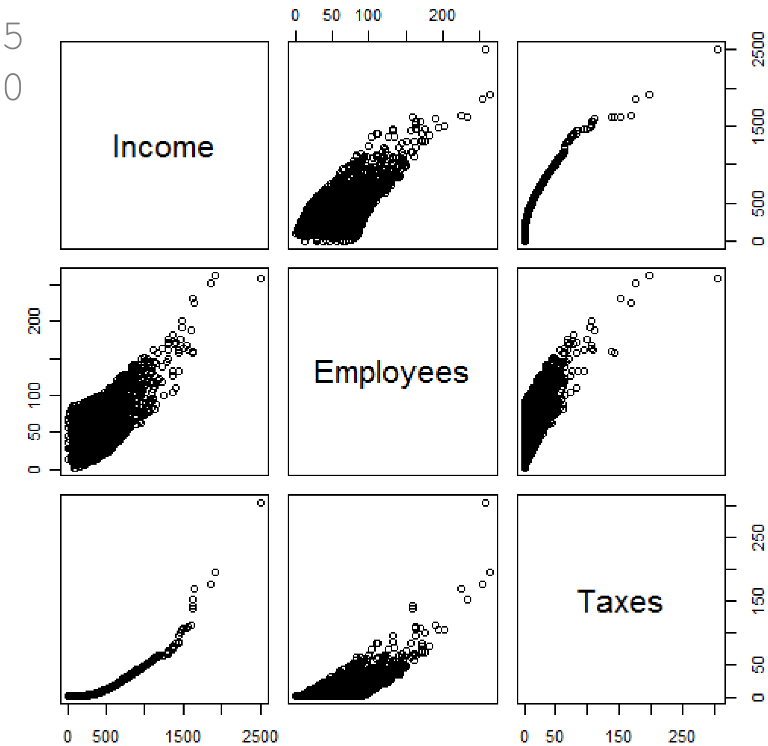
Big firms have higher incomes, provide a higher tax burden and employ more people than the small and medium enterprises. It is desirable that sampling frame contains the membership of each company to the industrial level because it is a good discriminant and allows implementation of appropriate sampling strategies to guide more precise estimates.

Marco and Lucy

It is also desirable to know the correlation between characteristics of interest. This can serve the time to bring the best sampling strategy.

```
> Datos <- data.frame(Income, Employees, Taxes)
> cor(Datos)
            Income Employees    Taxes
Income    1.000000  0.645536 0.916954
Employees 0.645536  1.000000 0.646855
Taxes     0.916954  0.646855 1.000000

> pairs(Datos)
```



Parameters of interest

Tabla 1.2: *Parámetros de la población discriminados por nivel industrial.*

			Ingreso	Impuestos	Empleados
Nivel	Grande	N total	83	83	83
		Suma	103.706	6.251	11.461
		Media	1.249	75	138
	Mediano	N total	737	737	737
		Suma	487.351	16.293	59.643
		Media	661	22	81
	Pequeño	N total	1.576	1.576	1.576
		Suma	444.160	6.110	80.846
		Media	282	4	51

Parameters of interest

Tabla 1.3: *Parámetros de la población discriminados por comportamiento publicitario.*

			Ingreso	Impuestos	Empleados
SPAM	no	N total	937	937	937
		Suma	397.952	10.593	59.600
		Media	425	11	64
	si	N total	1.459	1.459	1.459
		Suma	637.265	18.061	92.350
		Media	437	12	63

Parameters of interest

Tabla 1.4: *Parámetros de la población discriminados por nivel industrial y por comportamiento publicitario.*

		SPAM					
		no			si		
		N total	Suma	Media	N total	Suma	Media
Grande	Ingreso	26	31.914	1.227	57	71.792	1.260
	Impuestos	26	1.844	71	57	4.4.07	77
	Empleados	26	3.587	138	57	7.874	138
Mediano	Ingreso	291	190.852	656	446	296.499	665
	Impuestos	291	6.322	22	446	9.971	22
	Empleados	291	23.745	82	446	35.898	80
Pequeño	Ingreso	620	175.186	283	956	268.974	281
	Impuestos	620	2.427	4	956	3.683	4
	Empleados	620	32.268	52	956	48.578	51

Bernoulli sampling design

Suppose that you must select a sample with a Bernoulli sampling design. The expected sample size is N * pi = 400 companies.

As the population size is N = 2396, then the value is set to pi is 0.1669. To select the sample uses the `S. BE (N, prob)` of the `TeachingSampling` package whose parameters are `N`, the size population and `prob`, the value of the inclusion probability for each element of the population. This function uses the sequential algorithm that examine all elements of the population.

```
> # Uses the Marco and Lucy data to draw a Bernoulli sample
> data(Marco)
> data(Lucy)
> attach(Lucy)

> N <- dim(Marco)[1]
> # The population size is 2396. If the expected sample size is 400,
> # then, the inclusion probability must be 400/2396=0.1669
> sam <- S.BE(N,0.1669)
> # The information about the units in the sample is stored in an object
called data
> data <- Lucy[sam,]
> data
         ID Ubication  Level Zone Income Employees Taxes SPAM
7     AB007      c1k7  Small    A    490        22  10.5  yes
8     AB008      c1k8  Small    A    473        57  10.0  yes
. . .

> dim(data)
[1] 387   8
```

Note that the effective sample size is 387 companies

Estimating the parameters

Horvitz-Thompson estimator for the total

Once the data collection stage is done, we get a Lucy data file containing the values of the characteristics of interest for the selected companies.

The estimation stage is performed by using the function `E.BE (y, prob)` of the `TeachingSampling` package whose arguments are `y`, a vector or array containing the values of the characteristics of interest and `prob`, the probability of inclusion. In this case the length of each vector is n = 387.

This function gives the estimate of the total population and using the Horvitz-Thompson estimator, the estimate of the variance and estimated coefficient of variation .

```
> dim(data)
[1] 387   8

> sam <- S.BE(N,0.1669)
> # The information about the units in the sample is stored in data

> data <- Lucy[sam,]
> attach(data)
```

It is very important to use `attach` after the selection of the sample

```
> # The variables of interest are: Income, Employees and Taxes
> # This information is stored in a data frame called estima
> estima <- data.frame(Income, Employees, Taxes)

> E.BE(estima,0.1669)

                 Income    Employees        Taxes
Estimation 1.024661e+06 1.468484e+05 2.954164e+04
Variance   3.205513e+09 6.104305e+07 6.029255e+06
CVE        5.525459e+00 5.320456e+00 8.311841e+00
```

Alternative estimator for the total

With the help of the function `E.BE` it is possible to calculate the alternative estimate for the totals of interest. Just define the variable `n` indicating the effective sample size.

```
> N <- dim(Marco)[1]
> n <- dim(estima)[1]

> colSums(estima)
   Income Employees      Taxes
 171016.0   24509.0     4930.5

> (N/n)*colSums(estima)
   Income Employees      Taxes
1047965.1  150188.1    30213.5
```

As the alternative estimator is a ratio of estimators is not possible - at this point - to obtain an estimate for its variance and therefore can not calculate a cve .

Horvitz-Thompson estimator for the mean

It is also possible to estimate the mean by using the Horvitz-Thompson estimator. Of course, we have to estimate the variance and the cve.

```
> est.mean <- E.BE(estima,0.1669)[1,]/N
> est.mean
   Income Employees     Taxes
427.65504  61.28899  12.32957

> est.var <- E.BE(estima,0.1669)[2,]/N^2
> est.var
    Income  Employees      Taxes
558.372331  10.633165   1.050244

> est.cve <- 100*sqrt(est.var)/est.mean
> est.cve
   Income Employees     Taxes
 5.525459  5.320456  8.311841
```

Alternative estimator for the mean

It is also possible to calculate the alternative estimate for the mean of the characteristics of interest.

```
> (N/n)*colSums(estima)/N
   Income Employees     Taxes
437.38107  62.68286  12.60997
```

As with the alternative estimator for the total, it is not possible - at this point - to obtain an estimate for its variance and therefore cannot be calculated a cve.

Simple random sampling

Suppose that you must select a sample with a simple random design (SI).

1. It should be calculated the sample size of companies in the industrial sector.
2. It must be obtained statistical estimates of the total and average for population of the industrial sector.
3. Discriminated estimates should be obtained for all of the domains of interest.
4. Based on the results it should be proposed an economic policy for the industrial sector support.

The domains of interest are related to the advertising practices of companies. Then, there is the domain SPAM.YES for companies sending electronic publicity and SPAM.NO for companies that do not send these advertising.

Pilot sample

The sampling strategy to be used is as follows: The Horvitz-Thompson estimator applied to a simple random sampling design without replacement. You select a pilot `sample` of size 30 from the population. For this, once the data file LUCY is loaded, we use the sample function to extract the pilot sample. The characteristic of interest is the income of the companies, then we take the values of the variance and mean as estimates to be used for the calculation of sample size.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> sam <- sample(N,30)
> Ingresopiloto <- Income[sam]
> var(Ingresopiloto)
[1] 66952.62
> mean(Ingresopiloto)
[1] 455
```

Sample size: absolut error

The required estimates are such that :

- Absolute error: the margin of error for this study is 25 million dollars in total business income of the population.

- Confidence level: 95 %.

- By (3.2.16), n0 = 411.

$$n \geq \frac{n_0}{1 + \frac{n_0}{N}} \qquad (3.2.16)$$

- Then, n ≥ 351.

Sample size: relative error

The required estimates are such that :

• Relative error: the relative standard error must be less than 7% in total business income of the population

• Confidence level: 95 %.

• By (3.2.18), k0 = 446.

$$n \geq \frac{k_0}{1 + \frac{k_0}{N}} \qquad (3.2.18)$$

• Then, n ≥ 376.

In conclusion, we propose a sample size
n = 400

Simple random sampling without replacement

You must select a sample with a simple random sampling design without replacement (SI). To select the sample we use the function `S.SI (N, n)` of the package `TeachingSampling` whose parameters are `N`, population size and `n`, the sample size. This function uses the algorithm of Fan-Muller-Rezucha.

In this opportunity, it will be asked about income, taxes and number of employees in the fiscal year of interest, and also about the membership of the companies domains, i.e. if you send SPAM or not their customers or potential customers.

```
> N <- dim(Lucy)[1]
> n <- 400
> sam<-S.SI(N,n)
> # The information about the units in the sample is stored in an object
called data
> data <- Lucy[sam,]
> data
        ID Ubication  Level Zone Income Employees Taxes SPAM
1    AB001      c1k1  Small    A    281        41   3.0   no
3    AB003      c1k3  Small    A    405        68   7.0   no
7    AB007      c1k7  Small    A    490        22  10.5  yes
. . .

> dim(data)
[1] 400   8
```

Estimating the parameters

Horvitz-Thompson estimator for the total

Once the data collection stage is done, we get a Lucy data file containing the values of the characteristics of interest for the selected companies.

The estimation stage is done by using the function `E.SI(N, n, y)` of the `TeachingSampling` package whose arguments are the same as those of the function `S.SI` and `y`, a vector or array containing the values features of interest in the sample. In this case the length of each vector is n = 400. This function gives the estimate of the total population and using the Horvitz-Thompson estimator, the estimate of the variance and estimated coefficient of variation.

Horvitz-Thompson estimator for the total

```
> attach(data)
> # The variables of interest are: Income, Employees and Taxes
> # This information is stored in a data frame called estima

> estima <- data.frame(Income, Employees, Taxes)

> E.SI(N,n,estima)
                 Income    Employees        Taxes
Estimation 1.006769e+06 1.533440e+05 2.679028e+04
Variance   7.805793e+08 1.202052e+07 2.680269e+06
CVE        2.775100e+00 2.260971e+00 6.110996e+00
```

Horvitz-Thompson estimator for the mean

With the help of the function E.SI we can estimate the mean by using the Horvitz-Thompson estimator, also it is possible to estimate the variance of the estimator calculate the cve.

```
> est.mean <- E.SI(N,n,estima)[1,]/N
> est.mean
   Income Employees     Taxes
420.18750  64.00000  11.18125

> est.var <- E.SI(N,n,estima)[2,]/N^2
> est.var
      Income    Employees        Taxes
135.9700878   2.0938704   0.4668794

> est.cve <- 100*sqrt(est.var)/est.mean
> est.cve
   Income Employees     Taxes
 2.775100  2.260971  6.110996
```

Domains

```
> # The variable SPAM is a domain of interest
> Doma <- Domains(SPAM)
> # This function allows to estimate the parameters of the variables
   of interest for every category in the domain SPAM

> estima <- data.frame(Income, Employees, Taxes)
> SPAM.no <- estima*Doma[,1]
> SPAM.yes <- estima*Doma[,2]
```

Suppose that the domains of interest are the subgroups that send SPAM or not. This forms a partition of the population of industrial companies and note also that it is not know which firms tend to advertise through this medium. The function `Domains()` creates indicator variables for each domain of interest. Remember that these zeros and ones are multiplied with values of the characteristics of interest.

Horvitz-Thompson estimator for the total in each domain

```
> E.SI(N,n,SPAM.no)
                Income    Employees        Taxes
Estimation 3.656595e+05 5.710866e+04 9.500140e+03
Variance   7.495751e+08 1.544580e+07 1.190420e+06
CVE        7.487393e+00 6.881818e+00 1.148471e+01

> E.SI(N,n,SPAM.yes)
                Income    Employees        Taxes
Estimation 6.411097e+05 9.623534e+04 1.729014e+04
Variance   1.009908e+09 1.952392e+07 2.175746e+06
CVE        4.956882e+00 4.591440e+00 8.531113e+00
```

Note that the sum of the total estimated in the domains is equal to the HT estimate of the characteristics of interest. For example, `Income` must be

`365659.5 +641109.7 = 1006769`

It is important to perform this check!

Horvitz-Thompson estimator for the absolute size of the domains

With the help of the object DOMA and by using the function Domains it is possible to calculate the estimated absolute size for each of the two domains and obtain a corresponding cve.

```
> E.SI(N,n,Doma[,1])
                      y
Estimation   988.350000
Variance    2904.733402
CVE            5.453086

> E.SI(N,n,Doma[,2])
                      y
Estimation 1407.650000
Variance    2904.733402
CVE            3.828763
```

Horvitz-Thompson estimator for the mean of the domains

By using the above functions we can obtain an estimate for the average of each domain. As this is a ratio, we still cannot get a cve.

```
> E.SI(N,n,SPAM.no)[1,] / E.SI(N,n,Doma[,1])[1,]

   Income  Employees      Taxes
421.22424   61.59394   11.11818

> E.SI(N,n,SPAM.yes)[1,] / E.SI(N,n,Doma[,2])[1,]

   Income  Employees      Taxes
424.88511   62.34894   11.18085
```

Are there differences in average for companies that advertise via email?

Simple random sampling with replacement

Suppose that you must select a sample with a simple random design with replacement of sample size m = 400. There are several methods for the selection of a single sample with replacement, in the basic computing environment of R, the function `sample` allows to select such a sample when the replace option is set to `TRUE`.

```
sample(N,m, replace=TRUE)
```

In order to extract simple random samples with replacement, the `TeachingSampling` package uses a sequential algorithm based on the binomial distribution. The `S.WR` function has the following arguments: `N`, the size of the population and `m`, the size of the sample with replacement.

S.WR

```
> N <- dim(Marco)[1]
> m <- 400
> sam<-S.WR(N,m)
> # The information about the units in the sample is stored in an object
called data
> data <- Lucy[sam,]
> data
          ID Ubication  Level Zone Income Employees Taxes SPAM
16      AB016     c1k16  Small    A    340        12   5.0   no
25      AB025     c1k25  Small    A    365        49   6.0  yes
26      AB026     c1k26  Small    A    380        38   6.0   no
40      AB040     c1k40  Small    A    491        86  10.5  yes
45      AB045     c1k45  Small    A    365        53   6.0  yes
46      AB046     c1k46  Small    A    346        56   5.0   no
49      AB050     c1k49  Small    A    334        16   5.0   no
49.1    AB050     c1k49  Small    A    334        16   5.0   no
69 AB072     c1k69  Small    A    390        95   7.0  yes
...
```

Note that the company that is ranked as 49 in the sampling frame was selected twice in with-replacement sample.

Estimating the parameters

Hansen-Hurwitz estimator for the total

Once the data collection stage is done, we get a Lucy data file containing the values of the characteristics of interest for the selected companies.

The estimation stage is performed by using the function `E.WR(N, m, y)` of the `TeachingSampling` package whose arguments are those of the S.WR and y, a vector or array containing the values of the characteristics of interest in the sample. This function gives the estimate of the total population and using the Hansen-Hurwitz estimator, the estimate variance and coefficient of variation estimate.

Hansen-Hurwitz estimator for the total

```
> data <- Lucy[sam,]
> attach(data)

> # The variables of interest are: Income, Employees and Taxes
> # This information is stored in a data frame called estima

> estima <- data.frame(Income, Employees, Taxes)
> E.WR(N,m,estima)
                 Income    Employees        Taxes
Estimation 1.099207e+06 1.572734e+05 3.209143e+04
Variance   1.077487e+09 1.721914e+07 5.217604e+06
CVE        2.986253e+00 2.638459e+00 7.117813e+00
```

With the same sample size, the strategy that uses the sampling design simple random without replacement produces higher estimates for the coefficient of variation. It's the price you pay for duplicate information in the sample.

Hansen-Hurwitz estimator for the mean

With the help of the function E. WR it is possible to calculate the Hansen-Hurwitz estimation for the average features of interest, it is also possible to estimate the variance of the estimator and to calculate the cve.

```
> est.mean <- E.WR(N,m,estima)[1,]/N
> est.mean
   Income Employees     Taxes
458.76750  65.64000  13.39375

> est.var <- E.WR(N,m,estima)[2,]/N^2
> est.var
      Income   Employees       Taxes
187.6888683   2.9994246   0.9088611

> est.cve <- 100*sqrt(est.var)/est.mean
> est.cve
   Income Employees     Taxes
 2.986253  2.638459  7.117813
```

Horvitz-Thompson estimator for the total

As it is well known, once you define the selection probabilities for each element in the population, the inclusion probabilities are defined immediately. Therefore it is suitable to use the Horvitz-Thompson estimator to access to an estimate for the total of the characteristics of interest.

```
> # The vector of selection probabilities of units in the sample
> pk <- rep(1/N,m)
> # Computation of the inclusion probabilities
> Pik <- 1-(1-pk)^m

> HT(estima, Pik)
                [,1]
Income    1193283.34
Employees  170733.80
Taxes       34837.99
```

Estimates of variance and cve are not provided since the generic variance of the Horvtiz-Thompson estimator has a complex expression.

Design effect

The loss of efficiency in this strategy can be estimated with the Deff. Simply by doing the ratio of estimated variances it is possible to establish that - for this particular case - simple random sampling without replacement is better.

For simple random sampling without replacement:

```
                 Income    Employees        Taxes
Variance   7.805793e+08 1.202052e+07 2.680269e+06
```

For simple random sampling with replacement:

```
                 Income    Employees        Taxes
Variance   1.077487e+09 1.721914e+07 5.217604e+06
```

Estimates of the design effect:

```
                 Income    Employees        Taxes
Deff              1.371        1.433        1.944
```

Systematic sampling design

Note that the characteristics of interest are income, number of employees and taxes held in the last fiscal year and it can be assumed, correctly, that these features have no relation to the date of registration of the company in the sampling frame. Thus, it can happen that a young company, has a high yield, few employees and high return, but also can happen the contrary; in fact, this behavior is subject to the marketing used in each trading period and has low relation with the age of the business.

For the above reasons, it is assumed that the ordering of the sampling frame is completely random. It has been decided that the population will be partitioned into six groups, so that the effective sample size will be 399 or 400.

The sample selection is made by using the function `S.SY` whose arguments are `N`, the size of the population and `a`, the number of groups. This function assigns a random start and leaps, in this case, six in six elements to sweep the entire list.

```
> N <- dim(Marco)[1]
> a <- 6

> # The population is divided in 6 groups of size 399 or 400
> sam <- S.SY(N,a)
> data <- Marco[sam,]
> data

         ID Ubication  Level Zone
6     AB006      c1k6  Small    A
12    AB012     c1k12  Small    A
18    AB018     c1k18  Small    A
...
2385  AB912     c26k9    Big    E
2391  AB983    c26k15    Big    E

> dim(data)
[1] 399   4
```

Estimando los parámetros

Horvitz-Thompson estimator for the total

Once collected the information from the sample, we proceed to perform the estimate stage by using the function `E.SY` whose arguments are `N`, the population size and `y`, a dataset containing information on the characteristics of interest for each element in the sample.

```
> data <- Lucy[sam,]
> attach(data)

> estima <- data.frame(Income, Employees, Taxes)
> E.SY(N,a,estima)

                  Income    Employees         Taxes
Estimation 1.032540e+06 1.552320e+05 2.775300e+04
Variance   7.744526e+08 1.294529e+07 2.392375e+06
CVE        2.695197e+00 2.317793e+00 5.573201e+00
```

This is a conservative approximation to the variance assuming simple random sampling without replacement.

It has to be considered that the efficiency of this sampling strategy is larger than that performing simple random sampling.

Horvitz-Thompson estimator for the mean

With the help of the function E.SY it is possible to calculate the Horvitz-Thompson for the mean of the characteristics of interest, also it is possible to estimate the variance of the estimator and its corresponding cve.

```
> est.mean <- E.SY(N,a,estima)[1,]/N
> est.mean
   Income Employees     Taxes
430.94324  64.78798  11.58306

> est.var <- E.SY(N,a,estima)[2,]/N^2
> est.var
     Income   Employees       Taxes
134.9028862   2.2549572   0.4167308

> est.cve <- 100*sqrt(est.var)/est.mean
> est.cve
   Income Employees     Taxes
 2.695197  2.317793  5.573201
```

Intra-class correlation

$$\rho = 1 - \frac{n}{n-1}\frac{SCD}{SCT} \qquad (3.4.29)$$

This measure of correlation between pairs of elements of the groups formed takes a maximum value equal to one when SCE is null and takes a minimum value of (- 1 / n-1) when SCE is maximum. In particular, it is desirable for this strategy that ρ take values close to zero.

On the other hand, it is possible to show that the design effect, the ratio of the variances, is given by the following expression:

$$Deff = \frac{Var_{SIS}\hat{t}_{\pi}}{Var_{MAS}\hat{t}_{\pi}} = \frac{N-1}{N-n}[1+(n-1)\rho] \qquad (3.4.32)$$

Thus, we find that systematic sampling will be:

1. Just as efficiently by simple random sampling if ρ = (1 / 1-N).
2. Less efficient than simple random sampling if ρ> (1 / 1-N).
3. More efficient than simple random sampling it ρ <(1 / 1-N).

ANOVA

With the sums of squares is shown that this strategy is more efficient than simple random sampling. This suggests that the use of the term of the variance for a simple random sampling without replacement is a good approximation for the variance of systematic sampling because it overestimates the true variance.

```
> grupo <- as.factor(array(1:a,N))
> data(Lucy)
> attach(Lucy)
> anova(lm(Income~grupo))

Response: Income
            Df    Sum Sq Mean Sq F value Pr(>F)
grupo        5     12359    2472  0.0346 0.9994
Residuals 2390 170698187   71422

> n <- dim(data)[1]
> rho <- 1-(n/(n-1))*(170698187/(170698187+12359))
> rho
[1] -0.002439984
> rho < 1/(1-N)
[1] TRUE
```

Design effect

The gain in efficiency when using this design is nearly twenty-nine times since the design effect is approximately 0.034.

```
> Deff <- (N-1)*(1+(n-1)*rho)/(N-n)
> Deff
[1] 0.03464363
> 1/Deff
[1] 28.86534
```

On the other side, the true generic variance of the HT estimator for the characteristic of interest `Income` is

```
> VarHT <- N*12359
> VarHT
[1] 29612164
```

It is much lesser than the estimate yielded by the expression of simple random sampling without replacement.