IS 3001

Sampling Techniques

Group Project - Group 22

Group Members

S15352 -Fonseka G.T.A

S15665-K.K.D.T.R. Thathsarani

s15666-W.D.Dulakshi Thathsarani

s15355-K.K.D.S.N.Gunathilaka

Content

Introduction

The dataset “Billionaires” contains statistics on the world’s billionaires,including information about their industries , and personal details.

The dataset consists of both categorical and numerical variables and 2409 entries. finalWorth, personName ,age, country , industries , selfMade, status, gender , gdp_country , total_tax_rate_country and the population_country are the 11 variables in the dataset.

finalWorth variable represents the final net worth of the billionaire in U.S dollars .Country variable has 66 levels (countries). industries has 18 levels which are associated with the billionaire’s business interests.selfMade has two levels by representing as “True” and ”False” ,it indicates whether the billionaire is self-made. In this dataset gender also divides into two categories as “F” and “M”. “D”,“E”,”N”,”R”,”U”,"Split Family Fortune" are the six categories of status. Gross Domestic Product (GDP) for the billionaire’s country is represented by the gdp_country variable.Total tax rate and the population in the billionaire’s country are represented by the total_tax_rate_country and the population_country variables respectively.

This dataset was analyzed using Simple Random Sampling, Stratified Sampling, and Cluster Sampling separately by obtaining two samples for each design. All of these methods are explained in detail, in the next parts of the report.

Methodology

1.Sample size calculation

- Simple Random Sampling
  - ❖ First, we need to figure out how size of a sample we need. We can get that using the formula listed below.

$$n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2$$

  - ❖ Since the population is relatively small, we should use the finite population correction. Then the sample can be derived by the formula below

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

  - ❖ Here since we use R software for all the calculations, rsampcalc function that is included in sampler package is used for calculating the sample size.
  - ❖ We keep a margin of error of 3 and 5% type 1 error.
  - ❖ We get 740 as the sample size.

- Stratified Random Sampling

•In stratified Random sampling the population should be divided into strata(distinct subgroups).

•Then proportions of the sample were selected from each stratum to obtain the overall sample size .

•To select the overall sample size n we used the “rsampcalc” function in R

n=rsampcalc(nrow(Billionaires),3,95,0.5)

**Here we took tolerable margin of error as 3.

•We took our sample size (n) as 740..

•We used proportional allocation to find out the sample sizes for each stratum.(nh)

Strata (self made)	Population size (Nh)	Sample size (nh)
True	715	220
False	1694	520
Toal	2409	740

- Two Stage cluster Sampling

In two stage cluster sampling an SRS of a cluster is selected, then another SRS in each cluster is taken.

In our data set “Billionaires” clustering variable we used the variable “Country” which has 66 clusters.

Therefore, here to select the number of clusters used the formula below.

Rule of the thumb when selecting the number of clusters,

n=round(sqrt(M/2))

M is the witch is equal to the Total population divided into the number of clusters.

In our data set M=34; n=4

2.1 Stratification Variable

A stratification variable is a characteristic that divides the population of size N into H number of strata with stratum h has N_h(size of the h^{th} strata)sampling units. The values of N_1,N_2,….,N_H should be known .

$N_1 + N_2 + \ldots + N_H = N$ where N is the population size.

In our “Billionaires” data set , we selected “selfMade” as our stratification variable. Which divides the population into 2 distinct groups “True” (Self made) and “False” (not self made).

2.2 Cluster Variable

A cluster variable, in the context of statistics and data analysis, is a type of categorical variable used to group or categorize data into distinct clusters or groups.

In our “Billionaires “ data set , we selected “Country” as our Cluster variable. It have divided in to 66 categories. There are Algeria, Argentina, Armenia, Australia, Austria, Bahrain, Belgium, Brazil Cambodia, Canada, Chile, China, Colombia, Cyprus, Czech Republic, Denmark, Egypt, Finland France, Georgia, Germany, Greece, Hungary, India, Indonesia, Israel, Italy, Japan, Kazakhstan Latvia. Lebanon, Liechtenstein, Luxembourg, Malaysia, Mexico, Morocco, Nepal, Netherlands, New Zealand, Nigeria, Norway, Oman, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Singapore, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Tanzania, Thailand, Turkey, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Vietnam

To select the number of clusters, we used the formula below, which is the rule of the thumb when selecting the number of clusters, n=round(sqrt(34/2))

Below mentioned is the sample size of the different clusters selected in the first sample. The sampling is done 2 times.

An SRS of 4 clusters is selected. Then we took an SRS of each cluster .

```
> ClusterDetails
      Country Sample Size Population Size
1      Brazil          42              43
2  Uzbekistan           1               1
3 Netherlands          10              10
4  Kazakhstan           7               7
```

Results of the Study

Population Data

Mean	FinalWorth Population_country total_tax_rate_country	4749.855 512493689 43.84687
Total	FinalWorth Population_country total_tax_rate_country	11442400 1.234597e+12 105627.1
Proportion	Gender	Female 0.1178912 Male 0.8821088

Sample Data

1.Simple Random Sampling

Estimations :Sample 1

Sample size 740

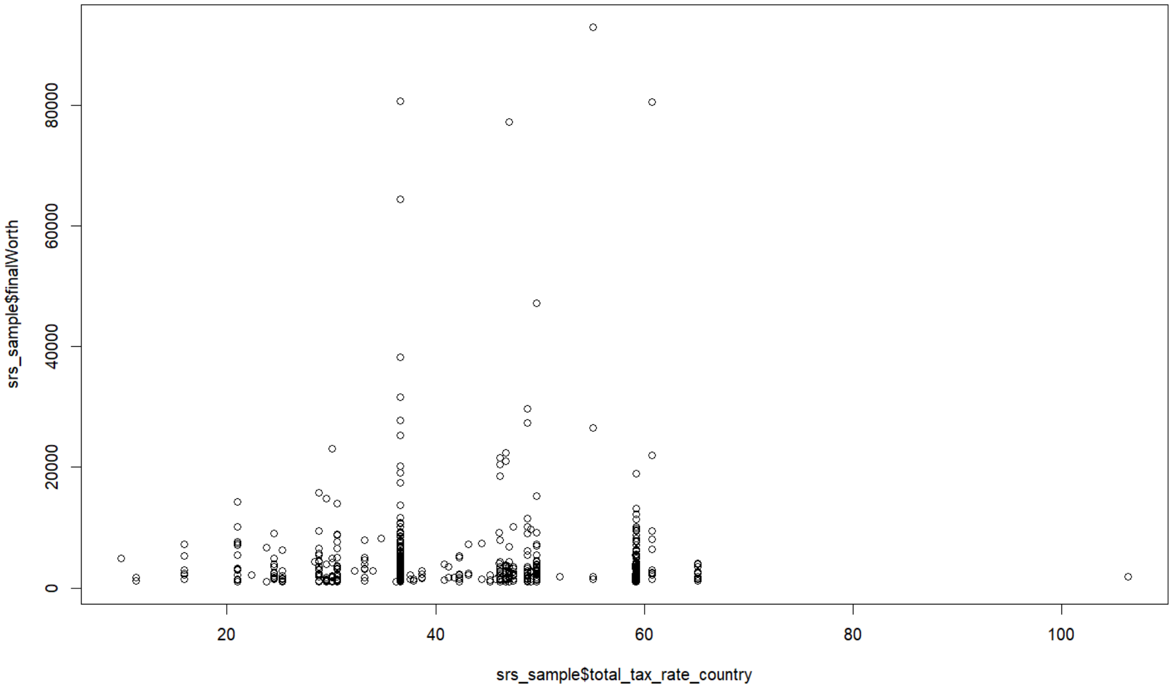
Mean		Mean	S.E
	FinalWorth	4274.9	281.85
	Population_country	511607152	20204001
	total_tax_rate_country	43.807	0.4524
Total		Total	S.E
	FinalWorth	3163400	208571
	Population_country	3.7859e+11	1.4951e+10
	total_tax_rate_country	32417	334.79
Proportion		Proportion	S.E
	Gender Female	0.11351	0.0117
	Gender Male	0.88649	0.0117

When we consider the actual population value and the sample values according to the mean, total and proportion we can say that,

- ➢ The sample mean final worth (4274.9) is lower than the population mean final worth (4749.855) by approximately 474.955, with a standard error of 281.85 indicating some uncertainty in the estimate.
- ➢ The sample mean for the population of the country (511,607,152) is slightly lower than the population mean (512,493,689) by approximately 886,537, with a standard error of 20,204,001 indicating a relatively large margin of error in the estimate
- ➢ The sample mean for the total tax rate of the country (43.807) is slightly lower than the population mean (43.84687) by approximately 0.04087, with a relatively small standard error of 0.4524, suggesting a relatively precise estimate of the population mean.

- ➢ The sample total of the total tax rate for the country (32,417) is substantially lower than the population total (105,627.1) by approximately 73,210.1, with a relatively small standard error of 334.79, suggesting a relatively precise estimate of the population total.
- ➢ The sample total of FinalWorth (3,163,400) is substantially lower than the population total (11,442,400) by approximately 8,279,000, with a relatively small standard error of 208,571, suggesting a relatively precise estimate of the population total.
- ➢ The sample total of the population in the country (3.7859e+11) is substantially lower than the population total (1.1234597e+12) by approximately 7.848e+11, with a standard error of 1.4951e+10 indicating a relatively small margin of error in the estimate.
- ➢ The sample proportions for gender (female: 0.11351, male: 0.88649) are close to the population proportions (female: 0.118, male: 0.882), with standard errors of 0.0117 indicating relatively small margins of error in the estimates, suggesting that the sample provides a good representation of the population's gender distribution.

Regression Estimation ,



Coefficients:

(Intercept)	total_tax_rate_country
4788.98	-11.74

According to this regression model,

The coefficient for "total tax rate" is -11.74, indicating that for each one-unit increase in the total tax rate, "Final worth" is estimated to decrease by 11.74 units. This model helps understand the relationship between "Final worth" and "total tax rate," with negative changes in "Final worth" associated with increases in the total tax rate.

Estimations :Sample 2

Sample size 740

Mean	Mean S.E FinalWorth 4581.5 303.3 Population_country 527291932 20499031 total_tax_rate_country 43.855 0.4264
Total	Total S.E FinalWorth 3390300 224438 Population_country 3.902e+11 1.5169e+10 total_tax_rate_country 32453 315.55
Proportion	Proportion S.E Gender Female 0.10946 0.0115 Gender Male 0.89054 0.0115

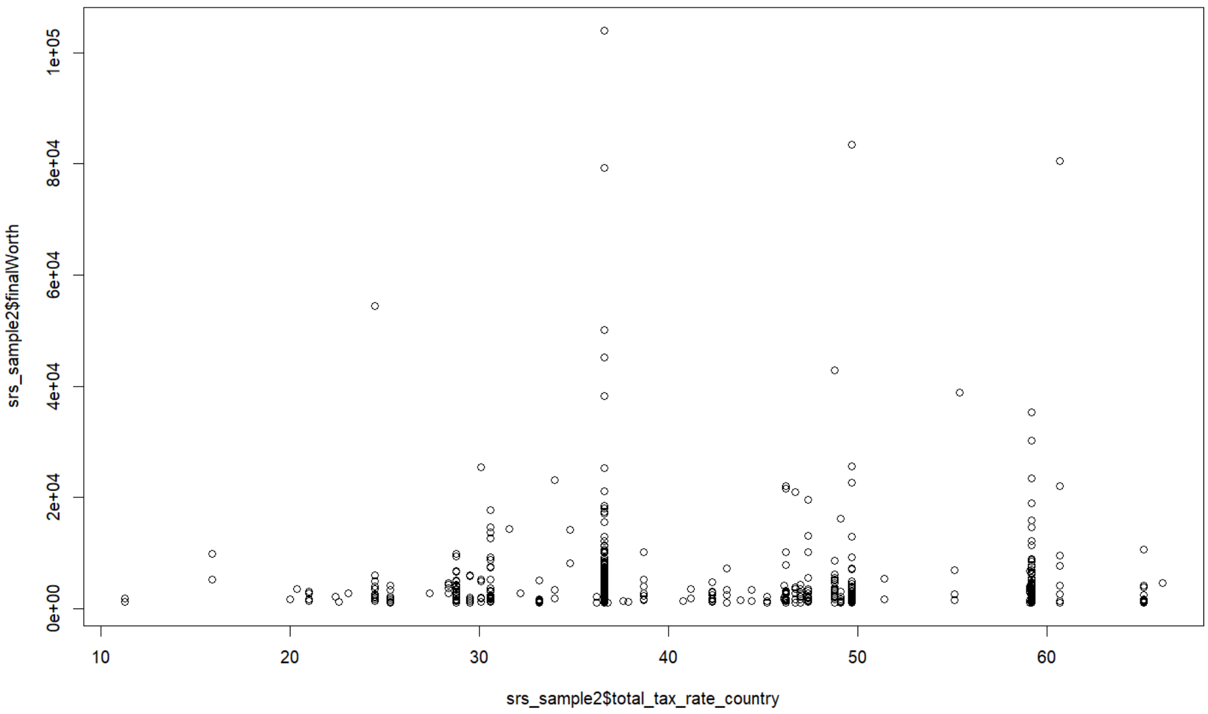
When we consider the actual population value and the sample values according to the mean, total and proportion we can say that,

- The sample mean for "Final worth" (4581.5) is slightly lower than the population mean (4749.855) by approximately 168.355, with a standard error of 303.3 indicating a moderate margin of error in the estimate. This suggests that the sample mean is in close proximity to the population mean, but there is some uncertainty in the accuracy of the estimate.
- The sample mean for the population of the country (527,291,932) is higher than the population mean (512,493,689) by approximately 14,798,243, with a standard error of 20,499,031 indicating some uncertainty in the estimate. This suggests that the sample mean may be an overestimate of the population mean, but the standard error is relatively small, providing a reasonably precise estimate.
- The sample mean for "total tax rate for the country" (43.855) is slightly higher than the population mean (43.84687) by approximately 0.00813, with a small standard error of 0.4264, indicating a relatively precise estimate. This suggests that the sample mean

provides a close estimate of the population mean, and the small standard error implies a low margin of error in the estimate.

- ➢ The sample total of FinalWorth (3,390,300) is substantially lower than the population total (11,442,400) by approximately 8,052,100, with a standard error of 224,438 indicating a relatively small margin of error in the estimate. This suggests that the sample total is quite different from the population total, and the small standard error implies a relatively precise estimate of this difference.
- ➢ The sample total of the population in the country (3.902e+11) is substantially lower than the population total (1.1234597e+12) by approximately 7.833397e+11, with a standard error of 1.5169e+10 indicating a relatively small margin of error in the estimate. This suggests that the sample total is quite different from the population total, and the small standard error implies a relatively precise estimate of this difference.
- ➢ The sample total of the "total tax rate for the country" (32,453) is substantially lower than the population total (105,627.1) by approximately 73,173.1, with a relatively small standard error of 315.55 indicating a relatively precise estimate of this difference. This suggests that the sample total provides a reasonably accurate estimate of the population total.
- ➢ The sample proportions for gender (female: 0.10946, male: 0.89054) are slightly lower than the population proportions (female: 0.118, male: 0.882), with standard errors of 0.0115 indicating a relatively small margin of error in the estimates. This suggests that the sample provides a fairly representative estimate of the population's gender distribution, with a slight underrepresentation of females.

Regression Estimation ,



Coefficients:

(Intercept)	total_tax_rate_country
5659.99	-24.59

The coefficient for "total tax rate" is -24.59, suggesting that for each one-unit increase in the total tax rate, "Final worth" is estimated to decrease by 24.59 units. This model helps understand the relationship between "Final worth" and "total tax rate," with negative changes in "Final worth" associated with increases in the total tax rate.

2.Stratified Random Sampling

Stratification Variable : selfMade

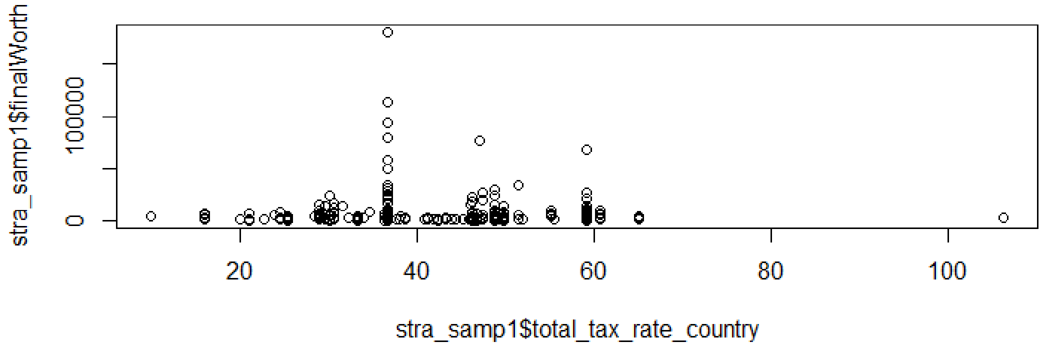
Sample size :740

Estimations :Sample 1

Mean	Mean S.E FinalWorth 4743.2 401.52 Population_country 505540130 19573738 total_tax_rate_country 43.822 0.439
Total	Total S.E FinalWorth 11407500 965657 Population_country 1.2158e+12 4.7075e+10 total_tax_rate_country 105392 1055.9
Proportion	Proportion S.E Gender Female 0.11486 0.0112 Gender Male 0.88514 0.0112

- When comparing estimated values from Stratified Random sample design with the actual population values we can see that ,
  - ➔ The variable finalWorth has nearly similar values for mean and total compared to the population mean and total of the variable finalWorth.
  - ➔ The estimated mean value for total_tax_rate_country is nearly the same as population total total_tax_rate_country.
  - ➔ The estimated proportions for gender variable have lower standard error (0.0112).
  - ➔ The estimated total value for population_country variable is approximately equal to the population total proportion_country value.

Regression Estimation



Coefficients :

(Intercept)	total_tax_rate_country
7452.61	-61.83

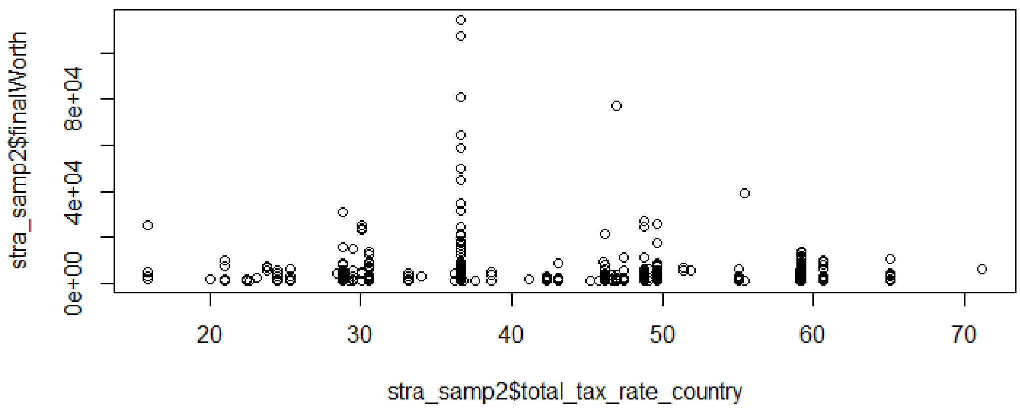
Calculated expected mean finalworth using regression model= 4743.096

Estimations :Sample 2

Mean		Mean	S.E
	FinalWorth	4735.7	344.85
	Population_country	505110276	19562296
	total_tax_rate_country	43.81	0.4391
Total		Total	S.E
	FinalWorth	11389300	829371
	Population_country	1.2148e+12	4.7047e+10
	total_tax_rate_country	105364	1055.9
Proportion		Proportion	S.E
	Gender Female	0.12838	0.0115
	Gender Male	0.87162	0.0115

- Here sample 1 and 2 are given nearly equivalent estimated values for the variables finalWorth, population_country and total_tax_rate_country .
- proportions for gender variables are very similar to actual proportions of gender.
- When comparing the sample 1 and sample 2 estimations with the population values, all three estimators mean, total and proportion are approximately equal.

Regression Estimation



Coefficients :

(Intercept)	total_tax_rate_country
8012.8	-74.8

Calculated expected mean finalworth using regression model= 4735.812

- The expected mean finalWorth obtained from the regression model for sample 1 is better estimation than the result obtained from sample 2 .

3.Cluster Sampling

- Clustering Variable : Country
- Number clusters in the population : 66
- Selected Clusters : 4
- Selected Sample 1 Cluster
- “France Uzbekistan Greece Australia”

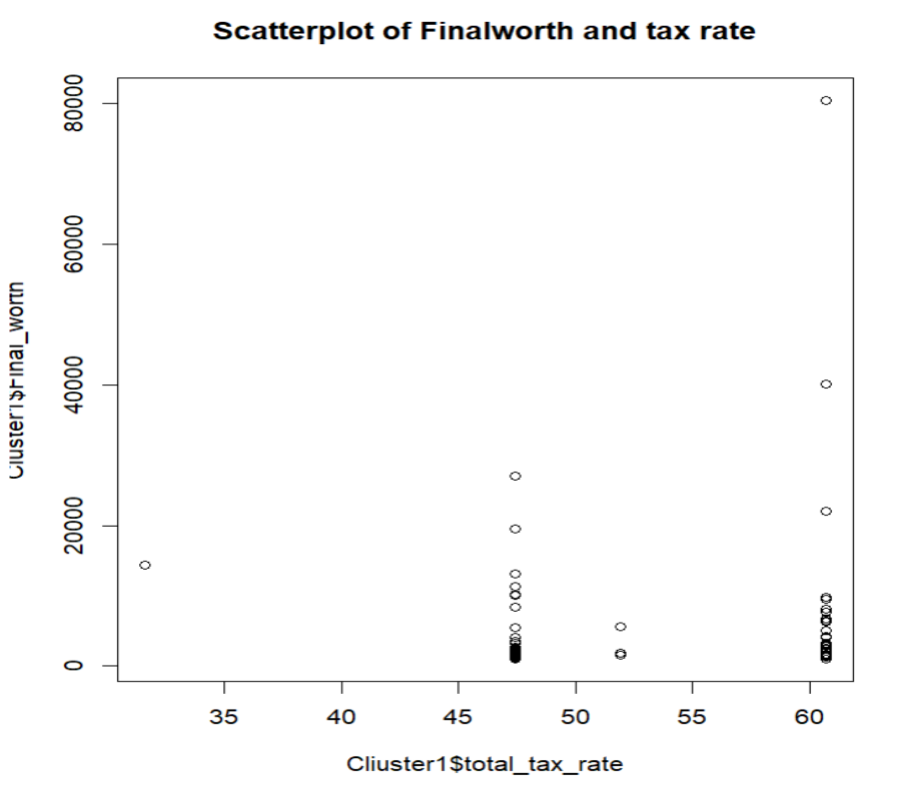
Country Sample	Size	Population Size
France	33	34
Uzbekistan	1	1
Greece	3	3
Australia	42	43

Estimations of Sample 1 Cluster

Mean		Mean	S.E
	FinalWorth	6107.1	1762.3
	Population_country	42638639	15778243
	total_tax_rate_country	52.954	5.0775
Total		Total	S.E
	FinalWorth	8162121	4556534
	Population_country	5.6987e+10	3.5369e+10
	total_tax_rate_country	70773	37332
Proportion		Proportion	S.E
	Gender Female	0.18941	0.0276
	Gender Male	0.81059	0.0276

- When comparing estimated values from Cluster sample design with the actual population values we can see that ,
    - ➔ The variable finalWorth has not nearly similar values for mean to the population mean of the variable finalWorth.
    - ➔ The population of the country is approximately equal to Actual population mean.
    - ➔ The estimated mean value for total_tax_rate_country is nearly the same as population total total_tax_rate_country.
    - ➔ The estimated proportions for gender variable have the same proportion values comparing actual proportions of the gender.
    - ➔ The estimated total value for population_country variable is not approximately equal to the population total proportion_country value.

Regression Estimation



```
Coefficients:
          (Intercept)  total_tax_rate_country
              -7494.1                   256.8
```

mean_Final_Worth = 5557.72 -49.68 * mean(data$total_tax_rate_country)

=3379.408

Ratio estimator

Ratios=

total_tax_rate_country

finalWorth 115.3274

SEs=

total_tax_rate_country
finalWorth 22.40642

- Selected Sample 2 Cluster

"South Korea" "Sweden" "Mexico" "Chile"

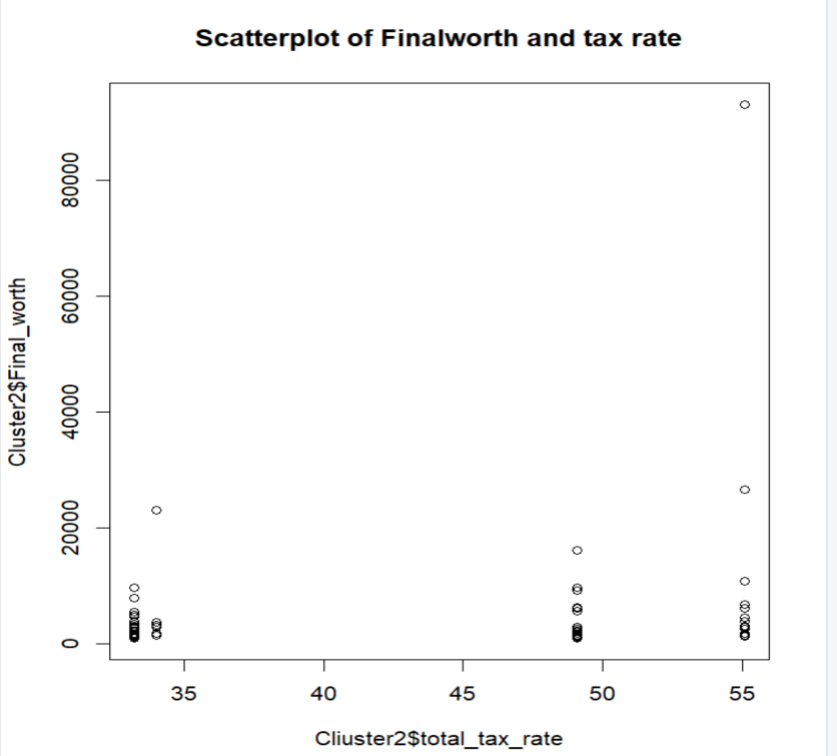
Country Sample	Size	Population Size
South Korea	29	29
Sweden	26	26
Mexico	13	13
Chile	6	6

Estimations of Sample 2 Cluster

Mean	Mean S.E FinalWorth 5059.5 1970.2 Population_country 47552435 22195246 total_tax_rate_country 42.699 5.6756
Total	Total S.E FinalWorth 6177600 1769569 Population_country 5.8062e+10 2.6408e+10 total_tax_rate_country 52135 14947
Proportion	Proportion S.E Gender Female 0.17568 0.0376 Gender Male 0.82432 0.0376

- When comparing estimated values from Cluster sample design with the actual population values we can see that ,
    - ➔ The variable finalWorth has nearly similar values for mean to the population mean of the variable finalWorth.
    - ➔ The population of the country is approximately equal to Actual population mean.
    - ➔ The estimated mean value for total_tax_rate_country is nearly the same as population total total_tax_rate_country.
    - ➔ The estimated proportions for gender variable have approximately the same proportion values comparing actual proportions of the gender.
    - ➔ The estimated total value for population_country variable is not approximately equal to the population total proportion_country value.And also total of FinalWorth and total_tax_rate country are the not approximately same value of their actual total values.

Regression Estimation




```
Coefficients:
           (Intercept)  total_tax_rate_country
               -5706.8                   252.1
```

mean_Final_Worth = 5557.72 -49.68 * mean(data$total_tax_rate_country)

=3379.408

Ratio Estimators

Ratios=

	total_tax_rate_country
finalWorth	118.4923

SEs=

	total_tax_rate_country
finalWorth	39.00957

Comparing Sample 1, Sample 2 and Actual values

Mean

	Population	Sample 1	Sample 2
FinalWorth	4749.855	6107.1	5059.5
Population_country	512493689	42638639	15778243
total_tax_rate_country	43.84687	52.954	42.699

Total

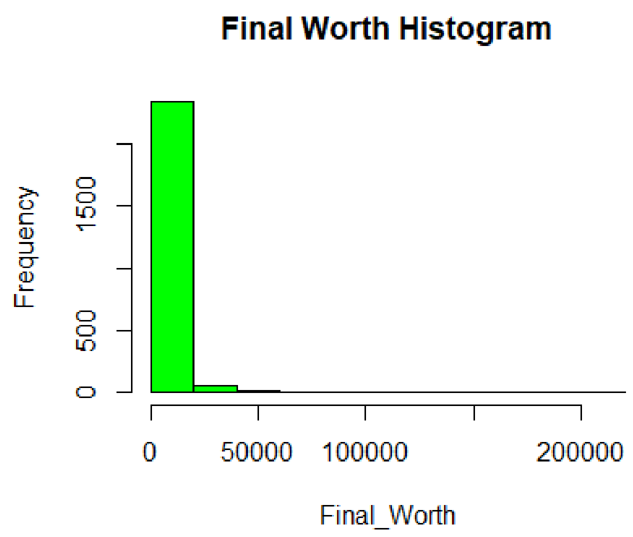
	Population	Sample 1	Sample 2
FinalWorth	11442400	8162121	6177600
Population_country	1.234597e+12	5.6987e+10	5.8062e+10
total_tax_rate_country	105627.1	70773	52135

Proportion-Gender

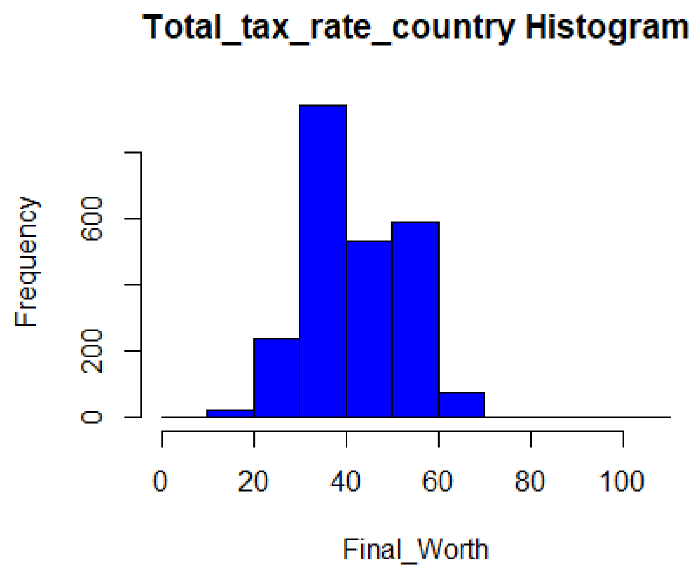
	Population	Sample 1	Sample 2
Female	0.1178912	0.18941	0.17568
Male	0.8821088	0.81059	0.82432

By comparing sample 1 and sample 2 Final worth of mean is approximately same for actual Final worth mean.Sample 1 and sample 2 mean population country is approximately the same but differ from mean population country. Mean Total tax rate country has some different population, sample 1 and sample 2. Total value of Final worth,total value of population country and total of total tax rate are different By comparing sample 1 sample 2 and population total.The proportion-Gender is approximately same in sample 1 and sample 2.And also approximately equal to population proportion of gender.

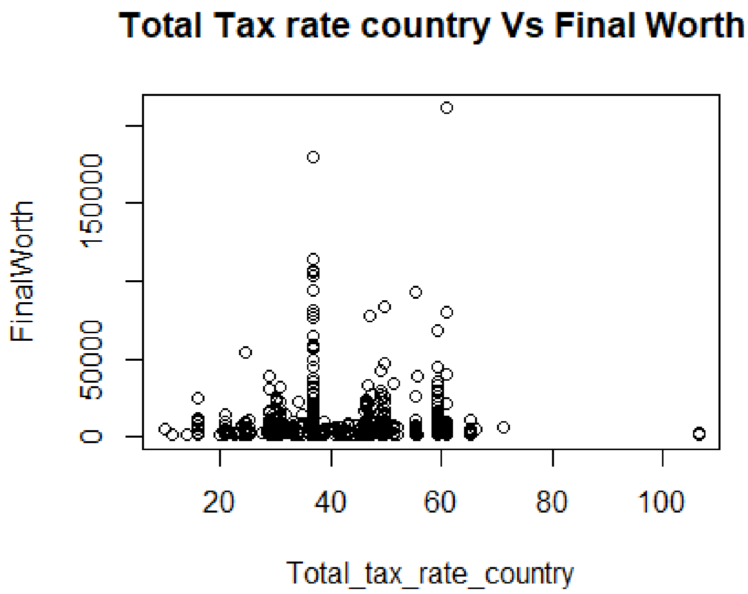
Graphical Analysis



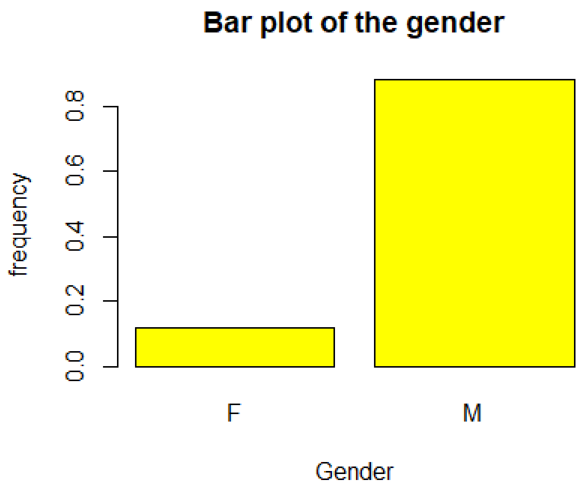
By Considering this Histogram in here Right Skewed sequence histogram represent. It’s Indicating that the majority of data points are clustered to the left with a tail extending towards the right.



In this graph a representative histogram of the Total_tax_rate_country data. This graph also has a Positive skewed distribution.



As in the above graph we are drawing the Final_Worth Vs Total_tax_rate_country graph we can see a linear pattern graph of the Final Worth and the Total_tax_rate_country between 0 and 50000. Here can also determine most of the FinalWorth data points are between 0 and 50000.



By considering the above bar plot showing that Proportion of the Male people are greater than the Proportion of the Female peoples.

Conclusion

By performing the above estimations and graphical analysis, we can make the following conclusions as bellow:

The results of this study that regards the sampling designs; simple random sampling, stratified random sampling and the two stage cluster sampling for the Billionaires are discussed above. Each of three sampling designs are built twice and compared with each other and with the actual population values. The results of this process illustrate that the estimated mean, total, proportion are suitable to explain the population with lower standard errors in all three sampling techniques.

- The Mean of the FinalWorth is around 4700 range
- The Mean of the Population country is around the 510000000
- The Mean of the total tax rate country is round up to the 43
- The Total number of final worth varying around the 11400000
- The Total number of Population country is around the 1.200000e+12
- The Total number of Total tax rate country is varying around 105000
- The proportion of Gender Male and Female approximately equal to the 0.11 and 0.88
- The proportion of the Male people is greater than the Proportion of the Female people.
- Most of the people on FinalWorth are between 0 and 50000.

As a whole, it is clear that results of the analysis do not differ significantly with the method of sampling. But the standard error of estimations in the Two-staged cluster sampling is higher when compared to the other two sampling methods. And the best Sampling method has most approximately equal value having to the actual population gives for us in the Stratified Sampling Technique. Therefore, Best Sampling method for analyzing this data set is the Stratified sampling technique.

R Code

Simple Random Sampling

```
#install.packages("survey")

#import libraries

library(survey)

library(sampler)

# Set the working directory to where your CSV file is located

setwd("C:/Users/Tharuka/Desktop/IS Project/new/")

# Read the CSV file

Data <- read.csv("Billionaires.csv")

# Actual values

attach(Data)

#Mean

#Final Worth

pop_mean_finalWorth=mean(finalWorth)

pop_mean_finalWorth

#Total Tax rate

pop_mean_total_tax_rate_country=mean(total_tax_rate_country)

pop_mean_total_tax_rate_country

#Country Population

pop_mean_population_country=mean(population_country)

pop_mean_population_country
```

```
#Total

#Final Worth

pop_total_finalWorth=sum(finalWorth)

pop_total_finalWorth

#Total Tax rate

pop_total_total_tax_rate_country=sum(total_tax_rate_country)

pop_total_total_tax_rate_country

#Country Population

pop_total_population_country=sum(population_country)

pop_total_population_country

#Proportion

#Gender

pop_proportion_gender=table(gender)/length(gender)

pop_proportion_gender

detach(Data)

#SRS_1

set.seed(123)

#sample size for SRS

srs_size=rsampcalc(nrow(Data),e=3,ci=95)

srs_size

#drawing a SRS 1

srs_sample=rsamp(Data,n=srs_size,rep =FALSE)

srs_sample

#Estimations SRS 1
```

```
attach(srs_sample)

#sample mean

#Final Worth

srs_sample_design=svydesign(id=~1,strata=NULL,data =srs_sample)

srs_sample_mean_for_finalWorth=svymean(~finalWorth,srs_sample_design)

srs_sample_mean_for_finalWorth

#Total Tax rate

srs_sample_design=svydesign(id=~1,strata=NULL,data =srs_sample)

srs_sample_mean_for_total_tax_rate_country=svymean(~total_tax_rate_country,srs_sample_design)

srs_sample_mean_for_total_tax_rate_country

#Country Population

srs_sample_design=svydesign(id=~1,strata=NULL,data =srs_sample)

srs_sample_mean_for_population_country=svymean(~population_country,srs_sample_design)

srs_sample_mean_for_population_country

#sample total

#Final Worth

srs_total_for_finalWorth=svytotal(~finalWorth,srs_sample_design)

srs_total_for_finalWorth

#Total Tax rate

srs_total_for_total_tax_rate_country=svytotal(~total_tax_rate_country,srs_sample_design)

srs_total_for_total_tax_rate_country

#Country Population

srs_total_for_population_country=svytotal(~population_country,srs_sample_design)

srs_total_for_population_country
```

```
#sample Proportion

#Gender

sample_prop=svymean(~gender,srs_sample_design)

sample_prop

#Regression estimation

#plot(srs_sample$total_tax_rate_country,srs_sample$finalWorth)

#fitting linear regression model

lm(finalWorth~total_tax_rate_country,srs_sample)

#-------------------------------------------------------------------

#Estimations SRS 2

#SRS_2

set.seed(321)

#sample size for SRS

srs2_size=rsampcalc(nrow(Data),e=3,ci=95)

srs2_size

#drawing a SRS 1

srs_sample2=rsamp(Data,n=srs2_size,rep =FALSE)

srs_sample2

attach(srs_sample2)

#sample mean

#Final Worth

srs_sample2_design=svydesign(id=~1,strata=NULL,data =srs_sample2)

srs_sample2_mean_for_finalWorth=svymean(~finalWorth,srs_sample2_design)

srs_sample2_mean_for_finalWorth
```

```
#Total Tax rate

srs_sample2_design=svydesign(id=~1,strata=NULL,data =srs_sample2)

srs_sample2_mean_for_total_tax_rate_country=svymean(~total_tax_rate_country,srs_sample2_design)

srs_sample2_mean_for_total_tax_rate_country

#Country Population

srs_sample2_design=svydesign(id=~1,strata=NULL,data =srs_sample2)

srs_sample2_mean_for_population_country=svymean(~population_country,srs_sample2_design)

srs_sample2_mean_for_population_country

#sample total

#Final Worth

srs2_total_for_finalWorth=svytotal(~finalWorth,srs_sample2_design)

srs2_total_for_finalWorth

#Total Tax rate

srs2_total_for_total_tax_rate_country=svytotal(~total_tax_rate_country,srs_sample2_design)

srs2_total_for_total_tax_rate_country

#Country Population

srs2_total_for_population_country=svytotal(~population_country,srs_sample2_design)

srs2_total_for_population_country

#sample Proportion

#Gender

sample2_prop=svymean(~gender,srs_sample2_design)

sample2_prop

#Regression estimation

plot(srs_sample2$total_tax_rate_country,srs_sample2$finalWorth)
```

```
#fitting linear regression model

lm(finalWorth~total_tax_rate_country,srs_sample2)
```

Stratified Sampling

```
data=read.csv("F:/3rd Year/IS 3001 Sampling Techniques/Billionaires.csv")

data

setwd("F:\\3rd Year\\IS 3001 Sampling Techniques")

getwd()

library(tidyverse)

glimpse(Billionaires)

names(Billionaires)

unique(Billionaires$personName)

unique(Billionaires$country)

unique(Billionaires$industries)

unique(Billionaires$selfMade)

unique(Billionaires$status)

unique(Billionaires$gender)

#Categorical variables

Billionaires$country<-as.factor(Billionaires$country)

Billionaires$industries<-as.factor(Billionaires$industries)

Billionaires$selfMade<-as.factor(Billionaires$selfMade)

Billionaires$status<-as.factor(Billionaires$status)

Billionaires$gender<-as.factor(Billionaires$gender)

levels(Billionaires$country)

levels(Billionaires$industries)
```

```
levels(Billionaires$selfMade)

levels(Billionaires$status)

levels(Billionaires$gender)

#Finding missing values.

colSums(is.na(Billionaires))

attach(Billionaires)

#install.packages("survey")

#install.packages("sampler")

library(survey)

library(readxl)

library(sampler)

#sample size n

size <- rsampcalc(nrow(Billionaires), 3, 95, 0.5)

size

#sample size for stratified sampling nh

str_size=ssampcalc(Billionaires,740,`selfMade`) #determine sample size by strata using proportional allocation

str_size

#draw stratifies samples without replacement using proportional allocation

stra_samp1=ssamp(Billionaires,740,`selfMade`)

stra_samp1

attach(stra_samp1)

#sample weights

stra_samp1$w=3.25
```

```
#Defining survey design object

stra_design=svydesign(id=~1,strata=`selfMade`,data=stra_samp1,weights=~w)

#estimate population mean of final worth

str_mean_worth=svymean(~finalWorth,stra_design)

str_mean_worth

#estimate population mean of population_country

str_mean_pop=svymean(~population_country,stra_design)

str_mean_pop

#estimate population mean of total tax rate

str_mean_tax=svymean(~total_tax_rate_country,stra_design)

str_mean_tax

#estimate total final worth

str_total_worth=svytotal(`finalWorth`,stra_design)

str_total_worth

#estimate total population_country

str_total_pop=svytotal(`population_country`,stra_design)

str_total_pop

#estimate total tax

str_total_tax=svytotal(`total_tax_rate_country`,stra_design)

str_total_tax

#estimate population proportion of gender

str_prop_gender=svymean(~gender,stra_design)

str_prop_gender

detach(stra_samp1)
```

```
#actual population values
attach(Billionaires)
population_mean_worth=mean(finalWorth)
population_mean_worth
population_mean_pop=mean(population_country)
population_mean_pop
population_mean_tax=mean(total_tax_rate_country)
population_mean_tax
population_total_worth=sum(finalWorth)
population_total_worth
population_total_pop=sum(population_country)
population_total_pop
population_total_tax=sum(total_tax_rate_country)
population_total_tax
pop_prop_gender=table(gender)/length(gender)
pop_prop_gender
detach(Billionaires)
#Regression estimation
plot(stra_samp1$total_tax_rate_country,stra_samp1$finalWorth)
#fitting linear regression model
lm(finalWorth~total_tax_rate_country,stra_samp1)
mean_worth_reg1=7452.61+(-61.83*43.822)
mean_worth_reg1
#Getting another stratified sampling
```

```
stra_samp2=ssamp(Billionaires,740,`selfMade`)

stra_samp2

attach(stra_samp2)

#sample weights

stra_samp2$w=3.25

#Defining survey design object

stra_design2=svydesign(id=~1,strata=`selfMade`,data=stra_samp2,weights=~w)

#estimate population mean of final worth

str_mean_worth2=svymean(~finalWorth,stra_design2)

str_mean_worth2

#estimate population mean of population_country

str_mean_pop2=svymean(~population_country,stra_design2)

str_mean_pop2

#estimate population mean of total tax rate

str_mean_tax2=svymean(~total_tax_rate_country,stra_design2)

str_mean_tax2

#estimate total final worth

str_total_worth2=svytotal(`finalWorth`,stra_design2)

str_total_worth2

#estimate total population_country

str_total_pop2=svytotal(`population_country`,stra_design2)

str_total_pop2

#estimate total tax

str_total_tax2=svytotal(`total_tax_rate_country`,stra_design2)
```

```
str_total_tax2

#estimate population proportion of gender

str_prop_gender2=svymean(~gender,stra_design2)

str_prop_gender2

detach(stra_samp2)

#Regression estimation

plot(stra_samp2$total_tax_rate_country,stra_samp2$finalWorth)

#fitting linear regression model

lm(finalWorth~total_tax_rate_country,stra_samp2)

mean_worth_reg2=8012.8+(-74.8*43.81)

mean_worth_reg2
```

Cluster Sampling

```
install.packages("survey")

install.packages("readxl")

install.packages("sampler")

library(survey)

library(readxl)

library(sampler)

#Link to the Dataset

setwd("C:\\Users\\Tashini Ramindi\\Desktop\\Cluster_sampling_2023")

data=read.csv("Billionaires.csv")

attach(Billionaires)

#Cluster sampling

set.seed(1234)
```

```
e=3

ci=95

#Obtaining a sample from two stage Cluster Sampling

#Selecting the number of clusters

#Cluserting variable = country

count_table=table(country)

count_table

srs_size=rsampcalc(nrow(data),e,ci)

srs_size

strata_size=ssampcalc(data,srs_size,country)

Strata_size

n1=1+1+6+7+88+26+2+2+9+1+2+75+16+6

n2=3+3+504+4+3+55+1+1+1+2+5+1+80

n3=1+43+1+7+3+37+1+10+2+3+29+25+754

n4=43+1+5+34+157+7+11+2+14+79+24+25+1

n5=11+40+7+1+25+1+13+3+5+46+26+6+1

(n1+n2+n3+n4)/66

n=round(sqrt(34/2))

n

#Number of clusters in the population

N=length(unique(data$country))

N

#Selecting the First Cluster Sample

#Selecting the clusters using SRS
```

```
clusters1 = sample(x = unique(data$country),size = n,replace = F)

clusters1

#Variable to save data after selecing clusters

Cluster1 = c()

#variable to save sample sizes

m=numeric(n)

#Variable to save population size of clusters

ClusterSize = numeric(n)

for (i in 1:n){

  #Dividing the dataset into clusters

  dat =  data[data$country==clusters1[i],]

  ClusterSize[i] = nrow(dat)

  #Selecting sample sizes for each cluster

  m[i] = rsampcalc(N = nrow(dat),e = e,ci = ci)

  #selecting a sample from each cluster and saving it

  Cluster1=rbind(Cluster1,rsamp(df = dat,n = m[i],rep = F))

}

ClusterDetails = data.frame(clusters1,m,ClusterSize)

colnames(ClusterDetails) = c("Country","Sample Size","Population Size")

ClusterDetails

View(Cluster1)

#Calculating sample weights

pw = numeric(0)

for (i in 1:nrow(Cluster1)){
```

```
  pw[i] =  (N*ClusterDetails[ClusterDetails$Country==Cluster1[i,]$country,]$`Population Size`)/
           (n*ClusterDetails[ClusterDetails$Country==Cluster1[i,]$country,]$`Sample Size`)
}
#Adding weights column to the main data frame
Cluster1=cbind(Cluster1,pw)
View(Cluster1)
# Select the most appropriate variables and estimate mean, proportion, total,
#Survey Design
#Clustering variables are ids and the country
Cluster_Design = svydesign(ids = ~country, weights = ~pw, data = Cluster1)
#Calculating mean, proportion, total
#Proportions
Cluster_Pgender1 = svymean(~gender,design = Cluster_Design)
Cluster_Pgender1
#Means
Cluster_mean_FW1 = svymean(~finalWorth,design = Cluster_Design)
Cluster_mean_FW1
Cluster_mean_TT1 = svymean(~total_tax_rate_country,design = Cluster_Design)
Cluster_mean_TT1
Cluster_mean_PP1 = svymean(~population_country,design = Cluster_Design)
Cluster_mean_PP1
#totals
Cluster_total_FW1 = svytotal(~finalWorth,design = Cluster_Design)
Cluster_total_FW1
```

```
Cluster_total_TT1 = svytotal(~total_tax_rate_country,design = Cluster_Design)

Cluster_total_TT1

Cluster_total_PP1 = svytotal(~population_country,design = Cluster_Design)

Cluster_total_PP1

#Selecting the second Cluster sample

set.seed(4567)

#Selecting the clusters using SRS

clusters2 = sample(x = unique(data$country),size = n,replace = F)

clusters2

#Variable to save data after selecing clusters

Cluster2 = c()

#variable to save sample sizes

m=numeric(n)

#Variable to save population size of clusters

ClusterSize = numeric(n)

for (i in 1:n){

  #Dividing the dataset into clusters

  dat =  data[data$country==clusters2[i],]

  ClusterSize[i] = nrow(dat)

  #Selecting sample sizes for each cluster

  m[i] = rsampcalc(N = nrow(dat),e = e,ci = ci)

  #selecting a sample from each cluster and saving it

  Cluster2=rbind(Cluster2,rsamp(df = dat,n = m[i],rep = F))

}
```

```
ClusterDetails = data.frame(clusters2,m,ClusterSize)

colnames(ClusterDetails) = c("country","Sample Size","Population Size")

ClusterDetails

View(Cluster2)

#Calculating sample weights

pw = numeric(0)

for (i in 1:nrow(Cluster2)){

  pw[i] =  (N*ClusterDetails[ClusterDetails$country==Cluster2[i,]$country,]$`Population Size`)/

            (n*ClusterDetails[ClusterDetails$country==Cluster2[i,]$country,]$`Sample Size`)

}

#Adding weights column to the main data frame

Cluster2=cbind(Cluster2,pw)

View(Cluster2)

#Survey Design

#Clustering variables are ids and the country

Cluster_Design = svydesign(ids = ~country, weights = ~pw, data = Cluster2)

#Calculating mean, proportion, total

#Proportions

Cluster_Pgender2 = svymean(~gender,design = Cluster_Design)

Cluster_Pgender2

#Means

Cluster_mean_FW2 = svymean(~finalWorth,design = Cluster_Design)

Cluster_mean_FW2

Cluster_mean_TT2 = svymean(~total_tax_rate_country,design = Cluster_Design)
```

```
Cluster_mean_TT2

Cluster_mean_PP2 = svymean(~population_country,design = Cluster_Design)

Cluster_mean_PP2

#totals

Cluster_total_FW2 = svytotal(~finalWorth,design = Cluster_Design)

Cluster_total_FW2

Cluster_total_TT2 = svytotal(~total_tax_rate_country,design = Cluster_Design)

Cluster_total_TT2

Cluster_total_PP2 = svytotal(~population_country,design = Cluster_Design)

Cluster_total_PP2

# Compare estimates with the actual values from the population.

# Compare the estimates obtained from the two samples under each design.

#Actual values from the Population

#Proportions

Pop_Agender = table(data$gender)/length(data$gender)

Pop_Agender

#Means

Pop_mean_FW = mean(data$finalWorth)

Pop_mean_FW

Pop_mean_TT = mean(data$total_tax_rate_country)

Pop_mean_TT

Pop_mean_PC = mean(data$population_country)

Pop_mean_PC

#Totals
```

```
Pop_total_FW = sum(data$finalWorth)

Pop_total_FW

Pop_total_TT = sum(data$total_tax_rate_country)

Pop_total_TT

Pop_total_PC = sum(data$population_country)

Pop_total_PC

#Comparing

Comp = data.frame("Cluster Sample 1" =
round(c(Cluster_mean_FW1,Cluster_mean_TT1,Cluster_mean_PP1,

                                        Cluster_total_FW1,Cluster_total_TT1,Cluster_total_PP1 ),

                                        digits = 3),

                    "Cluster Sample 2" =
round(c(Cluster_mean_FW2,Cluster_mean_TT2,Cluster_mean_PP2,

                                        Cluster_total_FW2,Cluster_total_TT2,Cluster_total_PP2 ),

                                        digits = 3),

                    "Population" = round(c(Pop_mean_FW,Pop_mean_TT,Pop_mean_PC,

                              Pop_total_FW,Pop_total_TT,Pop_total_PC),

                              digits = 3))

rownames(Comp) = c("Mean of Final worth","Mean of Total tax rate country","Mean of Population
Country",

                    "Total of Final worth","Total of Total tax rate country","Total of Population Country")

Comp

Comp_2 = data.frame("Cluster Sample 1" = round(c(Cluster_mean_FW1),digits = 3),

                    "Cluster Sample 2" = round(c(Cluster_mean_FW2), digits = 3),

                    "Population" = round(c(Pop_mean_FW), digits = 3))
```

```
Comp_2

Comp_3 = data.frame("Cluster Sample 1" = round(c(Cluster_Pgender1),digits = 3),
                    "Cluster Sample 2" = round(c(Cluster_Pgender2), digits = 3),
                    "Population" = round(c(Pop_Agender), digits = 3))


rownames(Comp_3)=c("Female","Male")

Comp_3

# Perform ratio or regression estimations

#Regression estimation

plot(Cluster1$total_tax_rate_country,Cluster1$finalWorth,
          main="Scatterplot of Finalworth and tax rate",
          xlab="Cliuster1$total_tax_rate",ylab="Cluster1$Final_worth")

RegressionLm = lm(finalWorth~total_tax_rate_country,data = Cluster1)

RegressionLm

mean_Final_Worth = 5557.72 -49.68 * mean(data$total_tax_rate_country)

mean_Final_Worth

#Ratio Estimation

r=svyratio(~finalWorth,~total_tax_rate_country,Cluster_Design)

r

predict(r,mean(data$total_tax_rate_country))

plot(Cluster2$total_tax_rate_country,Cluster2$finalWorth,
          main="Scatterplot of Finalworth and tax rate",
          xlab="Cliuster2$total_tax_rate",ylab="Cluster2$Final_worth")

RegressionLm = lm(finalWorth~total_tax_rate_country,data = Cluster2)
```

RegressionLm

mean_Final_Worth = 5557.72 -49.68 * mean(data$total_tax_rate_country)

mean_Final_Worth

#Ratio Estimation

r=svyratio(~finalWorth,~total_tax_rate_country,Cluster_Design)

R

predict(r,mean(data$total_tax_rate_country))

Graphical Analysis

hist(finalWorth, col="green", xlab="Final_Worth", main="Final Worth Histogram")

hist(total_tax_rate_country, col="blue", xlab="Final_Worth", main="Total_tax_rate_country Histogram")

plot(total_tax_rate_country, finalWorth, xlab="Total_tax_rate_country", ylab="FinalWorth", main="Total Tax rate country Vs Final Worth")

barplot(Pop_Agender,main="Bar plot of the gender", xlab="Gender", ylab="frequency",col="yellow")